Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Unraveling Green Innovation Drivers in High Carbon-Lock-in Process Industries: An Interpretable Machine Learning Approach with Temporal Validation

Hao Huang
Shandong Institute of Petroleum and Chemical Technology
Yujun Hu
hu342758@163.com

Shandong Institute of Petroleum and Chemical Technology
Xinqiang Li
PetroChina Xinjiang Oilfield Company Luliang Oilfield Operation Area
Yiwei Huang
Shandong Institute of Petroleum and Chemical Technology
Wenxuan Duan
Shandong Institute of Petroleum and Chemical Technology
Bowen Zhang
Shandong Institute of Petroleum and Chemical Technology
Anxue Li
Shandong Institute of Petroleum and Chemical Technology
Yang Li
Shandong Institute of Petroleum and Chemical Technology

Article

Keywords: green innovation, high carbon lock-in, petrochemical industry, XGBoost-SHAP, environmental regulation, threshold effect

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11206495/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Unraveling Green Innovation Drivers in High Carbon-Lock-in Process Industries: An Interpretable Machine Learning Approach with Temporal Validation

Hao Huang 1,2, Yujun Hu 1*, Xinqiang Li 3, Yiwei Huang 1, Wenxuan Duan 1, Bowen Zhang 1, Anxue Li 1,2, Yang Li 1

1 School of Economics and Management, Shandong Institute of Petroleum and Chemical Technology, Dongying, China

2 Key Laboratory of Digital Transformation and Application of Energy and Chemical Enterprises in Dongying City, Dongying, China

3 PetroChina Xinjiang Oilfield Company Luliang Oilfield Operation Area, Karamay, China

* Corresponding author: Yujun Hu. E-mail: hu342758@163.com. ORCID: 0009-0005-3957-5375

Hao Huang: ORCID 0009-0008-5230-2154; E-mail: 13805460321@139.com

Yujun Hu: ORCID 0009-0005-3957-5375; E-mail: hu342758@163.com

Xinqiang Li: ORCID 0009-0003-2264-2346; E-mail: liuqing@zcpei.cn

Yiwei Huang: ORCID 0009-0001-7580-0315; E-mail: yw2324515100@126.com

Wenxuan Duan: ORCID 0009-0003-9104-8076; E-mail: d3037059238@126.com

Bowen Zhang: ORCID 0009-0000-6574-2548; E-mail: zbw200506@126.com

Anxue Li: ORCID 0009-0008-4379-8063; E-mail: Anxue777@163.com

Yang Li: ORCID 0009-0008-1233-2631; E-mail: 2269932650@qq.com

Author contributions

Conceptualization: Hao Huang, Yujun Hu; Data curation: Xinqiang Li, Bowen Zhang; Formal analysis: Hao Huang; Investigation: Hao Huang, Yiwei Huang, Anxue Li; Methodology: Hao Huang, Wenxuan Duan; Project administration: Yujun Hu; Resources: Anxue Li; Software: Yujun Hu; Supervision: Yujun Hu; Validation: Xinqiang Li, Bowen Zhang, Yang Li; Visualization: Yiwei Huang, Bowen Zhang; Writing – original draft: Hao Huang; Writing – review & editing: Yujun Hu, Wenxuan Duan, Yang Li.

Unraveling Green Innovation Drivers in High Carbon-Lock-in Process Industries: An Interpretable Machine Learning Approach with Temporal Validation

Abstract

The petrochemical industry, characterized by severe carbon lock-in and high asset specificity, is one of the most challenging sectors for green transition. Existing studies predominantly rely on linear econometric models that struggle to capture nonlinear synergies and lack industry-specific evidence. Using 380 firm-year observations from Chinese A-share listed petrochemical enterprises (2014–2024), we construct a five-dimensional driver system with 20 indicators grounded in institutional theory and the resource-based view, employing an interpretable XGBoost–SHAP framework. This study makes three key contributions: (1) revealing the conditional boundary of institutional pressure under high asset specificity, showing that only command-and-control regulation forms positive synergies with internal resources; (2) adopting expanding-window rolling validation instead of random partitioning to test temporal generalization, and introducing Kendall's τ for the stability of SHAP feature importance rankings; and (3) uncovering single-threshold patterns across all core drivers in a high carbon-lock-in industry, providing micro-level evidence for optimizing regulatory instrument mixes. The model achieves a three-year average out-of-sample R^2 of 0.8352. Internal resource endowments—particularly firm size and R&D investment—dominate green innovation performance, while command-and-control regulation outperforms market-based instruments. These findings offer actionable insights for advancing the low-carbon transition of high-emission process industries, aligned with SDG 9 and SDG 13.

Keywords: green innovation · high carbon lock-in · petrochemical industry · XGBoost-SHAP · environmental regulation · threshold effect

1. Introduction

The petrochemical industry serves as a foundational pillar of China's national economy, yet its high energy consumption and emission intensity have long placed it at the core of environmental governance and the pursuit of carbon peak and carbon neutrality goals [1]. Owing to the indivisibility and long-cycle operation of core production installations, petrochemical

enterprises face extremely high switching costs for short-term process substitution, so their green technology innovation decisions depend heavily on external institutional pressures and internal resource endowments. Focusing on this typical high-carbon lock-in sector, this study systematically identifies the multidimensional driving mechanisms of green innovation in the petrochemical industry.

The driving mechanisms of corporate green innovation have long been interpreted through the dual lenses of institutional theory and the resource-based view (RBV). Wen et al.[2] verified the association between institutional pressure and green management practices in Chinese chemical firms, while Choi et al.[3] analyzed how Korean firms leverage internal capabilities to enhance eco-innovation. Institutional theory posits that organizational actions are shaped by regulative, normative, and cultural-cognitive pillars [4,5]. For petrochemical enterprises, institutional pressures manifest as command-and-control instruments (emission standards, energy consumption quotas) and market-based instruments (environmental taxes, emissions trading schemes), which exert markedly different effects on green innovation [6]. The RBV emphasizes that only resources possessing the VRIN attributes can form the foundation of sustained competitive advantage [7,8].

Importantly, these two theoretical pathways are not mutually exclusive. Petrochemical enterprises, characterized by high fixed asset specificity and substantial transformation costs, find it difficult to evade regulatory constraints through asset transfer, thereby making green technology innovation a more likely strategic response. Firms' technological absorptive capacity positively moderates the impact of regulatory pressure on green patent output [9], implying that external regulation can only be effectively transformed into green patent output when enterprises possess sufficient R&D human resources and prior technological accumulation.

Although the integration of institutional theory and the resource-based view provides a powerful analytical paradigm, the existing literature still exhibits notable limitations in variable coverage. From the environmental regulation research stream, studies employ single policy instruments as quasi-natural experiments: Wang et al.[10] examined command-and-control regulation using the new Environmental Protection Law as an exogenous shock; Yu et al.[11] explored the mechanisms of open environmental regulation; and Yu and Liu[12] analyzed the trade-off between green and digital innovation under the low-carbon city pilot policy. From the

corporate governance research stream, studies have separately examined board diversity and ownership structure [13], CEO age [14], and configurational combinations of board capital, CEO power and ownership structure [15]. From the internal resource research stream, some studies focus exclusively on endogenous capability variables such as R&D intensity, financial status and digitalization [16,17]. These three research streams each have their own emphases, yet all struggle to compare the relative importance of multiple internal and external factors within a unified framework. The multidimensional analytical framework constructed in this study is precisely designed to address this gap.

In terms of research methodology, existing studies predominantly employ traditional econometric regression models, with panel fixed-effects models and system GMM being the most representative methods [18,19]. However, their validity relies heavily on the assumption of linear relationships and the researcher's a priori specification of model structure, while the driving mechanism of green innovation involves the nonlinear superposition of multiple factors. For instance, Wang and Si[20] found an inverted U-shaped relationship between R&D intensity and green innovation performance, and the innovation incentive of environmental taxation likewise follows an inverted U-shaped pattern [21]. Traditional econometric models thus exhibit obvious limitations in capturing complex nonlinear relationships and higher-order interaction effects [22].

In recent years, machine learning methods have provided a new technical pathway for overcoming these limitations. Built on the gradient boosting tree framework, XGBoost is selected as the baseline prediction model because it naturally accommodates the zero-inflated, right-skewed distribution of green patent data [23]. The SHAP interpretability method, grounded in the Shapley value from cooperative game theory, decomposes the marginal contribution of each feature in an additive manner [24]. This prediction-interpretation dual capability has been preliminarily applied in green innovation research: Wang et al.[25] employed the XGBoost-SHAP framework to identify the green innovation performance of heavily polluting enterprises; Zhao et al.[26] revealed that firm size, R&D investment and technology utilization capability rank as the top three factors; and Wang et al.[27] applied multiple machine learning algorithms and SHAP analysis to reveal the nonlinear impacts of institutional elements on corporate green innovation. However, these studies leave room for improvement in industry focus, rigor of

temporal validation, and completeness of robustness testing. This study is precisely designed to address these gaps.

In summary, although existing research has made significant progress, two critical gaps remain. First, most studies focus on heavily polluting industries at the aggregate level or the manufacturing sector as a whole, without extending the multidimensional analytical framework to process industries with severe carbon lock-in and high asset specificity, where green innovation follows fundamentally different logic. Second, existing machine learning studies largely adopt random partitioning or cross-validation, lacking rigorous testing of cross-temporal generalization ability. To address these gaps, this study makes three key contributions. First, from a theoretical perspective, it reveals the conditional boundary of institutional pressure under high asset specificity: institutional pressure can be transformed into innovation output only when it has sufficient binding force and enterprises cross the resource threshold. Second, from a methodological perspective, it adopts expanding window temporal validation, combines multi-model Diebold-Mariano tests and parameter sensitivity analyses, and introduces Kendall's tau coefficient to test the stability of SHAP feature importance rankings. Third, from an empirical perspective, it uncovers single-threshold patterns across all core drivers and instrument heterogeneity in regulation-resource synergies in the petrochemical industry. Fig. 1 presents the technical roadmap of this study.

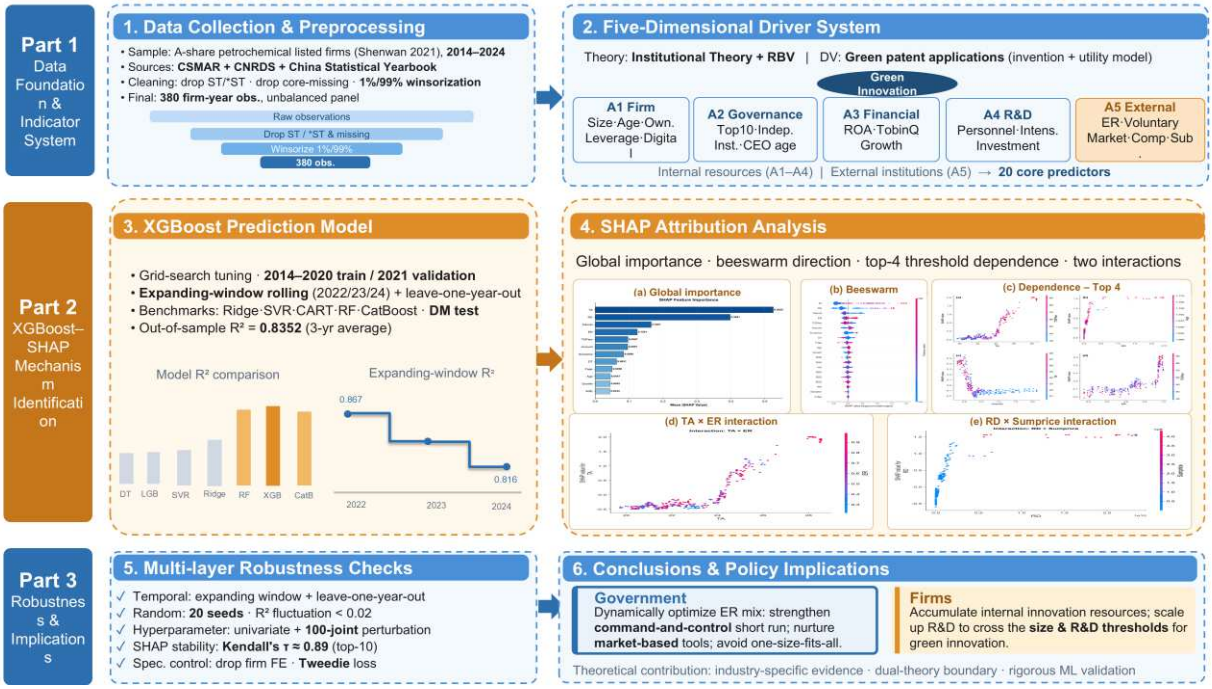


Fig. 1 Technical Roadmap

The remainder of this paper is organized as follows: Section 2 describes the research design; Section 3 reports the empirical results and SHAP analysis; Section 4 discusses the findings and policy implications; and the final section concludes.

2. Research Design

2.1 Sample and Data Sources

This study takes A-share listed petrochemical enterprises in China from 2014 to 2024 as the initial research sample, with industry classification referencing the 2021 version of the Shenwan Industry Classification Standard. To ensure data quality and eliminate confounding factors, a multi-step sample cleaning procedure is applied:

First, special treatment (ST) and *ST firms with abnormal financial conditions are excluded, as their operational instability may distort the relationship between R&D investment and green innovation output in the capital-intensive petrochemical sector.

Second, firm-year observations with excessive missing values in core variables are dropped to maintain the reliability of model estimation.

Third, all continuous variables are winsorized bilaterally at the 1st and 99th percentiles to mitigate the interference of extreme outliers caused by the wide asset gap between industry leaders and small firms.

After data cleaning, the final valid sample consists of 380 firm-year observations, forming an unbalanced panel dataset. Firm-level financial and governance data are mainly sourced from the CSMAR and CNRDS databases, while urban-level environmental regulation data are collected from the China Statistical Yearbook. To control for unobservable macro policy shocks and time-invariant firm heterogeneity, year fixed effects and firm fixed effects are incorporated into the baseline model, and a control test excluding firm fixed effects is designed in the robustness section to examine the stability of conclusions.

2.2 Variable Selection and Indicator System

Grounded in the dual perspectives of external institutional constraints and internal resource endowments, and in view of the high asset specificity and carbon lock-in attributes of petrochemical enterprises, this study constructs a driving factor indicator system by referring to the analytical framework developed by Wang et al.[25] for heavily polluting enterprises. The system covers five dimensions and 20 specific indicators, with the theoretical logic aligned with institutional theory (external environment dimension) and the RBV (firm characteristics, governance, finance and R&D dimensions).

2.2.1 Dependent Variable

The number of green patent applications filed by enterprises is selected as the proxy for green innovation performance, measured by the sum of green invention patent applications and green utility model patent applications. This indicator can intuitively reflect the output level of green technology R&D activities of enterprises, and is widely used in existing green innovation research. Relevant patent data are obtained from the CSMAR database.

2.2.2 Explanatory Variables

Five categories of driving factors are set as core explanatory variables, covering both internal firm characteristics and external environmental factors:

Firm characteristics (A1): Firm size, firm age, ownership nature, debt-to-asset ratio, and digitalization level, which capture the basic resource endowment and operational attributes of enterprises.

Corporate governance (A2): Top ten shareholder ownership, independent director ratio, institutional investor ownership, and CEO age, which reflect the governance structure and decision-making characteristics of firms.

Financial performance (A3): Return on assets, Tobin's Q, and firm growth, which portray the profitability, market valuation and development potential of enterprises.

Corporate R&D (A4): R&D personnel ratio, R&D intensity, and R&D investment, which measure the scale and intensity of enterprises' innovation input.

External environment (A5): Command-and-control environmental regulation, voluntary environmental regulation, market-based environmental regulation, market competition intensity, and government subsidies, which depict the external institutional pressure and policy support faced by firms.

It should be noted that variables including CEO duality, executives' political background and executives' academic background show zero within-sample variation in the petrochemical listed firm sample, and thus are excluded from the formal modeling to avoid invalid estimation interference. Detailed variable definitions and data sources are presented in Table 1. Extended variable definitions and the exclusion rationale are provided in the Supplementary Information.

Table 1 Indicator system of factors influencing corporate green innovation performance and data sources

Dimension	Indicator Name	Code	Definition	Data Source
Firm Characteristics (A1)	Firm Size	TA	Natural logarithm of total assets	CSMAR Database
	Firm Age	Fage	Natural logarithm of years since the firm's establishment	CSMAR Database
	Firm Ownership	SOE	1 for state-owned enterprises; 0 otherwise	CSMAR Database
	Asset-Liability Ratio	Debt	Total liabilities / total assets	CSMAR Database

	Digitalization Level	DT	Proportion of digital technology-related items in intangible assets, as disclosed in the notes to year-end financial statements	CNRDS Database
Corporate Governance (A2)	Top 10 Shareholders	TopTen	Sum of shareholding ratios of the top ten shareholders	CSMAR Database
	Independent Director Ratio	Indep	Ratio of the number of independent directors to total board size	CSMAR Database
	Institutional Shareholding	Inst	Proportion of shares held by institutional investors in the listed firm	CSMAR Database
	CEO Age	Age	Natural logarithm of the CEO's age	CSMAR Database
Financial Performance (A3)	Return on Assets	ROA	Net profit / total assets	CSMAR Database
	Market Value	TobinQ	Natural logarithm of Tobin's Q	CSMAR Database
	Firm Growth	Growth	Growth rate of sales revenue	CSMAR Database
Corporate R&D (A4)	R&D Personnel Ratio	RDP	Proportion of R&D personnel in total employees	CSMAR Database
	R&D Intensity	RDS	Ratio of R&D investment to total operating revenue	CSMAR Database
	R&D Investment	RD	R&D expenditure	CSMAR Database
External Environment (A5)	Command-and-Control Environmental Regulation	ER	Composite index of urban pollutant emissions	China Statistical Yearbook
	Voluntary Environmental Regulation	ISO14001	1 if the firm obtained ISO 14001 certification in the year; 0 otherwise	CSMAR Database
	Market-Based Environmental Regulation	Sumprice	Pollutant discharge fee before 2018; environmental protection tax after 2018	CSMAR Database
	Market Competition Intensity	HHI	Sum of squared market shares of firms in the industry	CSMAR Database
	Government Subsidy	Amount	Natural logarithm of total government subsidies received by the firm in the current period, as	CSMAR Database

*Note: All continuous variables are winsorized at the 1st and 99th percentiles to mitigate the influence of extreme outliers.

2.3 Model Specification: XGBoost–SHAP Analytical Framework

Traditional econometric models require pre-specification of linear functional forms, which are difficult to adapt to the complex driving mechanism of green innovation in petrochemical enterprises characterized by nonlinear superposition of multiple factors and high-order interaction effects. To address this limitation, this study constructs a two-step "prediction–interpretation" analytical framework. First, the XGBoost algorithm is applied to build a nonlinear prediction model of green innovation performance, to fully fit the complex functional relationships among multidimensional variables. Second, the SHAP additive interpretability model is introduced to conduct feature-level attribution of the black-box prediction results, so as to realize quantitative identification of driving mechanisms while ensuring prediction accuracy and economic interpretability.

2.3.1 Prediction Model: XGBoost

Proposed by Chen and Guestrin[23], XGBoost is an efficient implementation of the gradient boosting tree ensemble algorithm. Its core logic is to iteratively generate multiple classification and regression trees (CART), sequentially fit the residuals of the previous model, and finally integrate the outputs of all weak learners to form a high-precision strong learner. It has become one of the mainstream algorithms for structured data prediction.

The optimization objective of XGBoost consists of two components: a fitting loss term and a regularization term. The objective function is specified as follows:

$$\mathrm{Obj}(\theta) = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega(f_k)$$

Where $L\left(y_i, \hat{y}_i\right)$ denotes the loss function, which measures the fitting error between the predicted value $\hat{y}_i$ of the i-th sample and the actual number of green patent applications y_i. Given that green patent applications are non-negative count data with prominent zero-inflation and right-skewed distribution, the baseline model adopts the squared error loss function, and a

log(1+x) transformation is applied to the dependent variable to mitigate skewness. The Tweedie distribution loss function is introduced for a control test to verify the sensitivity of results to loss function specification.

The term $\sum_{k=1}^{K}\Omega(f_k)$ denotes the regularization term, which constrains model complexity and prevents overfitting. Its specific formulation is:

$$\Omega(f)=\gamma T+\frac{1}{2}\lambda\|w\|^2$$

where T is the number of leaf nodes in a single decision tree, w denotes the weight vector of leaf nodes, γ is the penalty coefficient for leaf node generation, and λ is the L2 regularization coefficient for leaf weights. With such regularization constraints, XGBoost effectively improves the out-of-sample generalization performance of the model.

The applicability of XGBoost to this study is reflected in three aspects. First, it requires no pre-specified functional form among variables and can automatically capture nonlinear effects and high-order interaction effects, which matches the complex driving characteristics of green innovation in petrochemical enterprises. Second, it exhibits strong robustness to missing values and outliers, and can automatically learn the optimal split direction for missing values via the sparsity-aware algorithm, which fits the data features of unbalanced panels of listed firms. Third, its parallel computing mechanism significantly improves training efficiency for high-dimensional samples, well adapting to the sample scale of this study with 20 driving variables and multi-year panel data.

In terms of model training and validation, the grid search method is adopted for hyperparameter tuning. Samples from 2014 to 2020 are used as the training set and samples from 2021 as the validation set, and the optimal hyperparameter combination is determined by taking the root mean squared logarithmic error (RMSLE) as the optimization objective. To avoid temporal data leakage and realistically simulate the policy scenario of "predicting future innovation performance using historical data", expanding window rolling validation and leave-one-year-out cross-validation are employed for out-of-sample testing, replacing the traditional random cross-validation scheme. Meanwhile, seven baseline models including linear regression, random forest, and support vector machine are selected for comparative analysis, and the Diebold-Mariano (DM) test is applied to verify the performance advantage of XGBoost.

2.3.2 Interpretability Method: SHAP

The inherent black-box nature of machine learning models restricts their output to prediction results alone, without providing quantitative insights into the mechanisms underlying variable effects, which cannot meet the demands of mechanistic analysis in academic research. To address this limitation, this study adopts the SHAP model proposed by Lundberg and Lee[24]. Grounded in Shapley value theory from cooperative game theory, the model conducts feature-wise attribution of XGBoost predictions by assigning each variable its marginal contribution to the final output, thus unlocking the interpretability of the otherwise black-box model.

The core principle of the SHAP model is to decompose the final predicted value of an individual sample into the sum of a base value and the SHAP values of all features, which satisfies the local additivity axiom. The specific formulation is:

$$\hat{y}_i = \Phi_0 + \sum_{j=1}^{m} \Phi_{i,j}$$

where $\hat{y}_i$ denotes the predicted green innovation performance of the i-th firm sample; Φ_0 is the base value, defined as the model prediction when all features take their sample mean values, representing the average prediction level across the full sample; $\Phi_{i,j}$ is the SHAP value of the j-th feature for the i-th sample, capturing the feature's marginal contribution to the sample's prediction. A positive $\Phi_{i,j}$ indicates that the feature elevates the predicted green innovation level and exerts a positive driving effect, while a negative value indicates that the feature reduces the predicted value and has a suppressive effect.

Compared with the native gain-based feature importance metric of XGBoost, SHAP values are built on a rigorous game-theoretic foundation and satisfy three core axioms: local accuracy, consistency, and missingness. Specifically, the consistency axiom guarantees that if a change in model structure increases the marginal contribution of a given feature, its corresponding SHAP value will not decrease. This property theoretically eliminates the inherent bias present in traditional feature importance measures, making SHAP more appropriate for identifying driving factors and inferring mechanisms in academic research. In this study, SHAP values are computed using TreeExplainer, an interpreter optimized for tree-based models. This method produces exact

analytical solutions rather than approximate estimates, delivering substantially higher attribution accuracy than general-purpose explainers.

Subsequent empirical analysis will apply SHAP from three perspectives: global feature importance ranking to identify core driving factors, SHAP dependence plots to explore nonlinear effect patterns, and interaction effect analysis to examine the synergistic mechanism between institutional pressure and internal resources.

2.4 Robustness Test Strategy

A multi-pronged robustness testing framework is designed from two dimensions: prediction performance robustness and interpretation conclusion robustness.

Robustness tests are designed in five dimensions: (1) temporal generalization, using expanding window rolling validation and leave-one-year-out cross-validation to examine the stability of out-of-sample prediction performance across different years;

(2) random perturbation, repeating model training with 20 distinct random seeds to rule out the interference of random initialization; (3) hyperparameter sensitivity, adjusting core hyperparameters via univariate and joint perturbation to assess whether the findings depend on specific parameter settings; (4) SHAP interpretation stability, using Kendall's tau to measure the consistency of SHAP feature importance rankings under different hyperparameter settings and sample partitions;

(5) supplementary control tests, including a control test excluding firm fixed effects and a test replacing the squared error loss with the Tweedie distribution loss.

SHAP interpretation stability test: the Kendall rank correlation coefficient (Kendall's tau) measures the consistency of SHAP feature importance rankings under different hyperparameter settings and sample partitions, addressing the common neglect of SHAP robustness in existing literature.

Supplementary control tests: a control experiment excluding firm fixed effects and a test replacing the squared error loss with the Tweedie distribution loss are conducted to confirm that conclusions are not confounded by model specification choices.

2.5 Overview of Analytical Strategy

This paper follows a five-step progressive analytical path: prediction performance validation, global feature importance analysis, nonlinear effect identification, interaction effect analysis, and micro-level heterogeneity discussion.

No strict causal assumptions are pre-specified: the feature system is constructed solely based on the integrated framework of institutional theory and the resource-based view, and all findings reflect correlational pattern recognition rather than strict causal inferences.

3. Empirical Results and SHAP Analysis

This chapter reports the full empirical results of the XGBoost-SHAP framework, covering descriptive statistics, prediction performance, SHAP attribution, interaction mechanisms, and robustness tests.

3.1 Descriptive Statistics

Consistent with the variable definitions in Table 1, this section verifies sample representativeness and data quality through the distribution characteristics of all core variables.

3.1.1 Dependent Variable

The dependent variable is the number of green patent applications, measuring the green innovation performance of sample enterprises, with a mean of 41.29, a median of only 2, a standard deviation of 161.70, a maximum of 1236, and a zero-value proportion of 36.84%.

The distribution presents significant zero-inflation and extreme right-skew: more than one-third of firms generate no green patents in a given year, while leading enterprises can achieve annual outputs exceeding one thousand, justifying the adoption of the nonlinear XGBoost model.

3.1.2 Core Explanatory Variables

(1) Internal Resource Dimension

After natural logarithmic transformation, total assets (TA) have a mean of 23.04, a median of 22.70, a standard deviation of 2.11, ranging from 19.13 to 28.65, reflecting an extremely wide span of firm size.

For R&D investment (RD), the mean is approximately RMB 1.426 billion while the median is only RMB 87 million, with a standard deviation of RMB 4.635 billion and a maximum of

RMB 31.115 billion, showing pronounced right-skew. The mean R&D intensity (RDS) is 2.61% and the mean R&D personnel ratio (RDP) is 12.18%.

Tobin's Q has a mean of 0.48, a median of 0.31 and a maximum of 2.46. The overall low valuation is consistent with the value stock attribute of the capital-intensive petrochemical industry.

(2) External Institutional Dimension

The mean value of command-and-control environmental regulation (ER) is 0.666 with a standard deviation of 0.258, ranging from 0.14 to 1, reflecting significant differences in regulation intensity across regions and years.

For market-based environmental regulation (Sumprice), the mean is approximately RMB 68.69 million while the median is only RMB 2.1589 million, with a standard deviation of RMB 240 million, presenting a strong right-skewed distribution.

The proportion of zero values for voluntary environmental regulation (ISO14001) reaches 78.95%, meaning only about 21% of sample firms have obtained ISO 14001 certification.

3.1.3 Control Variables

The mean of firm ownership (SOE) is 0.408, indicating state-owned enterprises account for 40.79% of the sample. The mean asset-liability ratio (Debt) is 48.65% with a median of 50.00%.

For the corporate governance dimension, the average shareholding ratio of the top ten shareholders is 60.08%, reflecting high ownership concentration.

For digitalization level (DT), the proportion of zero values is 28.42%. The mean market competition intensity (HHI) is 0.325 with a standard deviation of only 0.07. After logarithmic transformation, the mean of government subsidies (Amount) is 16.99.

Overall, the value ranges of all variables are consistent with the industrial reality of the petrochemical industry, providing a reliable data foundation for subsequent XGBoost modeling and SHAP attribution analysis.

3.2 Model Prediction Performance and Comparison

Before conducting mechanism identification, this section evaluates the prediction performance of the XGBoost model to verify its applicability to the petrochemical green

innovation sample, through expanding window rolling validation for temporal generalization ability and multi-model horizontal comparison for relative performance advantage.

3.2.1 Results of Expanding Window Rolling Validation

To avoid temporal data leakage and follow the standard logic of forecasting next-period outcomes with historical data, expanding window rolling validation is adopted for out-of-sample testing.

In the hyperparameter tuning stage, the 2014-2020 sample serves as the training set and the 2021 sample as the validation set, with the optimal hyperparameter combination determined via grid search (Supplementary Table S1).

In the rolling extrapolation stage, the training window expands progressively: the model is trained on 2014-2021 to predict 2022, refitted on 2014-2022 to predict 2023, and further expanded to 2023 to produce forecasts for 2024.

Table 2 XGBoost Rolling Prediction Results

Year	R^2	RMSLE	MAE	Spearman ρ
2022	0.8669	0.6627	18.22	0.8274
2023	0.8229	0.7311	10.11	0.8365
2024	0.8157	0.623	10.66	0.8822
Average	0.8352	0.6723	13.00	0.8487

As reported in Table 2, the three-year average out-of-sample R2 reaches 0.8352, explaining more than 80% of the variation in green innovation performance, with an average RMSLE of 0.6723.

Year by year, the out-of-sample R2 declines slightly from 0.8669 in 2022 to 0.8157 in 2024 but remains above 0.8; the mean absolute error narrows from 18.22 to 10.66, and the Spearman rank correlation coefficient rises steadily to 0.8822 in 2024.

Overall, the XGBoost model delivers stable intertemporal extrapolation performance, and an additional leave-one-year-out cross-validation yields largely consistent results.

3.2.2 Multi-model Comparison and Diebold-Mariano Test

To verify the relative applicability of XGBoost, six mainstream models (Ridge regression, SVR, single decision tree, Random Forest, LightGBM and CatBoost) are compared under the identical training-validation-test split and unified grid search tuning.

Table 3 Performance comparison of multiple models

Model Name	R^2	RMSLE	MAE	Spearman ρ
RandomForest	0.8398	0.6693	18.67	0.8698
XGBoost	0.8352	0.6723	13.00	0.8487
CatBoost	0.8276	0.6865	19.54	0.8667
Ridge	0.8021	0.7424	27.56	0.8312
SVR	0.7986	0.7483	26.07	0.8474
LightGBM	0.7986	0.7286	11.89	0.8542
DecisionTree	0.6960	0.9130	25.26	0.7758

As shown in Table 3, prediction performance presents a clear tiered pattern: three tree-based ensemble models (Random Forest, XGBoost and CatBoost) deliver the best overall performance with out-of-sample R2 all exceeding 0.82; Ridge regression, SVR and LightGBM rank second with R2 concentrated in 0.79-0.80; the single decision tree has the weakest fitting performance.

Within the first tier, Random Forest achieves the highest R2 and Spearman rank correlation coefficient, XGBoost has the smallest mean absolute error of 13.00, and CatBoost falls in the mid-level.

To further confirm the statistical significance of inter-group performance differences, this study employs the Diebold-Mariano (DM) test based on the squared error loss function. Detailed DM test results are provided in the Supplementary Information.

The DM test results show that the differences in prediction errors between XGBoost and the other two first-tier models fail to pass the 5% significance level, indicating no statistically significant difference in prediction accuracy among the three tree-based ensemble models.

Considering prediction performance and the demand for subsequent attribution analysis, XGBoost is finally selected as the core model: its accuracy ranks in the first tier, and its TreeExplainer delivers better stability and efficiency in SHAP value calculation.

3.3 Global Feature Importance Analysis of SHAP

On the basis of confirming the reliability of model prediction, this section applies the SHAP method to conduct global attribution analysis, identifying the relative importance of each driving factor and its direction of impact on green innovation performance.

3.3.1 Feature Importance Ranking

Fig. 2 presents the ranking of SHAP global feature importance; the mean absolute SHAP value measures the overall magnitude of a feature's impact on model predictions. Only the top 12 high-contribution variables are displayed.

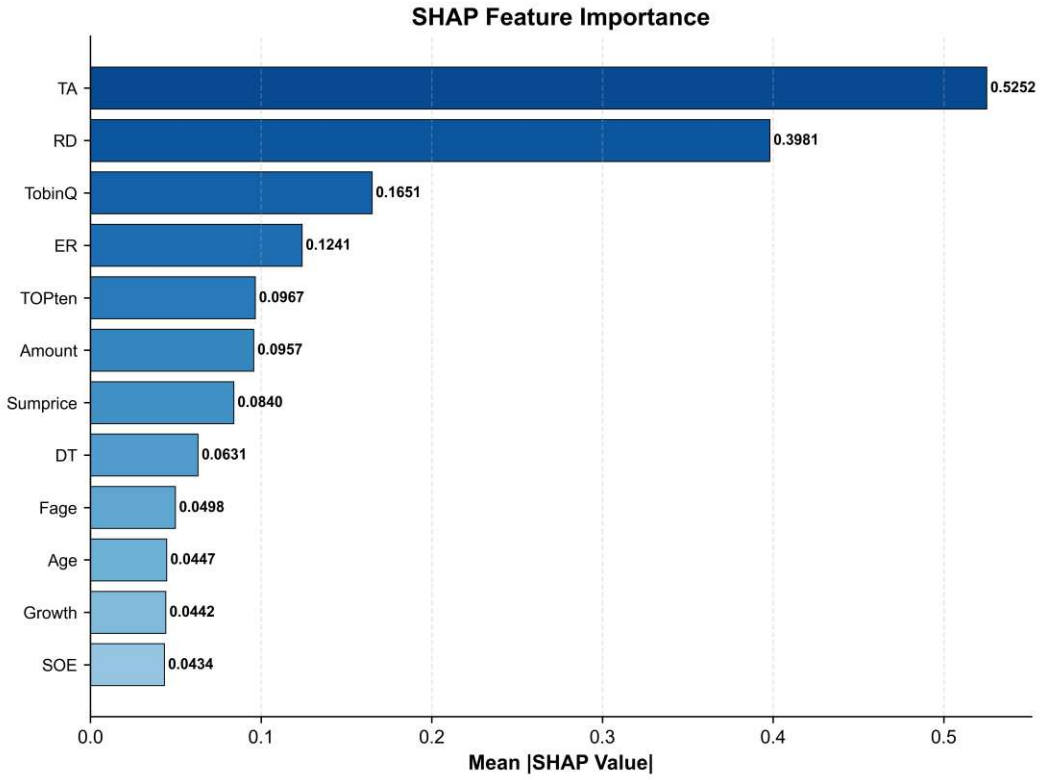


Fig. 2 Ranking of SHAP global feature importance

The impact intensity shows clear hierarchical differentiation. Total assets (TA) and R&D investment (RD) take a dominant lead, with mean absolute SHAP values of 0.5252 and 0.3981,

far exceeding all other variables; their combined contribution accounts for 47% based on all 20 features, and exceeds 53% among the top 12 features.

Market value (TobinQ) and command-and-control regulation (ER) rank 3rd and 4th with mean absolute SHAP values of 0.1651 and 0.1241. Variables ranking 5th to 8th are TopTen, subsidies (Amount), market-based regulation (Sumprice) and digitalization (DT), with values ranging from 0.06 to 0.1.

The remaining variables (firm age, CEO age, operating revenue growth rate, firm ownership, governance variables) all have mean absolute SHAP values below 0.05.

Aggregating by research dimension, the internal resource dimension accounts for approximately 77% of total contribution; the external environment dimension about 16%, and the corporate governance dimension approximately 7%.

Overall, the variation in green innovation among petrochemical enterprises is mainly determined by internal resource endowments, while the overall driving effect of external institutional pressure is relatively limited.

3.3.2 Feature Influence Direction and Distribution

Fig. 3 presents the SHAP beeswarm summary plot, reflecting the magnitude of feature values, their impact direction, and sample distribution characteristics: a positive SHAP value indicates the feature raises the predicted green innovation level, and vice versa.

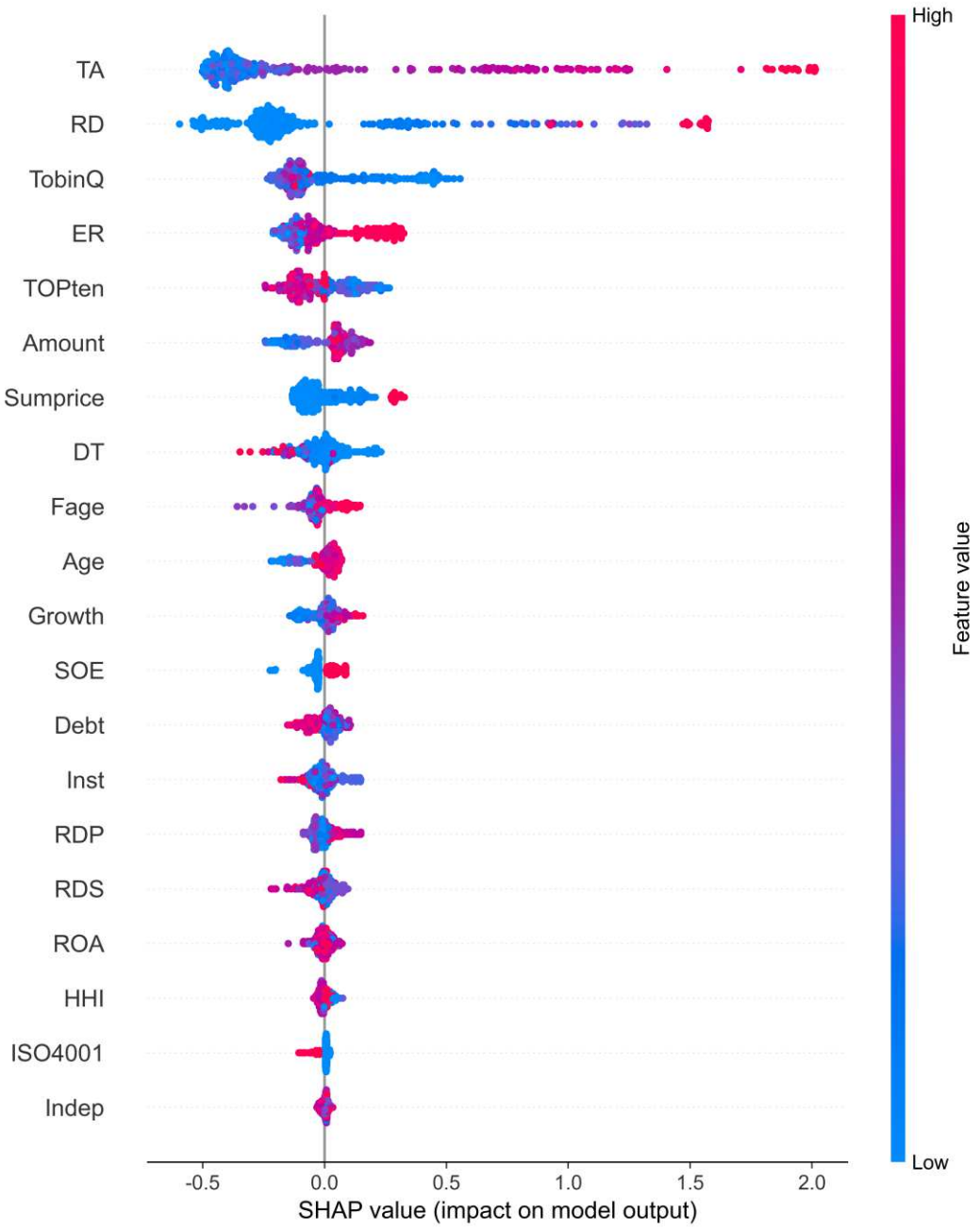


Fig. 3 SHAP beeswarm summary plot

For the top two core variables, TA and RD show an overall clear positive effect trend and are the only variables that can push the SHAP value of individual samples above 2.

Tobin's Q presents an obvious negative association: high-valuation samples are mostly concentrated in the negative SHAP interval while low-valuation samples correspond to higher positive values.

The effect of command-and-control regulation (ER) does not follow a uniform linear pattern: samples with high regulation intensity are scattered on both sides of zero, with most showing a positive innovation effect but some close to zero or negative.

Among the moderately important variables, subsidies (Amount) and market-based regulation (Sumprice) show an overall weak positive effect. High-value samples of TopTen mostly correspond to negative SHAP values, reflecting the inhibitory tendency of ownership concentration on green innovation.

For lower-ranked variables including firm age, CEO age, asset-liability ratio and corporate governance variables, scatter points are basically clustered around zero.

Overall, internal resource variables have clear effect directions and large impact magnitudes, serving as the main source of variation in green innovation performance; environmental regulation variables show significant firm heterogeneity and have not yet formed a unified linear driving effect.

3.4 Nonlinear Effect Analysis

To further explore the specific action pattern of core driving factors, this section constructs SHAP dependence plots (Fig. 4) for the top four core variables to identify their nonlinear effect characteristics and threshold effects.

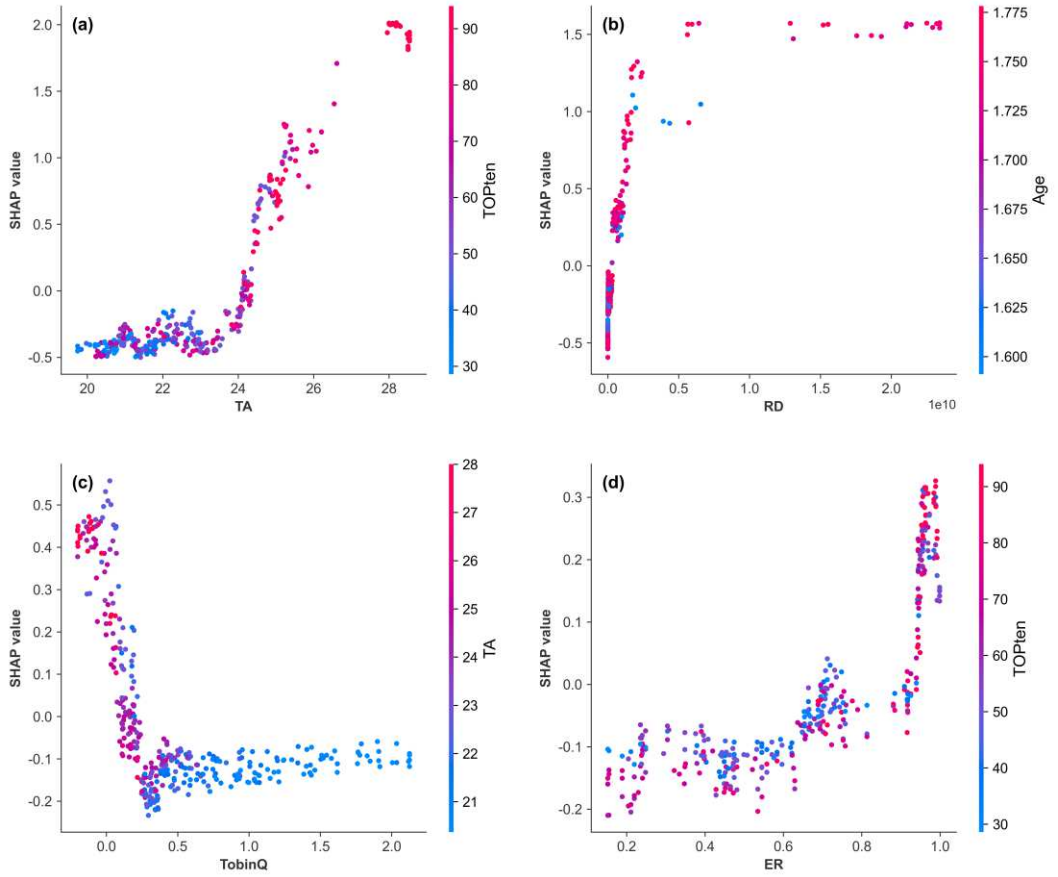


Fig. 4 SHAP dependence plots of core driving variables: (a) Total assets (TA); (b) Research and development investment (RD); (c) Tobin's Q; (d) Environmental regulation (ER)

3.4.1 Single-threshold Step Pattern of Firm Size

From the sample distribution of total assets (TA), the effect of firm size on green innovation does not increase linearly; instead, there exists a clear critical threshold. When log-transformed total assets are below 24, the SHAP values of samples mostly fall below zero, and scale expansion does not exert a significant positive pull effect on green innovation. Once total assets cross this threshold, SHAP values rise rapidly with increasing scale, and the upward trend persists through the maximum value interval of the sample without obvious marginal reversal.

This pattern is directly associated with the industrial characteristics of the petrochemical sector. Green technological transformation usually requires synchronous upgrading of core production facilities, with high upfront fixed investment and large sunk costs. Small and medium-sized enterprises lack sufficient scale to cover transformation costs, making it difficult to translate scale advantages into innovation output. Only firms that cross the scale threshold can dilute R&D and transformation costs via production capacity and build stable green innovation

capabilities, which also explains the structural cause of the bipolar differentiation of green innovation across the industry.

3.4.2 Input Threshold and Diminishing Marginal Returns of R&D Investment

The driving effect of R&D investment (RD) on green innovation also exhibits a clear input threshold. When R&D investment is at a low level, the corresponding SHAP values mostly fluctuate around zero, and some samples even show negative impacts. This indicates that scattered, low-level R&D investment can hardly be translated into actual green innovation outcomes. After breaking through the critical input scale, SHAP values rise rapidly with R&D investment, entering a stage of concentrated release of innovation output. In the highest R&D investment interval, however, the upward slope of the curve flattens slightly, showing signs of diminishing marginal returns.

This reflects the cumulative nature of green R&D in the petrochemical industry. Green technology breakthroughs require long-term technology reserves, team building and process verification. Scattered investment below the threshold cannot build a complete R&D system and is unlikely to produce substantive innovation results. When R&D scale reaches a relatively high level, constrained by the current green technology bottlenecks in the industry, the output efficiency of additional investment declines accordingly.

3.4.3 Negative Nonlinear Association of Market Value

Market value (TobinQ) shows an overall negative nonlinear association with green innovation: the lower a firm’s Tobin’s Q, the higher its corresponding SHAP value, meaning that low-valuation "value-type" enterprises generally have stronger green innovation capabilities. By interval, when Tobin’s Q is below 0.5, SHAP values rise rapidly with declining valuation, with a pronounced negative marginal effect. After Tobin’s Q exceeds 0.5, SHAP values stabilize in a slightly negative interval with significantly narrowed fluctuation.

This characteristic is directly related to the ownership structure of the petrochemical industry. Most low-valuation enterprises in the sector are state-owned backbone firms. Such enterprises have lower financing costs and their investment decisions are not constrained by short-term market value fluctuations, making them more willing to invest in long-cycle, high-risk green technology R&D. This explains the inverse pattern that lower valuation corresponds to higher green innovation performance.

3.4.4 Threshold-type Driving Effect of Command-and-Control Environmental Regulation

The effect pattern of command-and-control environmental regulation (ER) is not simply linear either, and also presents a critical threshold. When regulation intensity is in the low-to-medium range of 0.2 to 0.6, sample SHAP values mostly fluctuate below zero, and regulation has not yet formed stable green innovation incentives. When regulation intensity breaks through the critical value of around 0.7, SHAP values rise rapidly with increasing regulation intensity, showing a clear positive driving effect. Across the full sample interval, no obvious marginal decline is observed in the high regulation intensity segment, indicating that command-and-control regulation in the petrochemical industry is still in the stage where the innovation compensation effect dominates. Compliance costs have not yet offset the benefits brought by innovation, and there is still room for further improvement of policy intensity.

Overall, the core driving factors do not present a simple linear marginal effect, but all show a single-threshold characteristic. Only when the factor level crosses the critical value can it release a stable positive driving effect, which reflects the typical threshold attribute of green innovation activities in capital-intensive heavy chemical industries.

3.5 Interaction Effect Analysis

On the basis of single-factor nonlinear effects, this section further examines the interaction effects between the two types of environmental regulation and internal resource endowments, to verify whether external institutional pressure can form synergistic effects with internal resources.

3.5.1 Interaction Between Command-and-Control Regulation and Firm Size

Fig. 5 shows the SHAP interaction dependence plot between total assets (TA) and command-and-control environmental regulation (ER). The horizontal axis represents log-transformed total assets of firms, and the vertical axis represents the marginal SHAP value of TA. Scatter points colored from blue to red correspond to regulation intensity from low to high, which is used to identify the synergistic effect pattern of the two factors on green innovation performance.

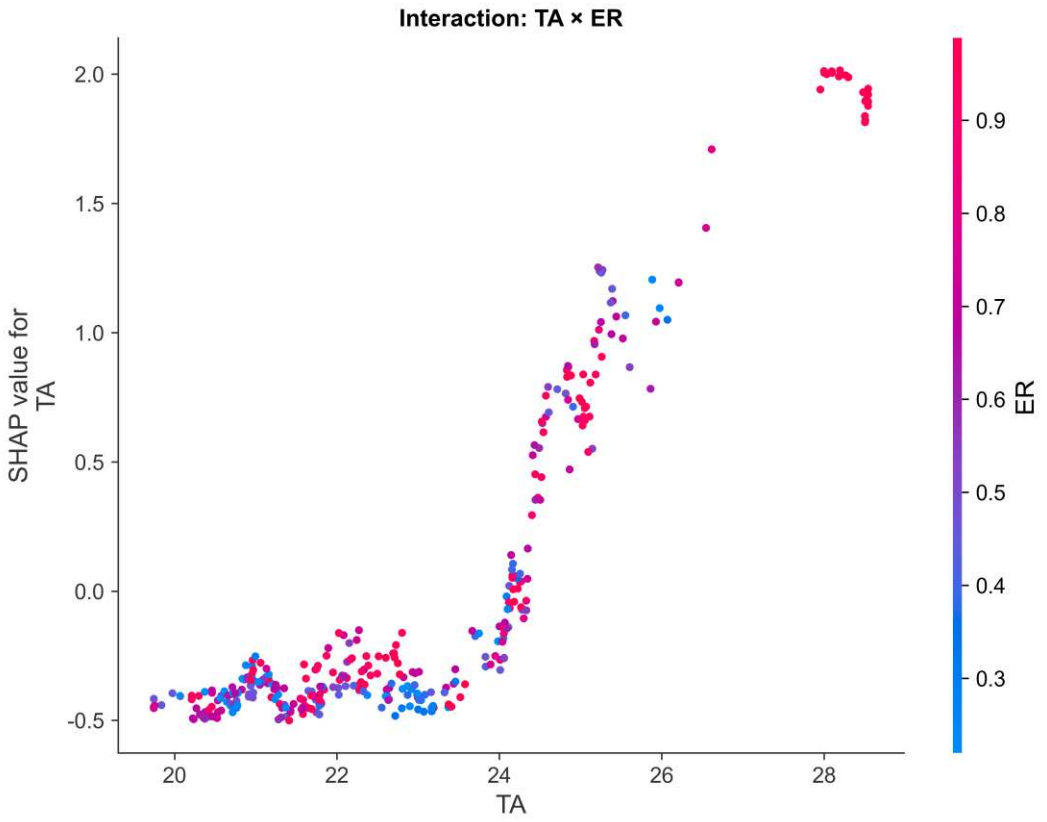


Fig. 5 SHAP interaction effect plot of TA and ER

There is an obvious critical threshold in the moderating effect of regulation on scale effect, echoing the stepwise characteristics identified in the previous univariate analysis. When log-transformed total assets are below 24, samples with different regulation intensities are interlaced. The SHAP values of TA generally fall below zero, and no clear stratification emerges between the high- and low-regulation groups. This indicates that when firm size has not reached the critical level, even with increased regulation intensity, scale expansion cannot drive green innovation, and regulatory pressure can hardly be translated into substantive innovation output.

Once total assets cross the critical threshold of 24, the color stratification of samples becomes immediately apparent. Under the same scale condition, the SHAP values of TA for samples with high regulation intensity are generally higher than those for samples with low regulation intensity. Moreover, the larger the firm size, the more obvious the gap in SHAP values between the two groups, and the synergistic strengthening effect of scale and regulation gradually amplifies. That is, the larger the firm size, the more prominent the marginal driving effect of command-and-control regulation on green innovation.

This result is consistent with the theoretical logic that institutional pressure exerts its effects through internal resource endowments. For small and medium-sized petrochemical enterprises, the sunk costs of green transformation of core facilities and low-carbon technology R&D are excessively high. Faced with regulatory constraints, they tend to adopt low-cost compliance approaches such as end-of-pipe treatment and phased production restrictions, and find it difficult to allocate resources to long-term green technology R&D. In contrast, leading enterprises that cross the scale threshold have sufficient capital, technology and talent reserves to transform compliance pressure into green R&D investment. Meanwhile, larger scale strengthens the ability to apportion transformation costs per unit of production capacity, leading to more significant innovation incentive effects brought by regulation.

3.5.2 Interaction Between Market-Based Regulation and R&D Investment

Fig. 6 presents the SHAP interaction effect between R&D investment (RD) and market-based environmental regulation (Sumprice). The horizontal axis represents the scale of corporate R&D investment, and the vertical axis represents the marginal SHAP value of RD. Scatter points colored from blue to red correspond to regulation intensity from low to high, which is used to judge whether market-based regulation can form synergistic innovation incentives with corporate R&D.

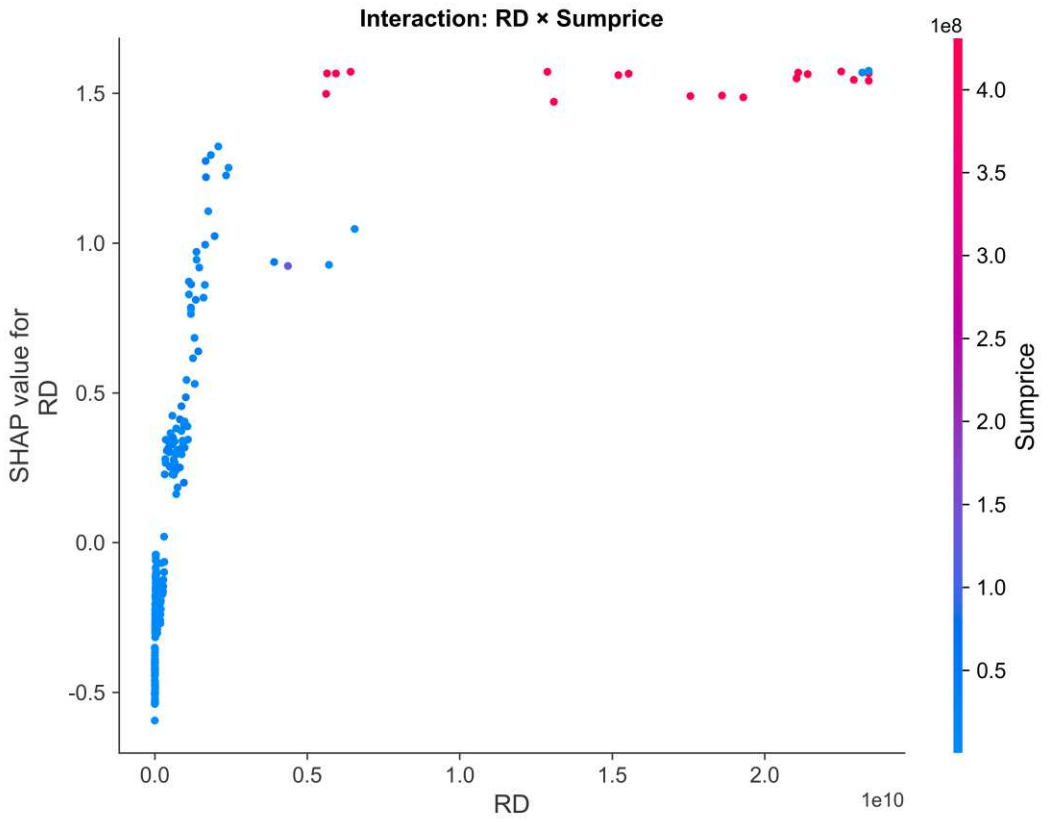


Fig. 6 SHAP interaction effect plot of RD and Sumprice

The threshold characteristic of R&D investment itself still exists: the SHAP values of samples in the low R&D interval basically fall below zero, and rise rapidly after crossing the input threshold to remain at a relatively high level. However, samples with different regulation intensities do not form a clear stratification pattern. In the low R&D interval, almost all samples are blue points with low regulation intensity, and no high-regulation samples are distributed. In the high R&D interval, samples with high and low regulation intensity are mixed together. Under the same R&D scale, the marginal contribution of RD in high-regulation samples is not systematically higher than that in low-regulation samples, and there is no synergistic strengthening feature as observed for command-and-control regulation in the previous section.

This indicates that current market-based environmental regulation has not yet formed effective positive synergy with corporate R&D investment, and the increase in regulation intensity has not significantly amplified the driving effect of R&D on green innovation. This result is directly related to the development stage of relevant policies. During the sample period, market-based tools mainly consisted of pollutant discharge fees and environmental protection

taxes [28], with overall low collection intensity and limited price guidance effect on corporate R&D decisions. In addition, the coverage of the national carbon market in the petrochemical industry [29] is still advancing, and long-term stable price expectations have not yet been formed. As a result, it is difficult to continuously stimulate corporate R&D investment, which naturally prevents the formation of a synergistic driving effect with internal R&D.

Overall, there are significant differences in the synergistic effects between the two types of environmental regulation and internal resources. Only command-and-control regulation with strong binding force can form a positive synergistic amplification effect with firm scale and R&D investment, while the incentive effect of market-based regulation is still limited by the current policy development stage.

3.6 Robustness Test Results

This section conducts an extensive robustness analysis covering prediction performance robustness, SHAP interpretation stability, and supplementary control tests.

3.6.1 Robustness Test of Prediction Performance

(1) Random perturbation robustness

To eliminate accidental deviation caused by random initialization, model training and prediction are repeated with 20 distinct random seeds; the standard deviation of out-of-sample R2 across the 20 experiments is only 0.0078. Univariate perturbation of eight core hyperparameters yields R2 fluctuations between 0.0076 and 0.0241, and joint perturbation with 100 random parameter combinations yields a standard deviation of only 0.0092, indicating that prediction performance is insensitive to hyperparameter settings and random initialization. Detailed results are provided in the Supplementary Information.

(2) Hyperparameter sensitivity analysis

(3) Model specification control test

Two control experiments are conducted by replacing core model specifications.

Firm fixed effects exclusion test: after removing individual firm fixed effects, the out-of-sample R2 reaches 0.8171, only slightly lower than the baseline. Tweedie distribution loss function test: replacing the squared error loss with the Tweedie loss raises the out-of-sample R2 to 0.8686, while the importance ranking, influence direction and nonlinear threshold characteristics of core variables remain consistent with the baseline model.

3.6.2 Stability Test of SHAP Interpretation Conclusions

Given that most existing machine learning studies overlook the stability of SHAP conclusions, Kendall's tau is adopted to measure the consistency of the importance ranking of the top 10 core features under different hyperparameter perturbations.

SHAP interpretation stability: under univariate perturbation of four core hyperparameters (max_depth, learning_rate, reg_alpha, reg_lambda), the Kendall's tau coefficients of the top 10 feature importance rankings are 0.9556, 0.9222, 0.9444 and 0.9556 respectively, all above 0.92; under joint perturbation, the average coefficient is 0.8904 and the minimum 0.8222. The importance ranking of core features and the threshold-type nonlinear patterns of TA and RD remain essentially unchanged. Detailed results are provided in the Supplementary Information.

Overall, the SHAP attribution conclusions are free from parameter-dependent contingency, and the identified driving mechanism pattern is highly stable.

3.6.3 Overall Evaluation of Robustness

Overall, the model maintains stable and high prediction performance under random perturbation, large-scale hyperparameter variation and model specification replacement, and the interpretation conclusions exhibit sufficient replicability.

4. Discussion

4.1 Interpretation of Core Empirical Findings

Grounded in institutional theory and the resource-based view (RBV), this study constructs a five-dimensional driver system covering 20 indicators, and adopts an interpretable XGBoost–SHAP framework with expanding window rolling validation to empirically examine the driving

mechanisms of green innovation performance for Chinese A-share listed petrochemical enterprises from 2014 to 2024. The core findings show that internal resource endowments dominate green innovation output in the sector, command-and-control environmental regulation exerts a significantly stronger incentive effect than market-based instruments, all core variables present a distinct single-threshold pattern, and only command-and-control regulation forms positive synergies with internal resources. These results not only align with the industrial attributes of high carbon lock-in and high asset specificity in the petrochemical industry, but also provide targeted empirical evidence for understanding the heterogeneous driving logic of green innovation in capital-intensive process industries.

First, the dominant role of internal resource endowments conforms to the core logic of the RBV, and presents more prominent industry-specific characteristics in the petrochemical sector. Global SHAP attribution results indicate that firm size and R&D investment rank as the top two drivers, with their combined contribution accounting for 47% of total feature contribution, and the overall share of the internal resource dimension reaching approximately 77%. This finding is partially consistent with Zhao et al.[26] and Wang et al.[25], who identified firm size and R&D investment as core drivers of green innovation among heavily polluting enterprises. However, the dominance of internal resources is far more pronounced in this petrochemical sample than in cross-industry samples of heavily polluting firms.

The underlying mechanism lies in the unique production attributes of the petrochemical industry: core production installations are indivisible and operate in long cycles, and green technological transformation requires synchronous upgrading of entire process systems, accompanied by extremely high upfront fixed investment and sunk costs. For petrochemical enterprises, green innovation is not a marginal adjustment of products or processes, but a systemic transformation that relies heavily on scale economies to dilute transformation costs and sustained R&D accumulation to achieve technological breakthroughs. This industrial reality directly determines that internal resource endowments become the fundamental constraint on green innovation output, which also explains the significant single-threshold effect of firm size and R&D investment identified in this study: only when resource scale crosses the critical threshold can enterprises bear the sunk costs of green transformation and translate input into substantive innovation output.

In addition, the negative nonlinear relationship between Tobin's Q and green innovation performance further reflects the unique ownership structure and innovation logic of China's petrochemical industry. Enterprises with lower market valuation, mostly state-owned backbone firms, generally show stronger green innovation capabilities: they bear dual responsibilities of industrial security and green transformation, have lower financing costs, and are less constrained by short-term capital market performance, making them more willing to deploy long-cycle, high-risk green technology R&D. This reveals the special innovation incentive structure of state-dominated high-carbon industries.

Second, the heterogeneous effects of different environmental regulation instruments reflect the stage characteristics of China's environmental governance system. This study finds that the innovation incentive effect of command-and-control environmental regulation is significantly stronger than that of market-based regulation, which differs from some cross-industry studies concluding that market-based instruments deliver more sustained innovation effects [6]. The divergence can be explained from two perspectives. On the one hand, for the petrochemical industry with clear emission boundaries and mature process standards, command-and-control instruments such as mandatory emission standards and energy consumption quotas have strong binding force and clear compliance benchmarks; under rigid regulatory constraints, enterprises can only achieve long-term compliance through green technological innovation, thus triggering the innovation compensation effect proposed by the Porter hypothesis [30]. On the other hand, during the sample period, China's market-based environmental instruments were still in the initial promotion stage for the petrochemical industry: the overall collection intensity of environmental protection tax [28] is relatively low, and the national carbon emissions trading market [29] has not yet fully covered the petrochemical sector, so the price signal is not strong enough to form stable long-term expectations for enterprises' R&D decisions.

Third, the single-threshold pattern and heterogeneous synergistic effects verify the conditional role of institutional pressure under high asset specificity (the “resource-gated transmission” logic) and deepen the integrated analytical framework of institutional theory and the RBV. Consistent with Kesidou et al.[9], which shows that technological absorptive capacity positively moderates the innovation effect of regulation, this study further reveals that in the petrochemical industry, regulatory pressure does not directly generate innovation output; it must first be absorbed by firms’ internal resource endowments and converted into innovation

investment, and this absorption operates only above critical resource thresholds. For firms below these thresholds, regulatory pressure is absorbed as compliance cost and channeled into low-cost reactive measures such as end-of-pipe treatment and output restrictions, yielding compliance effects without innovation effects, which explains why small and medium-sized enterprises below the scale threshold fail to generate substantive innovation incentives even under intensified regulation. Only when firms cross the scale and R&D thresholds do scale economies dilute transformation costs and complete R&D systems translate regulatory signals into technology directions, allowing the innovation offset effect to emerge. This also explains the observed instrument heterogeneity: only command-and-control regulation with strong binding force forms a positive synergistic amplification effect with firm size after crossing the scale threshold, whereas market-based regulation, constrained by weak price signals during the sample period, fails to stimulate enterprises to mobilize internal R&D resources for green innovation and thus cannot form a synergistic driving mechanism.

4.2 Theoretical Contributions

This study makes three key contributions to the green innovation literature, spanning theoretical, methodological, and empirical dimensions:

First, it advances the theoretical understanding of green innovation drivers by articulating a conditional boundary for the innovation effect of institutional pressure under high asset specificity, a “resource-gated transmission” mechanism. Rather than treating institutional pressure and internal resources as parallel driving forces, this study demonstrates that regulatory pressure does not directly generate innovation output; it must first be absorbed by firms’ internal resource endowments and converted into innovation investment, and this absorption operates only above critical resource thresholds. For firms below the scale and R&D thresholds, regulatory pressure is absorbed as compliance cost and channeled into low-cost reactive measures such as end-of-pipe treatment and output restrictions, yielding compliance effects without innovation effects. Only when firms cross these thresholds do scale economies dilute transformation costs and complete R&D systems translate regulatory signals into technology directions, allowing the innovation offset effect to emerge. This boundary logic is corroborated by three converging strands of evidence: the single-threshold patterns of firm size, R&D investment, and regulation intensity (Section 3.4); the interaction analysis showing that the

synergy between command-and-control regulation and firm size emerges only above the scale threshold (Section 3.5.1); and the instrument heterogeneity whereby market-based regulation, constrained by weak price signals during the sample period, fails to form such synergy (Section 3.5.2). This conditional-boundary perspective reconciles the conflicting findings in the literature on whether environmental regulation promotes or hampers green innovation: the net effect depends on where a firm sits relative to resource thresholds and on the binding force of the instrument, thereby offering a theoretical explanation for the heterogeneity documented in prior cross-industry studies.

Second, it improves the methodological rigor of machine learning empirical research in green innovation through two key innovations. Unlike most existing studies that adopt random cross-validation, this study introduces expanding window rolling validation to test the temporal generalization ability of the model, which better simulates the practical policy scenario of "predicting future green innovation trends with historical data" and avoids temporal data leakage. In addition, it applies Kendall's τ coefficient to test the stability of SHAP feature importance rankings across different training windows, addressing the critical limitation that existing machine learning studies often report SHAP results without verifying their robustness. These methodological improvements help reduce the contingency of machine learning empirical results and provide a reference paradigm for future research applying interpretable machine learning to environmental economics.

Third, it provides industry-specific empirical evidence for green innovation research by focusing on the petrochemical industry, a typical sector with severe carbon lock-in and high asset specificity. Most existing studies treat heavily polluting industries or manufacturing as a homogeneous whole, ignoring the substantial differences in transformation costs and innovation logic across sub-sectors. This study confirms that the driving mechanism of green innovation in process industries has clear particularities: internal resource dominance is more pronounced (accounting for approximately 77% of total contribution), command-and-control regulation effects are stronger than market-based instruments, and all core drivers exhibit single-threshold patterns. These findings refine the understanding of green innovation drivers in segmented industries and expand the applicable boundary of institutional theory and the resource-based view in high carbon-lock-in contexts.

4.3 Practical and Policy Implications

The findings carry clear guiding significance for optimizing environmental regulation policies and promoting the low-carbon transformation of petrochemical enterprises.

For government regulatory authorities, the instrument mix of environmental regulation should be dynamically optimized. In the short term, the constraint intensity of command-and-control regulation should be maintained and moderately improved by raising industry emission standards and energy consumption quota requirements. In the long term, market-based regulation systems should be accelerated, expanding carbon emissions trading coverage in the petrochemical industry, raising environmental protection tax standards, and forming stable long-term price signals. In addition, differentiated green innovation support policies should be implemented: special green innovation funds and industry-university-research cooperation platforms should be established to help small and medium-sized enterprises cross the scale and R&D input threshold, while leading enterprises should be incentivized to increase forward-looking green technology R&D.

For petrochemical enterprises, they should strengthen the accumulation of internal innovation resources by increasing R&D investment, expanding R&D teams, and breaking through the input threshold of green innovation, while promoting the deep integration of digital transformation and green technology R&D. They should also take the initiative to adapt to regulatory trends and transform compliance pressure into competitive advantages: leading enterprises should increase forward-looking green technology layout, while small and medium-sized enterprises can gradually improve their green innovation level through technology introduction and cooperative R&D.

4.4 Limitations and Future Research Directions

This study still has some limitations, which also point out directions for future research.

First, sample representativeness has certain constraints. This study only takes A-share listed petrochemical enterprises as samples, excluding a large number of non-listed small and medium-sized enterprises, which may lead to a certain overestimation of the overall innovation level of the industry.

Second, the measurement of green innovation performance can be further expanded. The number of green patent applications mainly reflects technological innovation output in the legal sense, but cannot fully cover process improvement, non-patented technologies and practical emission reduction effects.

Third, the causal identification of driving effects needs to be further strengthened. Based on machine learning methods, this study focuses on identifying correlation patterns and driving mechanisms; although temporal validation is adopted to avoid data leakage, strict causal inference as quasi-natural experiments do is still difficult to achieve.

Fourth, the dimension of driving factors can be further enriched. Future research can incorporate factors such as supply chain pressure, digital governance capability and executive environmental awareness into the analytical framework.All abbreviations used in this study are defined in Supplementary Table S2.

5. Conclusions

Against the backdrop of high carbon lock-in and high asset specificity in China's petrochemical industry, and in response to the limitations of existing studies on green innovation drivers (most of which rely on linear econometric models, fail to capture nonlinear synergistic effects, and lack targeted industry-specific evidence), this study takes unbalanced panel data of Chinese A-share listed petrochemical enterprises from 2014 to 2024 as the research sample. Grounded in institutional theory and the resource-based view, it constructs a driver system covering 20 indicators across five dimensions, and adopts a two-step "prediction–interpretation" interpretable machine learning framework of XGBoost–SHAP combined with an expanding window rolling validation scheme, to systematically identify the multidimensional driving mechanisms and action patterns of green innovation performance in petrochemical enterprises. The main conclusions are as follows.

First, the interpretable machine learning model achieves excellent fitting performance and sound temporal generalization ability for green innovation performance of petrochemical enterprises. Verified by expanding window rolling validation, the XGBoost model reaches a three-year average out-of-sample R^2 of 0.8352, which can stably explain more than 80% of the variation in green innovation performance. Compared with traditional linear econometric

models, it has notable advantages in capturing nonlinear effects and high-order interaction effects, providing reliable methodological support for identifying complex driving mechanisms. Through multi-dimensional robustness tests, both the model prediction performance and core attribution conclusions remain highly stable, indicating that the research results have good replicability.

Second, the green innovation performance of petrochemical enterprises presents a driving pattern of "internal resources as the dominant force, external institutions as the supplement". Internal resource endowments account for approximately 77% of the total contribution to the variation in green innovation. Among them, firm size and R&D investment are the top two core driving factors, with a combined contribution of nearly 50%, constituting the core support for industrial green innovation. The overall contribution of the external environment dimension is relatively limited, and the incentive effects of different regulatory instruments show obvious differentiation: the innovation-driven effect of command-and-control environmental regulation is significantly stronger than that of market-based environmental regulation, while voluntary environmental regulation represented by ISO 14001 certification has not yet formed substantive innovation incentives.

Third, core driving factors generally present a single-threshold effect, which is the structural root of the bipolar differentiation of green innovation across the industry. Firm size, R&D investment and command-and-control environmental regulation do not follow simple linear patterns, but exhibit typical threshold-driven characteristics: only when the variable level crosses the critical value can they release a stable and sustained positive driving effect. This feature is highly consistent with the industrial attributes of the petrochemical sector, such as heavy assets, long operation cycles and high sunk costs, and also explains the realistic pattern of marked gaps in green innovation levels between leading enterprises and small and medium-sized firms in the industry.

Fourth, there is significant instrument heterogeneity in the synergistic effects between environmental regulation and internal resources. Command-and-control environmental regulation can form a positive synergistic amplification effect with firm size, and the larger the firm size, the more prominent the innovation incentive effect of regulation. In contrast, current market-based environmental regulation has not yet formed effective synergy with enterprises'

internal R&D investment, and its price signal has not fully exerted a guiding role in enterprises' R&D decisions. This result verifies the theoretical logic that institutional pressure can only be transformed into innovation output on the basis of internal resource endowments, and also reflects that China's market-oriented environmental governance instruments are still in the stage of development and improvement for the petrochemical industry.

Focusing on the petrochemical industry, a typical high-carbon lock-in sector, this study supplements industry-specific heterogeneous evidence for research on green innovation driving mechanisms, deepens the integrated application of institutional theory and the resource-based view in the field of green innovation, and improves the conclusion reliability of machine learning empirical research through a rigorous temporal validation scheme. The findings can provide micro empirical support for government departments to optimize the portfolio of environmental regulation instruments and formulate differentiated green innovation support policies, and can also offer decision-making references for petrochemical enterprises to grasp regulatory trends and deploy green technology R&D. This study carries practical guiding significance for promoting the low-carbon transformation of the petrochemical industry and realizing industrial sustainable development.

References

State Council of the People's Republic of China. Action plan for carbon dioxide peaking before 2030 (Guo Fa [2021] No. 23) (2021).

Wen, X., Cheah, J. H., Lim, X. J. & Ramachandran, S. Why does "green" matter in supply chain management? Exploring institutional pressures, green practices, green innovation, and economic performance in the Chinese chemical sector. J. Clean. Prod. 427, 139182 (2023).

Choi, Y. C., Kim, T. & Lee, H. Impact of US and EU environmental regulations on eco-innovation in Korean battery firms: an analysis based on institutional theory and resource-based view. J. Clean. Prod. 521, 146217 (2025).

DiMaggio, P. J. & Powell, W. W. The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. Am. Sociol. Rev. 48, 147–160 (1983).

Scott, W. R. Institutions and Organizations (Sage, 1995).

Gebhardt, S., Hou, Y. & Yarime, M. Directing environmental innovation toward radical clean technologies for sustainable transitions: market-based vs. command-and-control policies. Res. Policy 55, 105446 (2026).

Barney, J. Firm resources and sustained competitive advantage. J. Manag. 17, 99–120 (1991).

Wernerfelt, B. A resource-based view of the firm. Strateg. Manag. J. 5, 171–180 (1984).

Kesidou, E., Krammer, S. M. S. & Wu, L. Subnational institutions, firm capabilities and eco-innovation. Ind. Corp. Change 33, 1460–1486 (2024).

Wang, S. Q., Liu, Z. Y., Tian, R. Z. & Jin, M. Environmental regulation, external attention and green innovation of heavily polluting enterprises: a difference-in-differences verification based on the implementation of the new Environmental Protection Law. China Environ. Sci. 46, 1164–1175 (2026a).

Yu, Z., Li, C. X. & Li, Y. K. How does open environmental regulation affect Chinese enterprises' green innovation? Syst. Eng. Theory Pract. 45, 2163–2187 (2025).

Yu, Y. & Liu, Y. Environmental regulations and firms' trade-offs between innovations: empirical evidence from the quasi-experiment in China. J. Environ. Manag. 370, 122819 (2024).

Ghazwani, M. The impact of board diversity and ownership on environmental innovation: the moderating role of CSR committees in G20 firms. Sustain. Dev. https://doi.org/10.1002/sd.70653 (2026).

Jiang, J., Tauni, Z. & Yousaf, U. B. The greyer the greener: CEO age and green innovation. Bus. Strategy Environ. 34, 7324–7348 (2025).

Shui, X., Zhang, M., Smart, P. & Ye, F. Sustainable corporate governance for environmental innovation: a configurational analysis on board capital, CEO power and ownership structure. J. Bus. Res. 149, 786–794 (2022).

Bataineh, M. J., Sánchez-Sellero, P. & Ayad, F. Green is the new black: how research and development and green innovation provide businesses a competitive edge. Bus. Strategy Environ. 33, 1004–1023 (2024).

Gropengießer-Arlt, L. & Zacharias, N. A. Accumulate or diversify ecological innovation assets? The effect of ecological innovation asset depth and breadth on firm financial performance. Bus. Strategy Environ. 34, 4001–4029 (2025).

Chen, Y. K., Liu, L. T. & Dong, J. R. Environmental regulation tools, regional differences and corporate green technology innovation: an analysis of Chinese provincial data based on system GMM and dynamic threshold. Sci. Res. Manag. 43, 111–118 (2022).

Ma, L. N. & Chen, Y. H. Research on the impact of environmental supervision on the increment and quality improvement of green innovation. Sci. Res. Manag. 47, 146–155 (2026).

Wang, L. & Si, Y. Analysis of the inverted “U” relationship between R&D intensity and green innovation performance: a study based on listed manufacturing enterprises in China. Sustainability 17, 7625 (2025).

Zhang, Q., Qi, L. & Loh, L. The inverted U-shaped effect of environmental taxation on green innovation: the roles of corporate environmental responsibility and green finance. Sustainability 17, 9915 (2025).

Zhu, M. & Rastelli, R. Green finance and carbon emissions: a nonlinear and interaction analysis using Bayesian additive regression trees. Preprint at https://doi.org/10.48550/arXiv.2510.19785 (2026).

Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794 (ACM, 2016).

Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. Preprint at https://doi.org/10.48550/arXiv.1705.07874 (2017).

Wang, X. L., Pang, M. Y., Jin, J. H., Tao, X. Y. & Dong, H. M. Research on influencing factors of green innovation performance of heavily polluting enterprises based on machine learning. Sci. Res. Manag. 46, 178–189 (2025).

Zhao, H., Wang, J. & Yuan, Y. How can enterprises’ green innovation persist? A study based on explainable machine learning. Sustainability 17, 10071 (2025).

Wang, T., Luo, K. & Yu, C. Unlocking the institutional foundations of green innovation: a machine learning analysis. Technol. Forecast. Soc. Change 225, 124552 (2026b).

National People’s Congress. Environmental Protection Tax Law of the People’s Republic of China (2016).

State Council. Interim regulations on the administration of carbon emissions trading (State Council Decree No. 775) (2024).

Porter, M. E. & van der Linde, C. Toward a new conception of the environment–competitiveness relationship. J. Econ. Perspect. 9, 97–118 (1995).

Declarations

Funding

No funding was received for this study.

Author contributions

Conceptualization: Hao Huang, Yujun Hu; Data curation: Xinqiang Li, Bowen Zhang; Formal analysis: Hao Huang; Investigation: Hao Huang, Yiwei Huang, Anxue Li; Methodology: Hao Huang, Wenxuan Duan; Project administration: Yujun Hu; Resources: Anxue Li; Software: Yujun Hu; Supervision: Yujun Hu; Validation: Xinqiang Li, Bowen Zhang, Yang Li; Visualization: Yiwei Huang, Bowen Zhang; Writing – original draft: Hao Huang; Writing – review & editing: Yujun Hu, Wenxuan Duan, Yang Li.

Competing interests

The authors declare no conflicts of interest.

Data availability

The data used in this study are obtained from the CSMAR database (https://www.gtarsc.com/), the CNRDS database (https://www.cnrds.com/), and the China Statistical Yearbook (https://www.stats.gov.cn/). Access to the CSMAR and CNRDS databases requires a paid subscription. The derived dataset supporting the findings of this study is available from the corresponding author upon reasonable request.

Code availability

The custom code used for XGBoost–SHAP analysis in this study is available from the corresponding author upon reasonable request.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf