Digital Resource Disclosure Configuration and Innovation Output: Evidence from Chinese A-Share Manufacturing Firms

Zihan Zhang

Shenyang Feiyue Experimental Middle School, Xian, China
zhangzihan2027@126.com

Abstract Digital disclosure configuration provides little stable information about subsequent patent applications beyond a released multidimensional digital-transformation index. We study this measurement question in a public panel of 20,858 firm-years from 3,073 Chinese A-share manufacturers. Annual reports permit analysis at scale, although their disclosure frequencies do not necessarily describe resources combined in practice. The primary proxy multiplies active-category breadth by entropy normalized over four technology categories, thereby weighting richness twice. Joint fixed-effects estimation gives a standardized proxy coefficient of -0.0030 (95% confidence interval from -0.0242 to 0.0181; 20,495 observations), compared with an index coefficient of 0.0791 and an interval from 0.0377 to 0.1205. Alternative configuration definitions yield similarly inconclusive full-panel estimates. The treatment of low-richness observations matters: among firm-years disclosing at least two categories, conditional evenness has a coefficient of -0.0906, an interval from -0.1706 to -0.0107, and 2,726 observations. Because this unadjusted sensitivity changes the population, it remains exploratory and establishes neither a full-versus-restricted difference nor harmful resource orchestration. The chronological prediction comparison also requires caution. Breadth ranks descriptively above the composite, but no uncertainty comparison accompanies that ranking. Patent applications are the observed outcome throughout; strategic innovation supplies the motivation for asking whether disclosed configuration contributes information beyond the broad index. The full-panel evidence is limited, and the interpretation of balance depends on how observations with few active categories enter the analysis.

Keywords digital transformation; resource configuration; corporate disclosure; patent applications; manufacturing firms

1. Introduction

Annual reports offer comparable evidence where firms' operational inventories are unavailable. The process they describe, digital transformation, encompasses technology and strategy as well as organizational change and value creation [1]. Researchers studying Chinese listed firms use these texts to measure digitalization and relate it to innovation potential, patent output, and innovation strategies [2–5]. What these associations mean depends in part on how document features become empirical variables [6].

Similar digital-transformation index values can accompany different distributions of language across artificial intelligence, big data, cloud computing, and blockchain. We therefore examine configuration separately from the overall level of transformation. Recombination theory emphasizes knowledge and coordination [7], while resource orchestration concerns managerial action beyond naming resources [8]. Unfamiliar combinations can also produce variable outcomes [9]. These arguments motivate the configuration question without presupposing a positive association between disclosure and patents.

Evidence on AI talk–action gaps illustrates how annual reports combine observable activity with strategic emphasis and communication [10]. Keyword counts, semantic embeddings, and composite readiness measures capture different aspects of digitalization [11, 12]. Their use does not, by itself, make disclosure a measure of implementation.

Our four technology-category frequencies describe a narrower object than the released index, which covers technology, strategy, organization, environment, outcomes, and applications. The primary configuration proxy multiplies breadth by entropy normalized with $log4$, so richness enters twice. Conditional evenness provides a

ISSN 3106-5783 (Print) ISSN 3106-7042 (Online)

sensitivity measure when at least two categories are active. These measurements concern disclosure and patent applications, the observed outcome; strategic innovation is the broader motivation for studying them.

RQ1 concerns the information that disclosed configuration contributes beyond the released index. We link year-t disclosure to year-$t+1$ invention applications using firm-and-year fixed effects; the primary full-panel coefficient is near zero. RQ2 examines whether estimates depend on the estimator, outcome, timing, financial treatment, category, or measurement definition. Conditional evenness is inconclusive across the full panel and negative in the smaller multi-category population. That sensitivity is exploratory, because it does not formally test a population difference. A separate chronological audit evaluates held-out predictive information without establishing predictive superiority. Both exercises examine measurement dependence, and neither design supplies causal identification.

2. Related Work

Disclosure diversity alone cannot establish implemented combinative capability. Kogut and Zander locate that capability in organizational knowledge and coordination [7]; Sirmon and colleagues examine how managers structure, bundle, and leverage resources [8]. Nor does recombination imply uniformly positive patent associations: unfamiliar technological combinations can accompany failed inventions as well as breakthroughs [9]. These accounts motivate examining configuration beyond disclosure volume, while keeping the textual measure distinct from the capability described by theory.

Our empirical question builds on associations between digital transformation and innovation potential, performance, and strategy in Chinese listed firms [2–5]. Their constructs and coefficients are specific to the study designs. Text-based work on green innovation, for example, uses different representations and identification choices [11, 13], which limits direct transfer to the present analysis.

Document features determine text-based variables [6]; embeddings and readiness indices need not convey equivalent information [11, 12]. Our data contain four frequency categories and a separate multidimensional index. Keeping that index in the model while varying configuration definitions makes this an incremental audit. Fixed-denominator entropy captures richness and allocation, whereas Pielou's normalization expresses evenness relative to observed richness [14]. Neither frequency-based measure recovers context or action, a limitation central to interpreting disclosure [10].

3. Data and Method

Public data and panel construction

Sang Chang contributed version 2 of the *Digital Transformation and Manufacturing Corporate Innovation Dataset*, published by Mendeley Data on 22 January 2026 under CC BY 4.0 [15]. We use its digital indicators, four technology-category frequency files, financial statements, and patent applications. The released digital-transformation index is used as supplied. It combines technology, strategy, organization, environment, outcomes, and applications, rather than representing a raw keyword total.

The population comprises A-share manufacturers with industry codes beginning with C. Disclosure years 2011–2021 were matched to financial variables and next-year patent applications. The primary outcome sums independently and jointly filed invention applications after listing. Firm size is log total assets, and the debt ratio is liabilities divided by assets. Other application definitions support scope and timing checks.

Missing next-year patent records were treated as missing outcomes, not zero applications, so 44 index rows were excluded. Deterministic joining produced 20,858 firm-years from 3,073 firms with complete primary variables. The estimation samples differ: 363 singletons are removed for firm-and-year fixed effects, leaving 20,495 observations, while additional exclusions for separation reduce the fixed-effects Poisson sample to 18,881 firm-years.

The zero audit checks the source's sparse-positive convention rather than independently verifying annual reports. All 20,858 rows could be matched to the source's four-category score table. For the 13,390 firm-years missing from positive term details, every category score equalled its year minimum. Reconstructed zeros coincided with those minima, and every positive count corresponded to a higher score, providing consistent internal evidence for the encoding.

Disclosure configuration

Let x_{itk} be firm i's year-t frequency, where k denotes one of the four categories. Let m_{it} be the number of categories with positive frequency. Breadth is

$$B_{it} = \frac{m_{it}}{4} \tag{1}$$

For $X_{it} = x_{it1} + x_{it2} + x_{it3} + x_{it4} > 0$, define $p_{itk} = x_{itk} / X_{it}$ and interpret $0log0$ as zero. Normalized category entropy is

$$H_{it} = -\frac{\sum_{k=1}^{4} p_{itk} log p_{itk}}{log 4} \tag{2}$$

The recombination proxy is $R_{it} = B_{it}H_{it}$ and is zero below two active categories. For m equally frequent categories, $H = log(m)/log(4)$; entropy already reflects richness, and

multiplication by breadth weights richness again. We retain the formula but interpret it only as a disclosure-based breadth-times-entropy proxy. Figure 1 separates its construction from the broader index and the subsequent evaluation lanes.

Because H uses $log(4)$, we also examine evenness conditional on observed richness. For $m_{it} \geq 2$, $J_{it} = H_{it} log(4) / log(m_{it})$ [14], and $A_{it} = B_{it} J_{it}$. A full-panel sensitivity encodes undefined J values for $m < 2$ as zero; these are bookkeeping values, not observed evenness. A restricted analysis estimates B and J jointly only when $m \geq 2$. Whole-panel means and population standard deviations are retained under restriction.

Association, sensitivity, and prediction

The primary log-linear fixed-effects ordinary least-squares model is

$$log(1 + Y_{i,t+1}) = \beta Q_{it}^{z} + \gamma D_{it}^{z} + \delta^{\top} C_{it}^{z} + \alpha_i + \lambda_t + \epsilon_{it}, \quad (3)$$

where Y is the subsequent invention-application count, Q is the focal configuration measure, D is the released index, C contains log assets and the debt ratio, and z denotes whole-panel standardization. Indices i and t identify firms and disclosure years. The coefficients β, γ, and δ correspond to the focal measure, index, and controls; α_i and λ_t are firm and year effects, and ϵ_{it} is the error. Uncertainty is clustered by firm. The primary specification sets $Q = R$. The standalone index model omits Q; breadth, entropy, alternative-composite, and conditional-evenness models substitute the stated measure while retaining D and the controls.

Sensitivity checks use fixed-effects Poisson, alternative patent scopes and timing, frequency trimming, category omissions, and five within-industry-year permutations. Financial checks winsorize or trim debt ratios at the first and 99th percentiles, exclude values above one, or omit the control; an observed-detail audit changes the population. The configuration audit compares R, A, and J in the full panel, then estimates B and J jointly for $m \geq 2$. Its p-values are unadjusted, with no formal population-difference test.

Deterministic local panel assembly read workbooks with openpyxl and joined records through Python dictionaries keyed by firm and year. Subsequent audits and estimation used pandas 2.3.3 and NumPy 2.2.6. Estimation ran on an AutoDL CPU under Python 3.10.8 and pyfixest 0.60.0, with firm-clustered CRV1 inference. Scikit-learn 1.7.2 median-imputed and standardized the released index, any added measure, log assets, debt ratio, and numeric year; only industry was one-hot encoded, with unseen categories ignored. Poisson regression modeled the raw next-year invention count by Newton–Cholesky with a fixed 10^{-6} penalty, 500-iteration limit, and 10^{-8} tolerance. Five recorded seeds governed permutations; annual mean-score intervals used 2,000 bootstrap resamples and year-specific seeds.

Chronological prediction uses 2011–2017 for training, 2018–2019 for validation, and 2020–2021 for testing. The baseline contains D, log assets, debt ratio, year, and industry; three models add B, H, or R. Preprocessing and coefficients are fitted only on training data. Validation checks the pipeline without model selection or tuning. Test Poisson deviance and mean absolute error describe held-out fit, with no uncertainty test for model rankings.

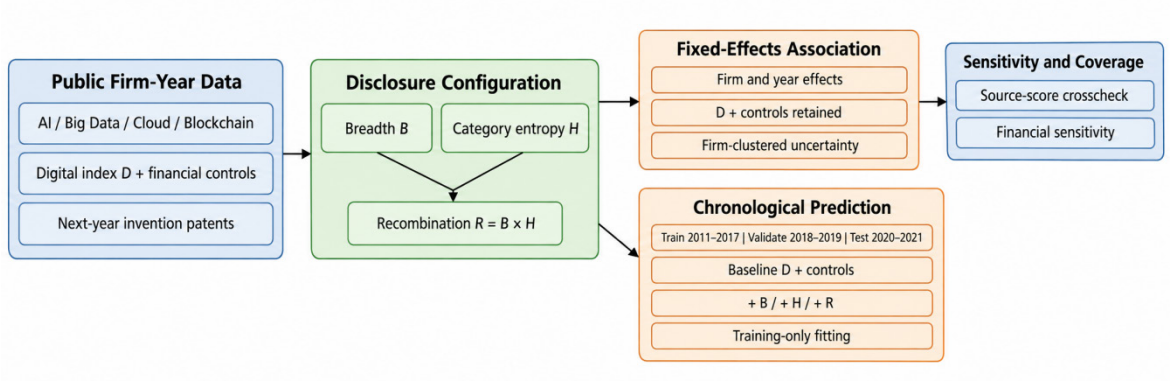


Figure 1. Workflow. Disclosed categories form breadth B, fixed-denominator entropy H, and $R = BH$; the released multidimensional index D remains separate. The right-hand lanes evaluate fixed-effects associations, measurement and coverage sensitivities, negative controls, and chronological prediction.

4. Results

Construct distribution

The panel contains 20,858 firm-years; 28.50% have no subsequent invention application. The outcome median is 3, its mean is 23.68, and its maximum is 8,448. Firm-years with zero through four active categories number 13,390, 4,358, 1,891, 1,167, and 52, respectively, leaving

ISSN 3106-5783 (Print) ISSN 3106-7042 (Online)

85.09% of recombination scores at zero. Figure 2 shows annual multi-category prevalence and mean scores.

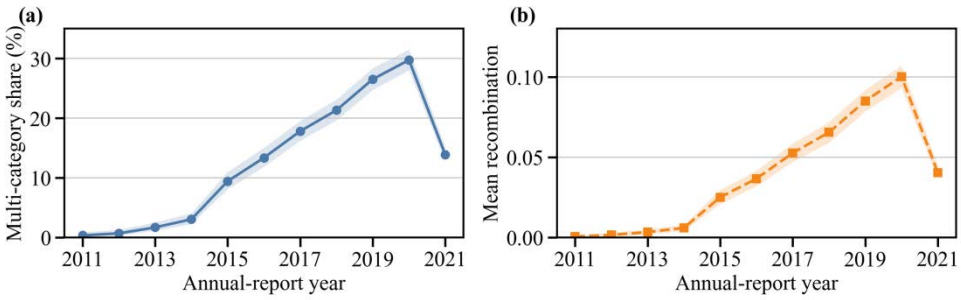


Figure 2. Annual evolution of the disclosure construct. Panel (a) gives the share with at least two active categories and Wilson 95% intervals. Panel (b) gives mean R with bootstrap 95% intervals.

Normalized entropy closely tracked the product: their Spearman correlation was 0.9996 overall and 0.9563 within positive-product rows. The financial data also required scrutiny. A median debt ratio of 0.375 accompanied a maximum of 178.35, with 114 values above one and six above ten. This right skew and the correlation between proxies motivated the financial and construct checks below.

Conditional associations

Table 1 separates the standalone index estimate from its coefficient in the joint recombination model. The released digital-transformation index has a coefficient of 0.0773 alone (95% CI $[0.0377, 0.1169]$; $p = 0.00013$) and 0.0791 with R (95% CI $[0.0377, 0.1205]$; $p = 0.00018$).

Table 1. Firm-and-year fixed-effects estimates for log $(1+$ subsequent invention-patent applications$)$. All reported variables and financial controls are standardized; uncertainty is clustered by firm. The standalone index row omits B, H, and R. Other rows include the released index, log assets, and the debt ratio. The symbol n denotes firm-years.

Reported term	Model	Coefficient	SE	95% CI	p value	n
Released index	Standalone	0.0773	0.0202	[0.0377, 0.1169]	0.00013	20,495
Released index	Index +R	0.0791	0.0211	[0.0377, 0.1205]	0.00018	20,495
Breadth (B)	Index +B	0.0043	0.0129	[−0.0211, 0.0297]	0.7394	20,495
Entropy (H)	Index +H	−0.0020	0.0106	[−0.0228, 0.0188]	0.8510	20,495
Recombination (R)	Index +R	−0.0030	0.0108	[−0.0242, 0.0181]	0.7799	20,495

The recombination coefficient is -0.0030 (95% CI $[-0.0242, 0.0181]$; $p = 0.7799$). The interval includes small associations of either sign, so the full-panel answer to RQ1 is inconclusive. Breadth and fixed-denominator entropy intervals also span zero conditional on the released index.

Sensitivity and predictive information

Figure 3 summarizes selected sensitivity intervals and the chronological audit. Fixed-effects Poisson, patent-scope, timing, group-company, frequency-trim, and four category-omission intervals span zero. Five permutation estimates range from -0.0058 to 0.0051, each with an interval crossing zero.

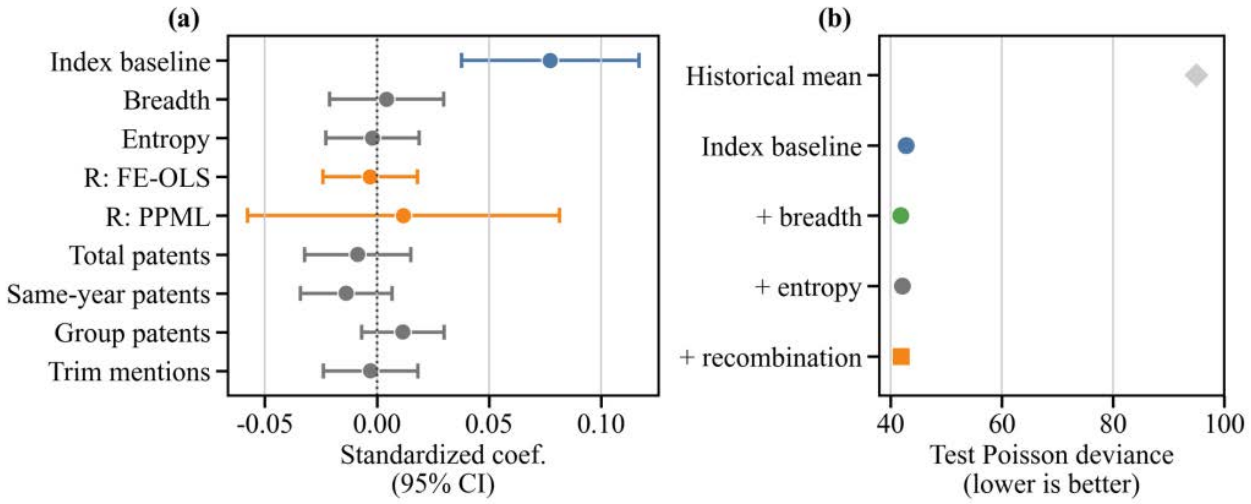


Figure 3. Incremental evidence beyond the released multidimensional index. Panel (a) reports focal estimates and 95% intervals. Panel (b) reports test Poisson deviance; lower values indicate better held-out fit. The ranking is descriptive and has no uncertainty comparison.

Table 2. Configuration and other sensitivities. Models include the released index and two-way fixed effects with firm-clustered uncertainty; n denotes firm-years and p values are unadjusted. Full-panel J encodes $m<2$ as zero; restricted rows require $m^3 2$.

Population/check	Term	Estimate	95% CI	p value	n
Winsorized debt ratio	R	−0.0027	[−0.0238, 0.0183]	0.7984	20,495
Trimmed debt ratio	R	−0.0013	[−0.0225, 0.0199]	0.9022	20,074
Debt ratio ≤1	R	−0.0025	[−0.0236, 0.0187]	0.8189	20,380
Debt-ratio control omitted	R	−0.0030	[−0.0242, 0.0181]	0.7803	20,495
Observed term-detail rows	R	−0.0195	[−0.0419, 0.0029]	0.0879	6,944
Full panel	A	−0.0020	[−0.0228, 0.0187]	0.8495	20,495
Full panel	J	−0.0010	[−0.0206, 0.0186]	0.9200	20,495
Full panel, joint	B	0.0111	[−0.0231, 0.0453]	0.5234	20,495
Full panel, joint	J	−0.0074	[−0.0336, 0.0188]	0.5783	20,495
m≥2, joint	B	−0.0350	[−0.1153, 0.0453]	0.3924	2,726
m≥2, joint	J	−0.0906	[−0.1706, −0.0107]	0.0264	2,726

Table 2 addresses the remaining RQ2 checks. Financial variants remain close to the primary estimate. The observed-detail estimate is more negative at −0.0195, but its interval spans zero and the population changes. Full-panel A, J, and joint $B-J$ intervals also span zero. The restricted $m^3 2$ joint model retains 2,726 of 3,110 input rows after 384 singleton removals. Breadth remains inconclusive, whereas conditional evenness is −0.0906 (95% CI [−0.1706, −0.0107]; unadjusted $p=0.0264$).

ISSN 3106-5783 (Print) ISSN 3106-7042 (Online)

This changed-population estimate is neither a formal contrast with the full panel nor evidence that resource orchestration is harmful.

The prediction audit evaluates incremental information, not deployment suitability. It uses 10,740 training, 4,409 validation, and 5,709 test observations; every Poisson regression converged in 11 iterations. Against historical-mean deviance of 95.0321, the index baseline achieved 42.7395. Extending that baseline with breadth, entropy, or recombination gave 41.7592, 42.1124, and 41.8921, respectively. These are descriptive comparisons because ranking uncertainty was not assessed. Breadth had the lowest test deviance, ahead of recombination by 0.1329.

5. Discussion

The results define a measurement boundary rather than a verdict on configuration. In the full panel, breadth times fixed-denominator entropy adds no stable within-firm association beyond the released index and controls. Its interval includes small values of either sign; this does not show that firms failed to combine resources.

Zero products account for 85.09% of observations, leaving H and R nearly collinear; breadth predicts slightly better. Richness affects both fixed-denominator entropy and its breadth multiplier. Conditional evenness separates allocation only when at least two categories are active, so handling low richness changes the interpretation. Without a population-difference test, its unadjusted, exploratory estimate supports neither a subgroup effect nor harm.

The released index spans strategy, organization, environment, technology, outcomes, and applications; R uses four technology-category frequencies. Their joint model tests incremental association, not a causal return to transformation.

Financial variants leave the estimates largely unchanged, but these checks and fixed effects do not eliminate time-varying confounding. Population dependence appears in the observed-detail and multi-category results. Source-score agreement addresses a separate issue: internal consistency of sparse-positive encoding, rather than independent verification of annual reports.

Interpreting disclosure balance as orchestration would require preregistered definitions, direct population tests, and text linked to operational records and quality-sensitive patent data. Annual-report frequencies presently cannot distinguish deployment from communication, and application counts record volume rather than quality. The evidence also concerns listed manufacturers, limiting its generalization beyond that population.

6. Conclusion

Full-panel configuration estimates are inconclusive beyond the released index, whereas conditional evenness is negative in the smaller multi-category population. This exploratory, unadjusted association is noncausal; breadth also predicts descriptively better than the composite. Inference therefore depends on construct algebra and population definition. Contextual text, operational records, and direct contrasts are needed before disclosed balance can represent resource orchestration.

References

[1] Vial, G., "Understanding digital transformation: A review and a research agenda," *The Journal of Strategic Information Systems*, 28, (2), 118–144, 2019, https://doi.org/10.1016/j.jsis.2019.01.003.

[2] Zhao, X., Chen, Q., Yuan, X., Yu, Y., and Zhang, H., "Study on the impact of digital transformation on the innovation potential based on evidence from chinese listed companies," *Scientific Reports*, 14, (1), 6183, 2024, https://doi.org/10.1038/s41598-024-56345-2.

[3] Jin, Z., and Qin, J., "Does digital transformation promote innovation performance? Evidence from listed chinese firms," *Journal of Accounting, Business and Finance Research*, 18, (1), 12–25, 2024, https://doi.org/10.55217/102.v18i1.753.

[4] Zhu, N., "Digital transformation and corporate innovation: Evidence from chinese listed companies," *Advances in Economics, Management and Political Sciences*, 280, (1), 229–238, 2026, https://doi.org/10.54254/2754-1169/2026.35470.

[5] Pu, T., "Does digital transformation propel innovation strategies? An empirical investigation based on machine learning and the dynamic GMM approach," *E+M Ekonomie a Management*, 28, (2), 81–96, 2025, https://doi.org/10.15240/tul/001/2025-5-012.

[6] Gentzkow, M., Kelly, B., and Taddy, M., "Text as data," *Journal of Economic Literature*, 57, (3), 535–574, 2019, https://doi.org/10.1257/jel.20181020.

[7] Kogut, B., and Zander, U., "Knowledge of the firm, combinative capabilities, and the replication of technology," *Organization Science*, 3, (3), 383–397, 1992, https://doi.org/10.1287/orsc.3.3.383.

[8] Sirmon, D. G., Hitt, M. A., Ireland, R. D., and Gilbert, B. A., "Resource orchestration to create competitive advantage: Breadth, depth, and life cycle effects," *Journal of Management*, 37, (5), 1390–1412, 2011, https://doi.org/10.1177/0149206310385695.

[9] Fleming, L., "Recombinant uncertainty in technological search," *Management Science*, 47, (1), 117–132, 2001, https://doi.org/10.1287/mnsc.47.1.117.10671.

[10] Long, H., and Hao, Y., "AI talk–action gaps and ESG rating sensitivity: Evidence from chinese A-share firms," *Oeconomia Copernicana*, 17, (2), 457–513, 2026, https://doi.org/10.24136/oc.4220.

[11] Yin, X., and Chen, Y., *A semantic embedding and dynamic PSM-DID framework for causal analysis of digital transformation on green patent quality*, *Proceedings of the 2025 5th international conference on big data, artificial intelligence and risk management*, 525–529, ACM, 2025, https://doi.org/10.1145/3800227.3800305.

[12] Trequattrini, R., Cuozzo, B., Manzari, A., and Schimperna,

M., "Measuring digital-sustainability readiness in industrial SMEs: A composite index and empirical evidence," *Journal of Global Responsibility*, 17, (2), 293–321, 2026, https://doi.org/10.1108/JGR-04-2025-0120.

[13] Li, X., Tang, K., and Ding, Y., *How does digital transformation empower green innovation in chinese enterprises: Evidence based on Python text analysis*, *Proceedings of the 2025 3rd international conference on internet of things and cloud computing technology*, 239–244, ACM, 2025, https://doi.org/10.1145/3776865.3776906.

[14] Pielou, E. C., "The measurement of diversity in different types of biological collections," *Journal of Theoretical Biology*, 13, 131–144, 1966, https://doi.org/10.1016/0022-5193(66)90013-0.

[15] Chang, S., *Digital transformation and manufacturing corporate innovation dataset (original data)*, Mendeley Data, version 2, doi: 10.17632/b464y5cjkp.2; Mendeley Data, 2026, https://doi.org/10.17632/b464y5cjkp.2.