Firm-Pair Merger Prediction in Technology and Bio Tech Sectors: A Topological and Machine Learning Approach

Group 5:
Arthur Khamkhosy, Ananya Ravichandran, Torrin Pataki, Amie Le Hoang, Duncan Harrop

April 27, 2026

Contents

1 Introduction

Predicting mergers and acquisitions (M&A) is notoriously difficult. Traditional screening models by investment banks and corporate development teams focus heavily on fundamental metrics such as valuation ratios (e.g., Tobin's q), free cash flow, and target size relative to the acquirer. While these metrics successfully capture the financial feasibility of a transaction, they fail to capture the underlying strategic rationale, namely technological and research & development (R&D) synergy. In the technology and life science sectors, large-cap acquirers frequently bypass traditional value investments in favor of targets that complete their ecosystem or clinical pipeline. M&A is driven by firm incentives for R&D, in some cases incentivizing large firms to engage in M&A as an alternative to internal R&D [1].

This project develops a quantitative M&A prediction pipeline that combines traditional corporate finance screening with advanced Natural Language Processing (NLP). NLP has been used in the M&A space before in analysis of firm 10-Ks [2]. We extend this analysis by analyzing firm patent filings as a proxy for R&D. By embedding the patent text of 7,485 deduplicated firms into a geometric space using PatentSBERTa [3], we created a new measurement of technological overlap between two firms. This approach provides clear implications such as generating target lists that map the strategic supply chains of large-cap technology and biopharmaceutical acquirers, bridging the gap between theoretical corporate strategy and quantitative data.

Furthermore, this framework moves beyond static target origination by introducing a dynamic strategic simulator. By merging the patent portfolios of an acquirer (A) and a target (x), we can project a synthetic post-merger probability distribution (C). This $A + x = C$ topological application allows us to calculate the shift in semantic proximity (ΔBC) between the newly formed entity and the broader competitive landscape. Consequently, the model not only identifies who a firm should acquire, but quantitatively reveals the underlying macroeconomic motive such as bridging a moat to neutralize a specific market rival.

2 Methodology

2.1 Data Scope: 2020 Macroeconomic Baseline

From a macroeconomic perspective, 2020 represented an inflection point characterized by unprecedented quantitative easing and a Zero Interest Rate Policy (ZIRP). This environment of extremely low cost of capital catalyzed a historic wave of global M&A activity. Simultaneously, COVID-19 pandemic forced a sudden, transformational shifts in corporate strategy, accelerating semiconductor/cloud consolidation to support digital infrastructure, and driving aggressive biopharmaceutical R&D acquisitions.

By using pre-merger data from 2020, we establish a robust "clean room" with limited data leakage and look-ahead bias. This allows the algorithmic pipeline to be stress-tested against the subsequent historic wave of mega-deals executed in late 2020 and 2021. Evaluating the model during a period of peak consolidation proves its ability to separate genuine technological synergy from the noise of an active, high liquidity financial market.

2.2 Data Architecture & PatentSBERTa Embeddings

To capture unobserved technological heterogeneity across our dataset of 7,485 deduplicated technology and biotechnology firms, we vectorize patent portfolios using PatentSBERTa [3], a transformer model optimized for technical language. For each patent, we generate two distinct 768-dimensional embeddings. First, we concatenate the title and abstract to encode both the technological base found in the title (following [4]) and the patent's novel claims. Second, following the technological overlap framework of [5], we vectorize citations by averaging the title-abstract embeddings of all cited patents to represent the foundational knowledge upon which the patent builds. The primary embedding is then appended to the average citation vector and projected via Uniform Manifold Approximation and Projection (UMAP) onto a 50-dimensional manifold. This reduction preserves the local and global topological structures of the latent space while ensuring computational tractability for high-volume pairwise operations.

2.3 Topological Mapping with Gaussian Mixture Models

We model each firm's patent portfolio as a distribution rather than a single average vector. Specifically, we apply a Bayesian Gaussian Mixture Model (GMM) to the 50-dimensional patent embeddings.

The probability density of a firm's patent portfolio is defined as:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

Where K is the number of identified technology clusters, π_k represents the mixing weight (the proportion of the firm's portfolio dedicated to that cluster), μ_k is the cluster centroid in the embedding space, and Σ_k is the covariance matrix dictating the spread of that technology area.

Relying on a Bayesian Dirichlet Process prior, firms naturally prune unused components bounded by a maximum threshold. Stability in the prior can be found in A.3. Following extensive convergence testing, we empirically locked the production specification at $K_{max} = 15$. K_{max} convergence analyses can be found in A.7.

2.4 Redefined Technological Overlap: Bhattacharyya Coefficient

To measure the synergy between any two firms, we calculate the Bhattacharyya Coefficient (BC), which quantifies the geometric overlap of their respective GMM distributions. The production matrix utilizes a linear-weighted BC formula bounded in $[0, 1]$. This metric allows us to identify deep structural R&D synergies even between firms that share zero direct patent citations, effectively revealing hidden technological supply chains. The distribution of BC across the dataset can be found in A.8. The BC is calculated as follows:

$$BC(A, B) = \sum_{i=1}^{K_A} \sum_{j=1}^{K_B} \pi_i \pi_j e^{-D_B(\mathcal{N}_i, \mathcal{N}_j)} \tag{1}$$

Where the Bhattacharyya Distance (D_B) for components with diagonal covariance is:

$$D_B(\mathcal{N}_i, \mathcal{N}_j) = \frac{1}{8} \sum_{d=1}^{D} \frac{(\mu_{i,d} - \mu_{j,d})^2}{\sigma_{i,j,d}^2} + \frac{1}{2} \sum_{d=1}^{D} \ln \left(\frac{\sigma_{i,j,d}^2}{\sigma_{i,d}\sigma_{j,d}} \right) \tag{2}$$

The variables are defined as follows:

- $\mu_{i,d}, \mu_{j,d}$: The **mean** positions of clusters i and j in dimension d, representing the "center" of a specific technological focus.

- $\sigma_{i,d}, \sigma_{j,d}$: The **standard deviation** (variance) of the clusters, representing the "breadth" or diversity of the firm's RD in that niche.

- $\sigma_{i,j,d}^2$: The **pooled variance**, calculated as $\frac{\sigma_{i,d}^2 + \sigma_{j,d}^2}{2}$, used to stabilize the distance calculation.

2.5 Directional Complementarity

To capture capability-acquisition motives that the symmetric BC may overlook, we introduce a directional feature: $Comp(A \to B)$. This metric measures "gap-filling" by identifying whether Firm B concentrates its R&D in technological zones where Firm A is underweight relative to the global patent landscape.

Unlike a simple dissimilarity measure, this "Market-Relative" approach asks if Firm B provides strength specifically where Firm A underinvests compared to the global average ($p_{market,k}$). The measure is asymmetric ($Comp(A \to B) \neq Comp(B \to A)$) and is defined as:

$$Comp(A \to B) = \sum_{k=1}^{K} \max(0, p_{market,k} - p_{A,k}) \cdot p_{B,k} \tag{3}$$

3 Model Performance and Results

3.1 Strategic Topology: Synthetic Merger Application ($A + x = C$)

Beyond the static origination of target pipelines, the topological geometry of the PatentSBERTa embeddings enables dynamic strategic simulations. Specifically, we can model the macroeconomic competitive shift that occurs post-acquisition.

Let A represent the patent portfolio of the acquirer, and x represent the patent portfolio of the target firm. We simulate a merger by computationally combining the patent embeddings into a synthetic merged portfolio C, such that $A \cup x = C$. A new Gaussian Mixture Model is then fitted over this combined synthetic portfolio C.

With the new synthetic GMM established, we recalculate the pairwise Bhattacharyya Coefficient against other firms in the economy. The strategic intent of the merger can be quantified by observing the delta:

$$\Delta BC_{firm} = BC(C, firm) - BC(A, firm)$$

Theoretically, high baseline BC scores identify firms occupying identical technological spaces (direct competitors). Therefore, if ΔBC is large and positive for a specific firm, it implies the acquirer executed the merger in an explicit attempt to become more similar to that specific competitor. Due to the computational constraints of refitting the global $N \times N$ matrix, this simulated ΔBC is calculated exclusively against the acquirer's top 20 existing baseline competitors.

To evaluate this synthetic application, we analyzed four specific historical transactions. While ΔBC was positive for some acquirer-comparator pairs, ΔBC was close to 0 or even negative for other pairs. Large, positive increases in BC are highlighted and analyzed for interpretability. Analysis of directional complementarity as detailed in 2.5 is also included.

3.1.1 Case Study 1: AMD & Xilinx (Computing to Communications)

AMD (CPUs/GPUs) acquired Xilinx (FPGAs) in 2022 to penetrate the communications infrastructure market. $Comp(AMD \to Xilinx) = 0.0019$. The low directional complementarity score validates the interpretation of the merger as an *ecosystem expansion* into a new domain (FPGAs) rather than a correction of existing market-average weaknesses while the stronger reverse score ($Comp(Xilinx \to AMD) = 0.0060$) confirms the broader CPU/GPU platform absorbed the niche, not vice versa.

- **AIL Technologies:** Represents the combined entity's pivot toward RF sensors and 5G signal processing.

Table 1: Strategic Shift (ΔBC) for AMD following the synthetic integration of Xilinx.

Comparator	**Rank in Acquirer's Top 20**	**BC (Acquirer, Comp.)**	**BC (Target, Comp.)**	**BC (Synthetic, Comp.)**	ΔBC
AIL Technologies	6	0.0047	0.0014	0.0321	**0.0274**
Commodore Intl Ltd	7	0.0045	0.0010	0.0090	**0.0045**
Micron Technology Inc	13	0.0031	0.0009	0.0060	**0.0029**

- **Micron Technology Inc:** Indicates convergence in data center chiplet architecture, where FPGAs require high-bandwidth memory (HBM) integration to compete with Intel.
- **Commodore Intl Ltd:** Mechanical shift driven by shared legacy FPGA circuitry.

3.1.2 Case Study 2: Analog Devices & Maxim (Strategic Mimicry)

Analog Devices (ADI) acquired Maxim in 2021 to challenge Texas Instruments (TI) in the broad-market analog sector. $Comp(AnalogDevices \rightarrow Maxim) = 0.0043$. The strong signal identifies ADI as the primary beneficiary of portfolio rebalancing, with the minimal reverse score (0.0014) confirming Maxim acting strictly as the gap-filler. This confirms the deal's nature as a consolidation where the acquirer patches market-relative gaps.

Table 2: Strategic Shift (ΔBC) for ADI following the synthetic integration of Maxim Integrated.

Comparator	**Rank in Acquirer's Top 20**	**BC (Acquirer, Comp.)**	**BC (Target, Comp.)**	**BC (Synthetic, Comp.)**	ΔBC
iWatt	9	0.0081	0.0604	0.0122	**0.0041**
National Semiconductor	4	0.0094	0.0083	0.0131	**0.0037**

- **National Semiconductor:** Direct evidence of *strategic mimicry*. By absorbing Maxim, ADI's distribution converged with the "general-purpose analog" profile defined by National Semiconductor (a key TI acquisition).
- **iWatt:** Represents a *capability acquisition* in digital power management, bridging ADI's linear expertise with Maxim's PMIC patents.

3.1.3 Case Study 3: Merck and Co. & Acceleron (Niche Pivot)

This acquisition demonstrates the model's sensitivity to small-target pivots in high-stakes Biotech clusters (TGF-beta proteins). $Comp(Merck \rightarrow Acceleron) = 0$. This distinguishes the acquisition as a targeted clinical-asset deal rather than a broad strategic rebalancing of the R&D portfolio. Conversely, $Comp(Acceleron \rightarrow Merck) = 0.0092$ reflects the specialized biotech gaining the safety of a diversified platform while Merck secures a singular asset.

Table 3: Strategic Shift (ΔBC) for Merck & Co. following the synthetic integration of Acceleron Pharma.

Comparator	**Rank in Acquirer's Top 20**	**BC (Acquirer, Comp.)**	**BC (Target, Comp.)**	**BC (Synthetic, Comp.)**	ΔBC
Bristol-Myers Squibb	12	0.0090	0.0004	0.0371	**0.0280**
Aventis Pharmaceuticals	13	0.0077	0.0003	0.0313	**0.0236**
SmithKline Beecham (GSK)	8	0.0127	0.0006	0.0278	**0.0151**

- **Bristol-Myers Squibb:** Merck captured the TGF-beta signaling space previously dominated by the BMS-Acceleron partnership (*Reblozyl*).

- **Aventis/GSK:** Reflects Merck's entry into Pulmonary Arterial Hypertension (PAH). The shift toward these legacy respiratory leaders validates the model's ability to map expansion into uncharted RD territories.

3.1.4 Case Study 4: HP & Poly (Model Stress Test)

Despite Poly holding only 9 patents, the model captured nuanced shifts in HP's hardware-centric portfolio toward enterprise services. $Comp(HP \to Poly) = 0.0081$. The signal suggests some underlying technological gap-filling, though the high reverse score ($Comp(Poly \to HP) = 0.0204$) indicates Poly was largely absorbed into HP's broader hardware platform, with sales-channel bundling being the primary driver.

Table 4: Strategic Shift (ΔBC) for HP following the synthetic integration of Poly.

Comparator	**Rank in Acquirer's Top 20**	**BC (Acquirer, Comp.)**	**BC (Target, Comp.)**	**BC (Synthetic, Comp.)**	ΔBC
Aprisma Management Tech	16	0.0042	0.0000	0.0059	**0.0017**
Guidance Software Inc	14	0.0046	0.0000	0.0062	**0.0016**
Oracle Corp	18	0.0038	0.0000	0.0053	**0.0015**
IBM	15	0.0042	0.0000	0.0056	**0.0014**

- **Aprisma/Guidance Software:** High interpretability regarding "Poly Lens" software; HP's distribution moved toward network manageability and endpoint visibility.
- **Oracle/IBM:** Secondary evidence of strategic mimicry as HP transitions from a hardware vendor to an enterprise solutions provider.

3.2 Predictive Modeling using BC

[5] find technological overlap to be a significant predictor in determining firm-pair mergers. A natural extension of this project is to test the predictive power associated with the constructed pairwise BC scores. Due to limited time and compute, a logistic regression is conducted with a limited sample. The regression is computed as follows:

$$\ln\left(\frac{P_{i,j}}{1-P_{i,j}}\right) = \beta_0 + \beta_1 BC_{i,j} + \beta_2 \text{TargetPatents}_j + \beta_3 \text{SizeRatio}_{i,j} + \beta_4 \text{SizeRatio}_{i,j} * BC_{i,j}$$
$$+\beta_5 \text{ValuationGap}_{i,j} + \beta_6 \text{ValuationGap}_{i,j} * BC_{i,j} + \beta_7 \text{PEGap}_{i,j} + \beta_8 \text{EBITDAMarginGap}_{i,j} \quad (4)$$

Where: *Valuation Gap = Target Tobin's q − Acquirer Tobin's q*

There are 46 M&A events from 2020 included in the dataset. For every true M&A target ($P_{i,j} = 1$), there exists 20 fake targets ($P_{i,j} = 0$). The input data is limited to public firms with available financial data from Compustat. Descriptive statistics by quartile are shown as follows:

Table 5: Descriptive Statistics by BC Quartile

BC Quartile	**Tobin's q**	**Size Ratio**	**Valuation Gap**	**PE Gap**	**EBITDA Margin Gap**
Q1 (Lowest BC)	3.213	30.771	0.634	-4.745	0.015
Q2	3.194	6.491	-0.040	6.011	-0.039
Q3	3.499	6.212	0.360	62.070	-0.0216
Q4 (Highest BC)	5.652	4.236	2.587	17.978	-0.280

The descriptive statistics reveal a strong positive correlation between technological similarity and market valuation, as the average Tobin's q nearly doubles from 3.213 in the lowest quartile to

5.652 in the highest. As firms move into the highest technological overlap (Q4), the average size ratio drops significantly to 4.236, suggesting that deals between highly similar firms involve more balanced partners rather than lopsided "tuck-in" acquisitions. Furthermore, the substantial increase in the valuation gap (2.587) and the sharply negative EBITDA margin gap (-0.280) in the top quartile indicate that highly similar targets are often premium-valued assets with lower current margins, potentially signaling that acquirers are prioritizing the capture of high-growth technological synergies over immediate operational profitability.

Table 6: Weighted Logistic Regression Coefficients and Odds Ratios

Variable	**Coefficient**	**Odds Ratio**
Intercept	-2.0764	—
BC	-1.1494	0.3168
Target Patents	-6.0383	0.0024
Size Ratio	-3.5400	0.0290
Size Ratio * BC	0.5360	1.7092
Valuation Gap	-0.5235	0.5924
Valuation Gap * BC	-0.1623	0.8502
PE Gap	-0.2830	0.7535
EBITDA Margin Gap	-0.0820	0.9213

The model utilizes L2 normalization to prevent overfitting due to the small sample size. The model accuracy is represented using Hit@5 and MRR. The model demonstrates significant predictive accuracy with a Hit@5 of 0.72 and an MRR of 0.43. This implies that for 72% of historical M&A events in the validation set, the true target was inside its top 5 recommendations. A random model would only be able to achieve 23.8% accuracy. Looking at Table 6, large negative coefficients for target patent counts and size ratios confirm that sheer size act as the largest deterrent to acquisition. However, the negative coefficient for BC suggests that firms in similar technology spaces are unlikely to merge. This is likely due to the linear nature of the logistic regression conflating high technological overlap with large-scale "un-acquirable" firms. This forces the model to penalize similarity scores to minimize error, masking the benefits of technological synergy. Interestingly enough, by adding interaction terms ($SizeRatio * BC$, $ValuationGap * BC$), BC has a direct impact on the size and valuation variables. The positive coefficient on $SizeRatio * BC$ suggests high technological synergy mitigates the penalty of target size. The negative coefficient on $ValuationGap * BC$ suggests that for high BC firm-pairs, the market is more sensitive to changes in the spread of Tobin's q (represented by the valuation gap). These observations provide new insights on the relationship between overlapping technology spaces and M&A market dynamics.

4 Extensions

4.1 Strategic Topology Extension

The case studies in 3.1 are a litmus test for the model's ability to model strategic intent. Due to limited compute, the methodology also lacks target identification. To expand on this study, it is only natural to bring in additional M&A events from SDC Platinum, run the model on a larger scale, and justify results using a more statistically sound methodology.

4.2 ML Predictive Modeling

The current predictive model in 3.2 is limited in scope. In order to develop a more robust model, additional data (in regards to M&A events and private firms) needs to be added. Additionally, a natural extension of this predictive model would be to use a ML model that captures non-linear relationships in input variables (e.g. XGBoost). Studies like [6] can be leveraged with the addition of our constructed pairwise BC scores to identify the feasibility of BC as a feature.

4.3 Directional Complementarity Extension

To further ground the directional complementarity framework in economic theory, we propose three primary extensions. First, by analyzing technological "overweights" rather than just gaps, the model can identify opportunities for strategic diversification away from overexposed or declining sectors. Second, to transition from general market averages to firm-specific needs, we suggest benchmarking against the patent distributions of industry leaders as a proxy for value-maximizing "target" portfolios. Finally, we propose a dynamic "active share" measure to track how a firm's market-relative gaps evolve over time, allowing us to test whether proactive realignment serves as a precursor to significant market value gains.

5 Conclusions

This project developed a quantitative tool that bridges the theoretical gap between corporate strategy and machine learning. By using PatentSBERTa embeddings and Gaussian Mixture Models, we developed a mathematically rigorous framework to aggregate patent embeddings and model patent portfolios. The model was then used to compare patent portfolios through the Bhattacharyya Coefficient and Directional Complementarity scores.

To address model capability and performance, we conduct a series of case studies using the newfound metrics. The case studies showed high interpretability consistent with economic theory and market analyses. We furthermore create a predictive model using BC alongside Compustat financials. While the current results show little evidence of predictive power with pairwise BC scores over patent counts and relative firm sizes, we provide the framework for extensions that better utilize the BC and directional complementarity scores.

References

[1] G. M. Phillips and A. Zhdanov, "R&D and the incentives from merger and acquisition activity," *The Review of Financial Studies*, vol. 26, no. 1, pp. 34–78, 2013.

[2] G. Hoberg and G. Phillips, "Text-based network industries and endogenous product differentiation," *Journal of political economy*, vol. 124, no. 5, pp. 1423–1465, 2016.

[3] H. Bekamiri, D. S. Hain, and R. Jurowetzki, "PatentSBERTa: A deep NLP based hybrid model for patent distance and classification using augmented SBERT," *Technological Forecasting and Social Change*, vol. 206, p. 123536, 2024.

[4] D. Barkemeyer, J. Bena, and X. E. Li, "Technology life cycles: Investment, growth, and firm value," *Growth, and Firm Value (March 04, 2026)*, 2026.

[5] J. Bena and K. Li, "Corporate innovations and mergers and acquisitions," *The Journal of Finance*, vol. 69, no. 5, pp. 1923–1960, 2014.

[6] Y. Zhao, X. Bi, and Q.-P. Ma, "Predicting mergers & acquisitions: A machine learning-based approach," *International Review of Financial Analysis*, vol. 99, p. 103933, 2025.

[7] U.S. Patent and Trademark Office, "Patentsview data feed." `https://www.patentsview.org/`, 2026. Accessed: 2026-04-27.

[8] L. Kogan, D. Papanikolaou, A. Seru, and N. Stoffman, "Technological innovation, resource allocation, and growth," *Quarterly Journal of Economics*, 2017.

[9] Refinitiv., "Sdc platinum: Mergers acquisitions database [data set]," *London Stock Exchange Group*, 2026.

[10] S. G. M. Intelligence., "Compustat north america daily [data set]," *Retrieved from Wharton Research Data Services (WRDS)*, 2026.

[11] D. Barkemeyer, J. Bena, and X. Li, "Technology life cycles: Investment, growth, and firm value," *SSRN*, 2026.

A Robustness Checks & Pipeline Audits

To ensure the mathematical integrity of the embeddings, the topological mapping, and the ranking algorithms, we conducted five separate comprehensive data audits.

A.1 Gaussian Audit

Because UMAP dimensional reduction can sometimes cause dense embeddings to collapse into point masses (rendering covariance matrices non-invertible), we audited 7,170 GMM components. The results confirmed that 100% of the evaluated components exhibited a mathematically reasonable variance and 100% were non-degenerate. This validates that the 50-D UMAP space supports proper spherical clustering.

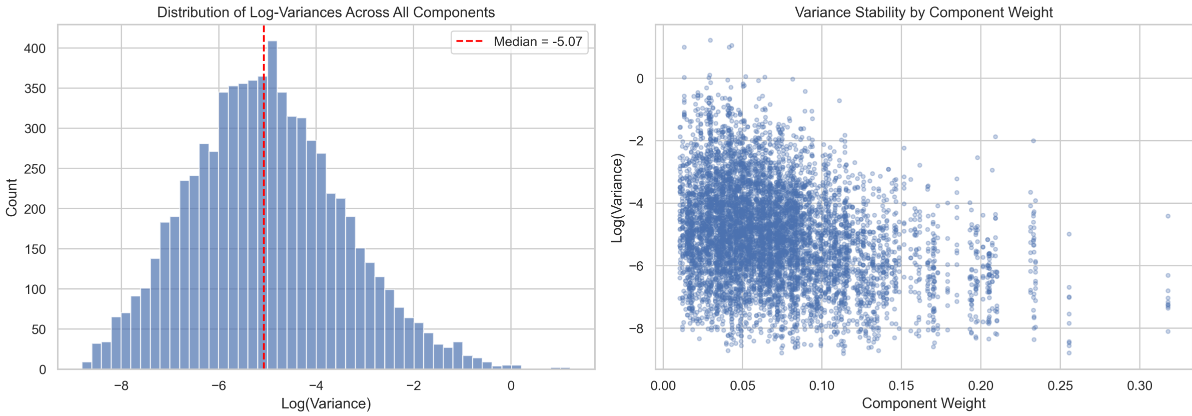

Figure 1: Gaussian Variance Check

A.2 Bootstrap Ranking Stability

A ranking model is only viable if its top recommendations are stable against minor data variations. We executed a bootstrap resampling procedure across 100 firms. The pipeline achieved a mean stability score = 97.77. Furthermore, 72 out of the 100 tested firms maintained a perfectly stable Top-1 ranking target (SD = 0.0) across all bootstrap iterations, proving the ranking hierarchy is structurally robust and free from stochastic noise.

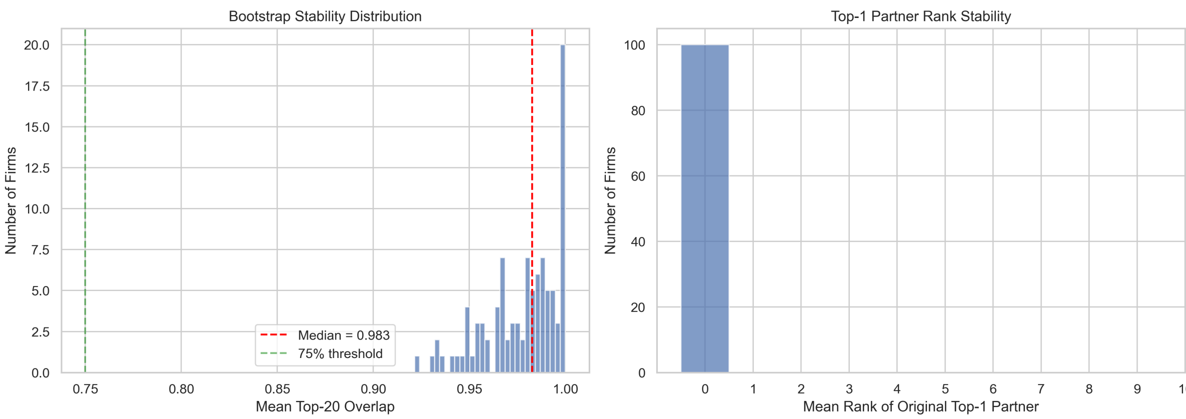

Figure 2: Ranking Stability

A.3 Prior Sensitivity: Dirichlet Process & Entropy

We audited the Dirichlet Process prior across the entire universe of 7,949 firms to ensure the model was not overfitting components. The algorithm naturally stratified the economy into two tiers: 6,304 firms were adequately modeled by a *single Gaussian* (indicating highly concentrated R&D), while 1,645 complex firms utilized the full GMM structure. For the GMM tier, the average number of active components was 10.18 (well below the $K_{max} = 15$ ceiling), with a mean entropy of 0.7375. This proves the Bayesian prior successfully and organically prunes unused dimensions.

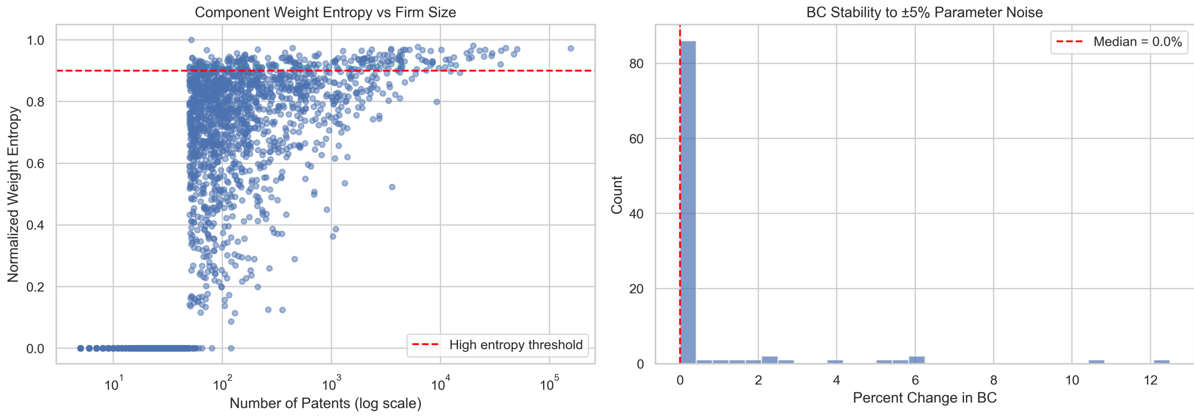

Figure 3: Prior Sensitivity Analysis

A.4 Prior Sensitivity: Hyperparameter Perturbation

To ensure the final BC synergy scores were not hyper-sensitive to Bayesian initialization parameters, we conducted a perturbation audit across 100 random firm pairs. Modifying the prior hyperparameters resulted in a mean absolute BC change of just 1.57×10^{-6} and a median percentage change of 8.35×10^{-20}. The resulting synergy matrix is mathematically sound.

A.5 External Validity & Correlation Analysis

To confirm that the topological BC score measures genuine strategic synergy, we evaluated 1,000 firm pairs against traditional metrics. The BC score demonstrated a near-perfect correlation (0.9909) with traditional patent citation overlap. Furthermore, the mean BC for firms within the same industry (SIC2 match) was 0.060, compared to 0.00068 for firms in different industries. This confirms the geometry naturally captures industry boundaries while allowing for cross-industry matching when citations are absent.

A.6 Theoretical Constraint: GMM vs. von Mises-Fisher (vMF)

A theoretically purer alternative to the Euclidean GMM would be a von Mises-Fisher mixture model, which natively respects spherical embedding data. However, the exact BC between two vMF distributions relies on the modified Bessel function of the first kind. In 50-dimensional space, computing these integrals results in catastrophic exponential overflow, making exact similarity calculations analytically intractable without computationally prohibitive Monte Carlo estimations. The Euclidean GMM remains the optimal choice.

A.7 Convergence in K_{max}

The GMM distributions converge at $K_{max} = 10$. We expand the $K_{max} = 15$ to allow for richer firm-level characterization in larger firms.

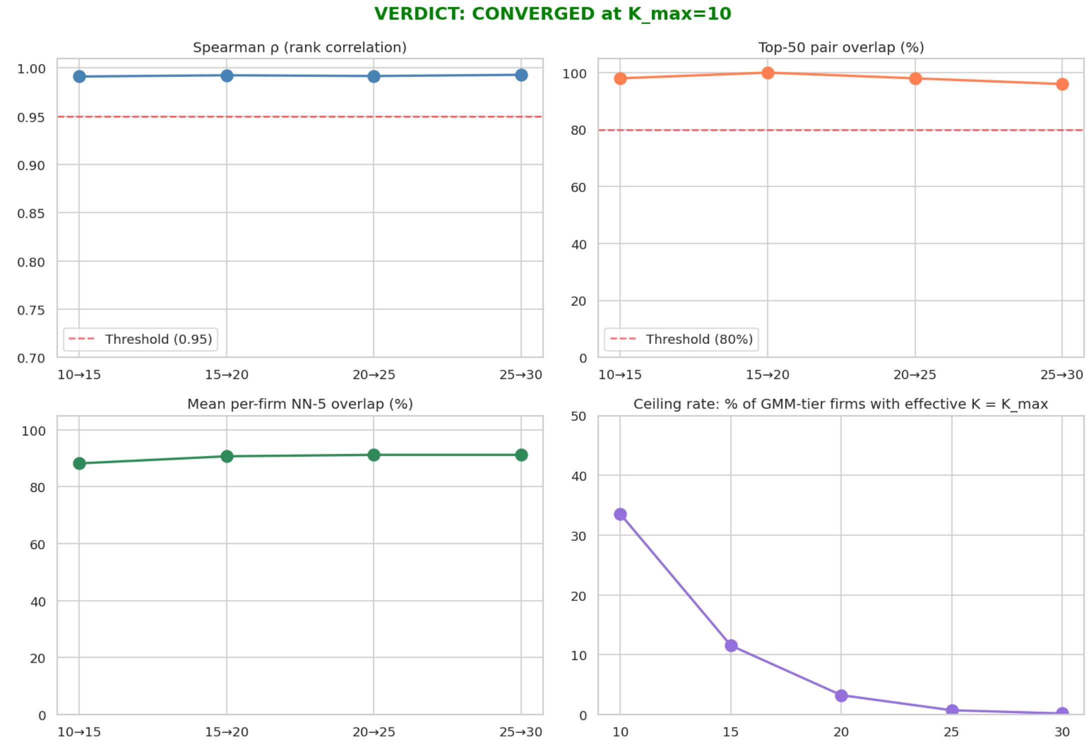

Figure 4: Analyses for K_{max}

A.8 BC Distribution

The distribution of BC over the dataset is right-tailed. This is expected as the BC is calculated for every pair and many firms are different in their technology focus.

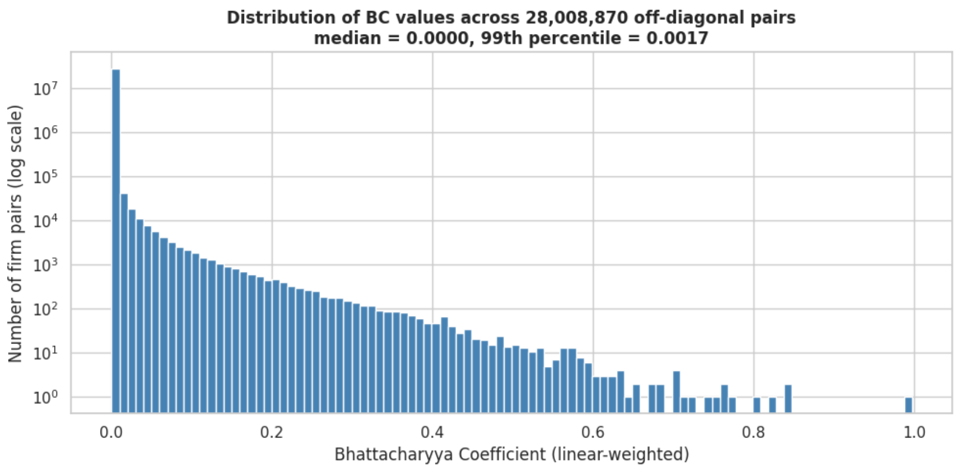

Figure 5: BC Distribution

A.9 Directional Complementarity: Formula Derivation

The intuitive formulation $\mathrm{Comp}(A \to B) = \sum_k (1 - p_{A,k}) \cdot p_{B,k}$ is algebraically equivalent to $1 - \mathrm{BC}(A, B)$ when weights sum to 1, adding no information beyond dissimilarity. The market-relative gap formulation resolves this: the gap vector $\max(0, p_{\mathrm{market},k} - p_{A,k})$ is not a normalised probability distribution, so the algebraic symmetry collapses. Validation across all 28 million firm pairs confirms genuine asymmetry, with a maximum $|M_{ij} - M_{ji}| = 0.039$.