# Technology Trajectories and the Returns to Frontier R&D: Evidence from ARPA-E

Joshua Krieger, Jacquelyn Pless

June 16, 2026

**EXTENDED ABSTRACT**

Innovation policy and R&D strategy face a fundamental allocation problem: scarce capital must be deployed not just across projects, but also across positions in the technology landscape. Some investments push into crowded areas where inventive activity, scientific attention, and private capital are already expanding. Others target mature domains where technical progress and the associated returns may be more predictable but less transformative. Still others attempt to create new technological directions in areas with little prior activity, where private incentives may be weakest but the potential social returns may be substantial. These choices around how actively managed R&D programs allocate capital across technology areas are central to the design of public R&D programs and to decisions by firms, investors, and research organizations deciding whether to follow technological momentum, differentiate from it, or build entirely new opportunities.

Despite the importance and ubiquity of this allocation problem, relatively little is known about how the returns to R&D depend on the trajectory of the technology area (or technological "neighborhood"). Existing work evaluates funded projects as isolated treatments, or characterizes them by observable recipient and project features (Howell, 2017; Myers and Lanahan, 2022; Azoulay et al., 2019). Yet similar projects may have very different consequences depending on whether they enter a growing, crowded, mature, or undeveloped technological neighborhood. Entering a rapidly expanding area may raise private returns by providing complementary knowledge, suppliers, thicker labor markets, and downstream adoption opportunities. Entering a sparse or less-developed area may generate weaker near-term measurable outcomes, but potentially larger social value if the funding opens a new technological direction. Understanding this tradeoff is essential for evaluating public innovation awards and agencies and for designing R&D portfolios, more generally.

This paper studies how the state of the technological opportunity set shapes the returns to frontier R&D funding. We ask three questions. First, where do actively managed public R&D programs choose to intervene in technology space? Do such programs allocate awards toward areas with strong prior momentum, toward mature and technologically dense areas, or toward sparse and relatively undeveloped areas where public funding may be needed to create new innovation paths? Second, how do direct project outcomes vary across these technology area trajectories? In particular, do awards in high-momentum neighborhoods generate more knowledge, invention, and follow-on financing than otherwise similar awards in less-developed or more mature neighborhoods? Third, do public R&D awards merely select into technology areas that were already improving, or do they shift the subsequent trajectory of inventive activity in the surrounding neighborhood?

We examine these questions in the context of the Advanced Research Projects Agency-Energy (ARPA-E), a U.S. Department of Energy agency tasked to fund and advance "high-potential, high-impact energy technologies that are too early for private-sector investment"[1] through actively managed programs. Actively managed research programs—where a program manager determines program topics and project portfolios—make up a substantial and growing portion of government-funded R&D. In 2025, the U.S. "ARPA" programs had public budgets exceeding $6.3 billion.[2] ARPA-E is an especially useful setting because program directors do not simply choose individual projects; they construct and manage portfolios intended to redirect and accelerate technological search within a mission-oriented domain. Empirically, that structure creates rich variation in both award-level choices and program-level portfolio strategies. Some programs concentrate in narrow, high-momentum technology areas, while others diversify across emerging, mature, and undeveloped domains. This variation allows us to study not only whether particular projects succeed, but also how the composition of an R&D portfolio shapes aggregate scientific and inventive outcomes.

Our empirical strategy involves developing a new measure of technology position that maps each ARPA-E award into a continuous technology space using Cooperative Patent Classification (CPC) embeddings constructed from co-occurrence

[1] https://www.energy.gov/advanced-research-projects-agency-energy-arpa-e

[2] The Defense Advanced Research Projects Agency (DARPA) is the largest of such agencies, with a 2025 public budget of $4.37 billion. ARPA-E (part of the Department of Energy) and ARPA-H (National Institutes of Health) had 2025 budgets of $450 million and $1.5 billion, respectively. The Intelligence Advanced Research Projects Activity (IARPA) has a classified budget. The Agriculture Advanced Research and Development Authority, part of the Department of Agriculture was authorized in 2018, but remained unfunded. Recently, the United Kingdom established its own actively managed frontier R&D agency, Advanced Research and Invention Agency (ARIA), and Germany established its Federal Agency for Disruptive Innovation (known as SPIN-D)

in patents and taxonomy distance. We classify award descriptions into a vector of CPC codes using a model developed by the European Patent Office (Hähnke et al., 2025), then calculate each award's CPC embedding distance relative to the existing patent corpus at the time of funding. We validate the approach by comparing the resulting novelty scores for the subset of awards with granted patents to the novelty of their patents and breakthrough measures from Kelly et al. (2021) (see Figure 7). Beyond a measure of backward-looking novelty, for each award, we also measure the neighborhood's prior inventive activity, recent growth, and subsequent evolution (through event studies). This approach separates two dimensions that are often conflated: whether a project is technologically novel, and whether it enters an area with strong pre-existing momentum. The Appendix provides additional technical detail on the measurement approach.

The mapping into technology space also allows us to move beyond descriptive comparisons among funded awards. We construct comparison groups from the broader corpus of energy-related technology classes and from matched or synthetic-control technology neighborhoods that were similar prior to funding but not associated with ARPA-E awards. This design allows us to compare the post-award evolution of ARPA-E-associated neighborhoods to counterfactual technology areas with similar pre-award trajectories. We use these comparisons to distinguish selection into already-growing technology areas from potential acceleration or redirection of subsequent inventive activity.

The first part of the analysis characterizes the distribution of ARPA-E awards by their technology novelty and trajectories using similarity to the patent corpus. Our descriptive analysis reveals substantial heterogeneity across ARPA-E awards and programs in both technological novelty (see Figures 1–3) and technology area trajectory (see Figures 5) and 6. Across multiple versions of our novelty measures, we find the distribution of ARPA-E awards is more novel than the technologies represented in a random sample of patents from the same broad energy technology fields.

Next, the analysis estimates how direct project outcomes vary by project novelty and their technology area trajectories. We examine project outcomes like patents, publications, citations, and venture funding. By comparing direct outcomes across technology area types, we can ask how the private and measurable returns to R&D funding depend on the state of the surrounding technology area at the time of award. We find that patenting activity is decreasing in award novelty, while scientific publishing increases in novelty (see Figure 7). For the subset of startup funding recipients, we find that project subsequent private capital funding for projects is decreasing in novelty. Controlling for overall novelty, our preliminary results show that

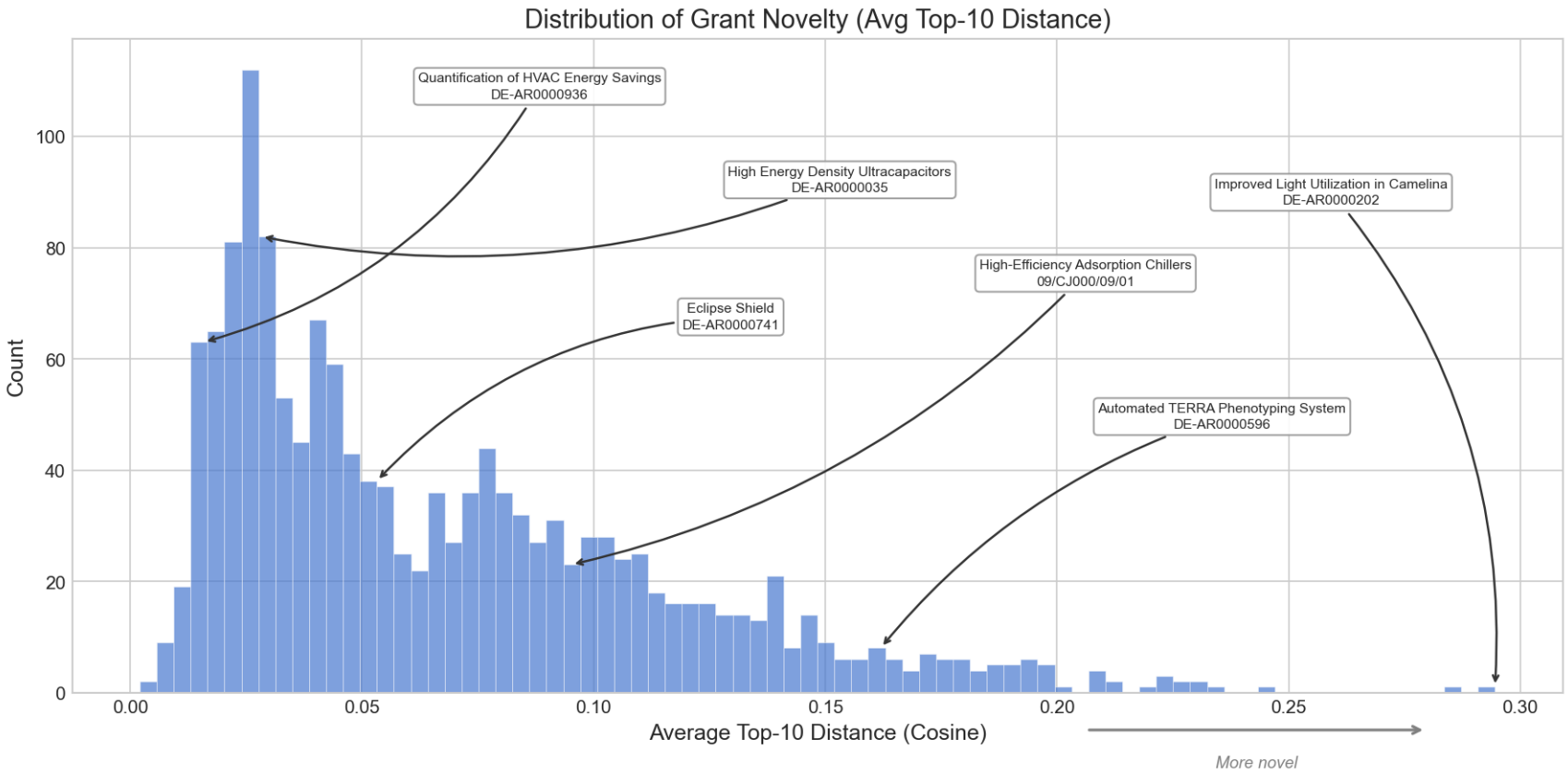
greater "momentum" in a technological area prior to awards is associated with more publications per funded project but is only weakly associated with more patenting.

Third, we evaluate technology-area-level and spillover outcomes. We compare the post-award evolution of ARPA-E-associated technology neighborhoods to the pre-award trend and against control groups based on comparable patents' technology neighborhoods (see Figures 8– 11). We observe substantial heterogeneity in the relationship between pre-award momentum and post-award inventive activity. We find that roughly half of awards go to technology areas with stable/stagnant activity leading up to the award, and half of those experience a substantial increase in inventive activity after the award. By contrast, the other 50% of awards had crowding into the neighborhood prior to the ARPA-E award, yet only half of those saw continued entry into the relevant technology neighborhood. In a differences-in-differences specification, we find that ARPA-E projects' technology fields experience substantially more crowding in, on average, than control technology areas, even after adjusting for technology novelty (see Table 1 and Figure 11). In ongoing analyses, we are evaluating technology area trends vs. synthetic control groups and how much of the crowding in is due to new entrants vs. incumbents increasing their activity.

Finally, we examine how these award-level choices combine into program portfolios. ARPA-E programs vary not only in technological diversity, but also in the diversity of technology area trajectories represented by their awards. Some programs concentrate many projects in a single rapidly growing or emerging area, while others combine projects across mature, emerging, and undeveloped areas (see Figures 12 and 13). We measure both technological diversity and trajectory diversity within programs, then compare program-level outcomes including total patents, publications, and growth in associated technology areas.

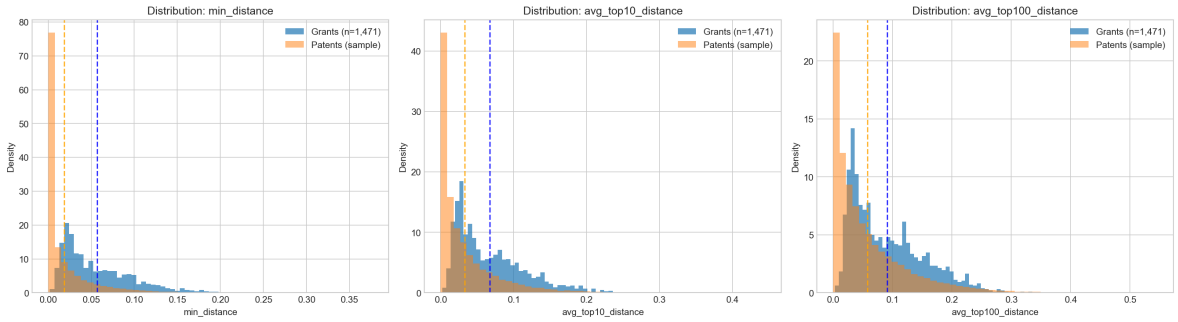
This paper contributes to the economics and management of innovation by introducing technology-neighborhood trajectories as a way to evaluate R&D returns. The results clarify a core allocation problem for public agencies, firms, and investors: whether to follow technological momentum, exploit mature domains, or create opportunities in sparse areas where private incentives may be insufficient. By linking award novelty, neighborhood momentum, spillovers, and portfolio structure, the paper shows how an R&D investment's position in technology space shapes both measurable project outcomes and broader innovation trajectories.

**Figure 1. Distribution of Grant Novelty (Avg. Top-10 Distance)**



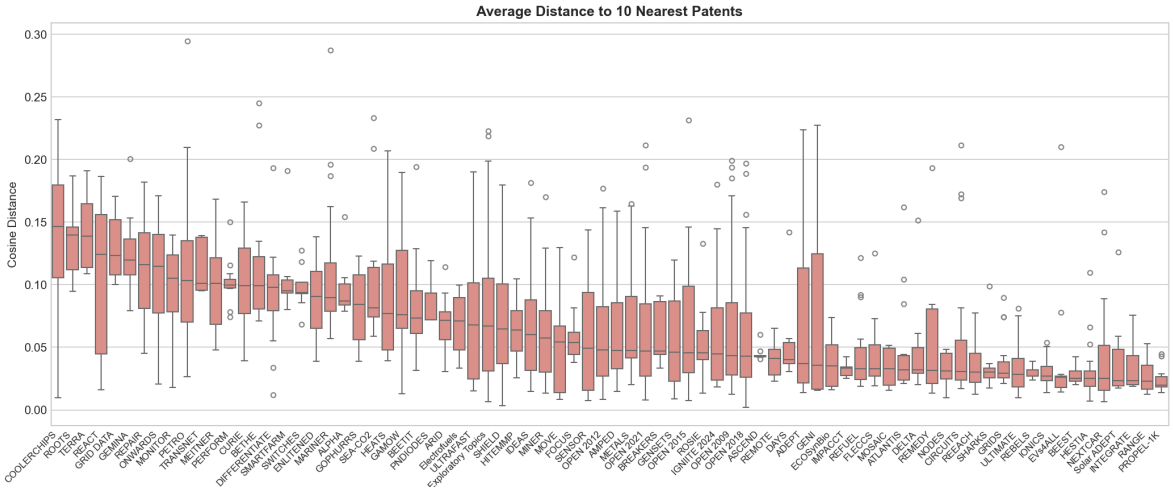

The distribution of ARPA-E award novelty measured by average cosine distance to the 10 nearest patents. Representative projects are annotated at the 5th, 25th, 50th, 75th, and 95th percentiles to illustrate the range of technology novelty across the portfolio.

**Figure 2. Distribution of Grant Novelty vs. Patent Novelty for Energy Patent Corpus**



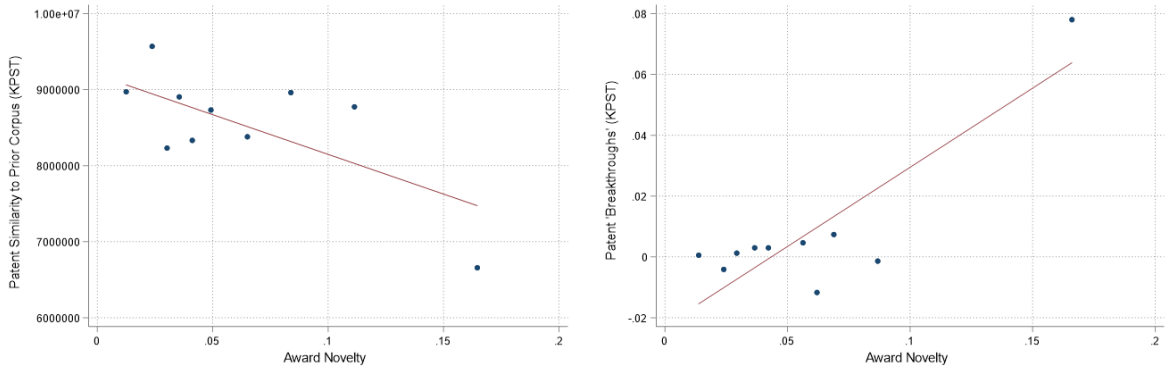

Note: Distribution of distance-based novelty metrics for ARPA-E awards (blue) versus the patent corpus (orange). The patent corpus, in our case, is the set of all USPTO patents submitted on or after January 1, 2000, that share at least one CPC class with one of the ARPA-E awards. Across all three metrics, awards are shifted rightward, indicating that ARPA-E projects occupy more novel regions of the technology space than typical patents.

**Figure 3. Program-Level Novelty Distribution Variation**



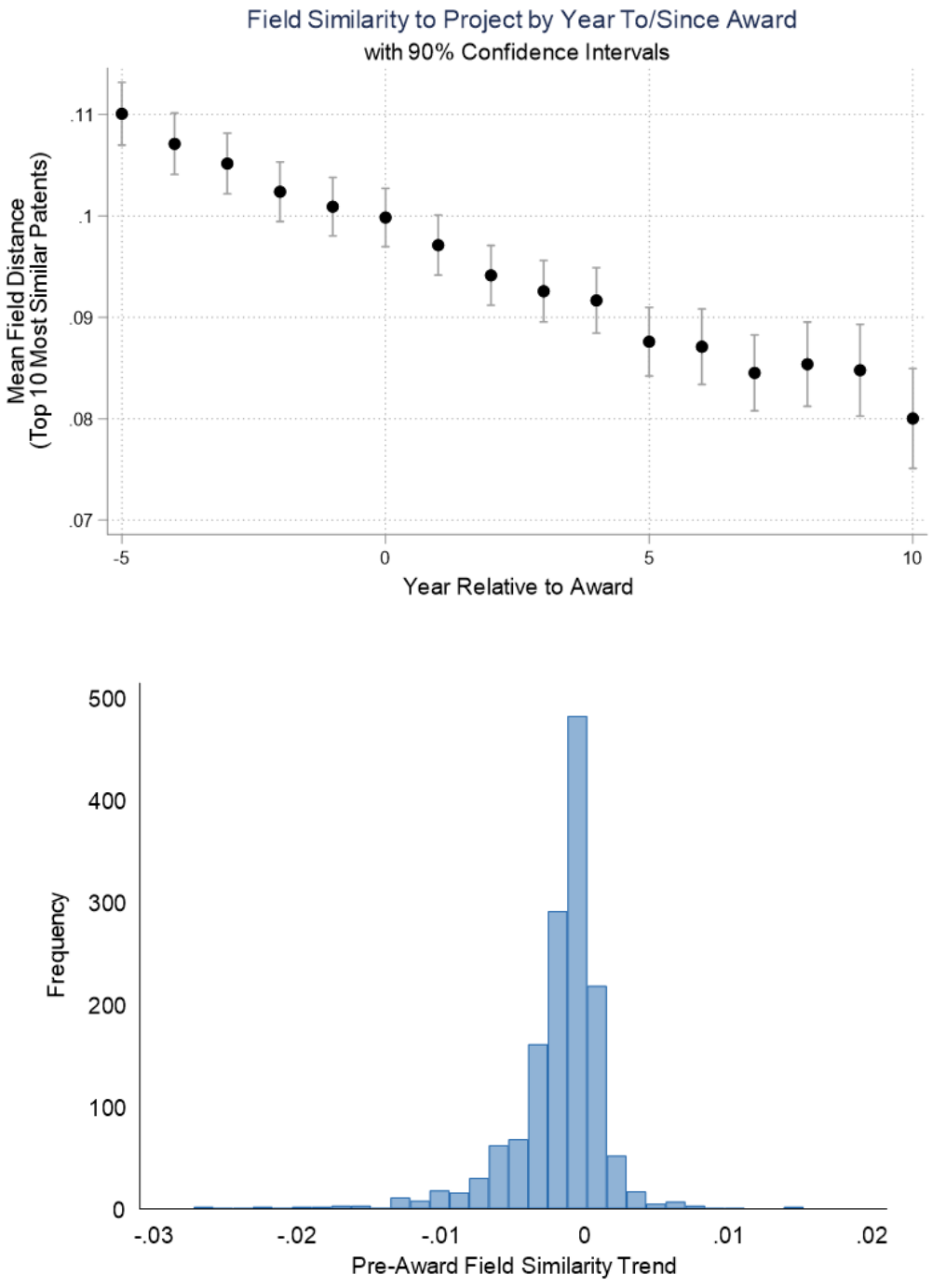

Note: Box plots of novelty metrics across all ARPA-E programs. Programs are sorted by median novelty.

**Figure 4. Novelty Validation: Award Patent Novelty and KPST Patent Text Similarity**




Note: The binscatter plots compare the 783 ARPA-E patents (from 274 awards) that overlap with the KPST dataset (Kelly et al., 2021). The first panel shows the negative correlation between our measure and the KPST backward patent text similarity measure. More novel awards, according to our CPC embedding method, generate patents that are less similar to the previous patent corpus. The right panel shows that the patents from the top decile of award novelty are more than eight times more likely to generate a "breakthrough" patent (as determined by the KPST patent text measures).

**Figure 5. Tech Neighborhood Award-Preceding Trajectories: Overall and Distributions**




Panel A shows the average neighborhood technology similarity by years relative to ARPA-E awards. The analysis is based on OLS regression of panel data for the full set of awards. Neighborhood similarity is calculated as the average patent similarity (based on our CPC embedding method) for the top 10 most similar patents to the award in each year. Panel B shows the distribution in the slope of pre-award neighborhood similarity in the 5 years leading up to an ARPA-E award

**Figure 6. Tech Neighborhood Award-Preceding Trajectories Compared to Control Patents' Pre-Grant Trend**

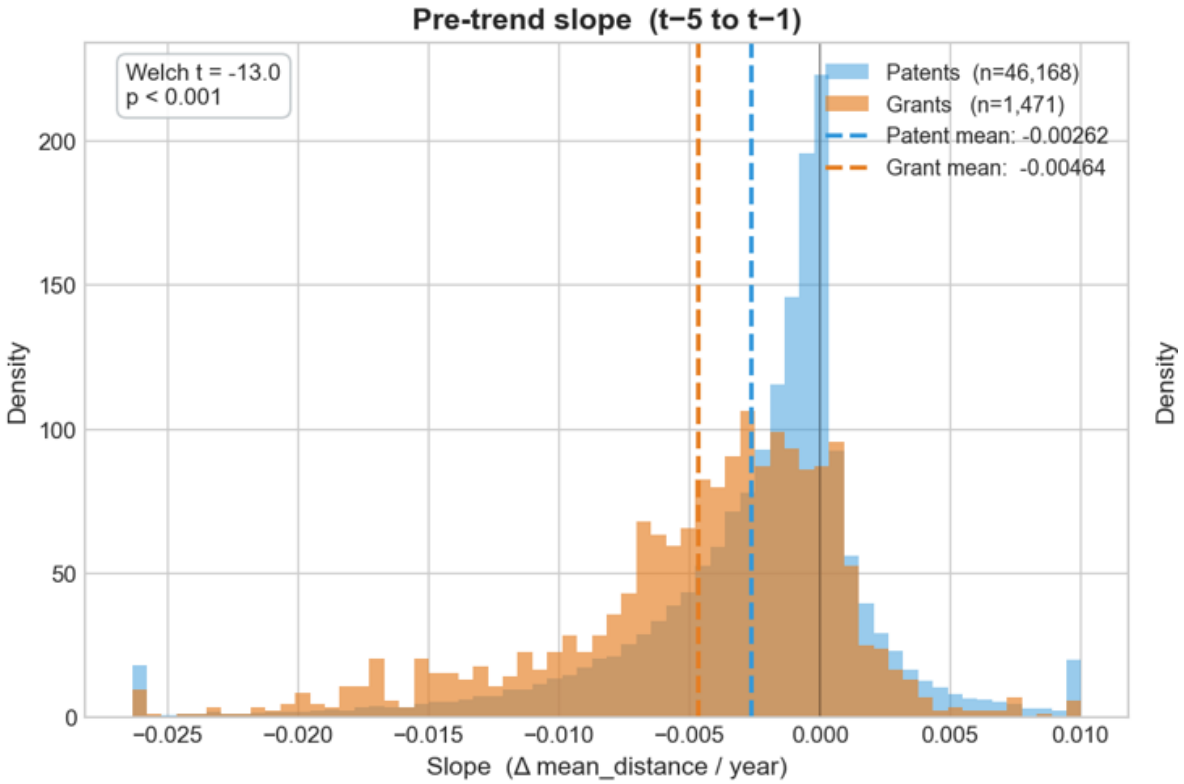
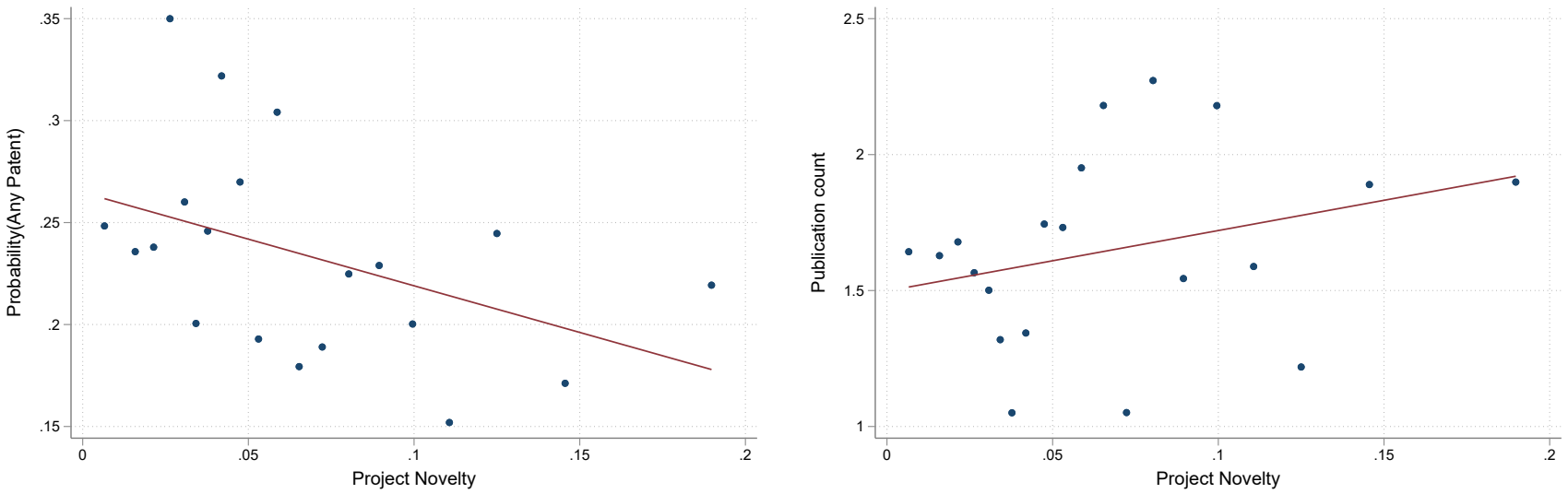


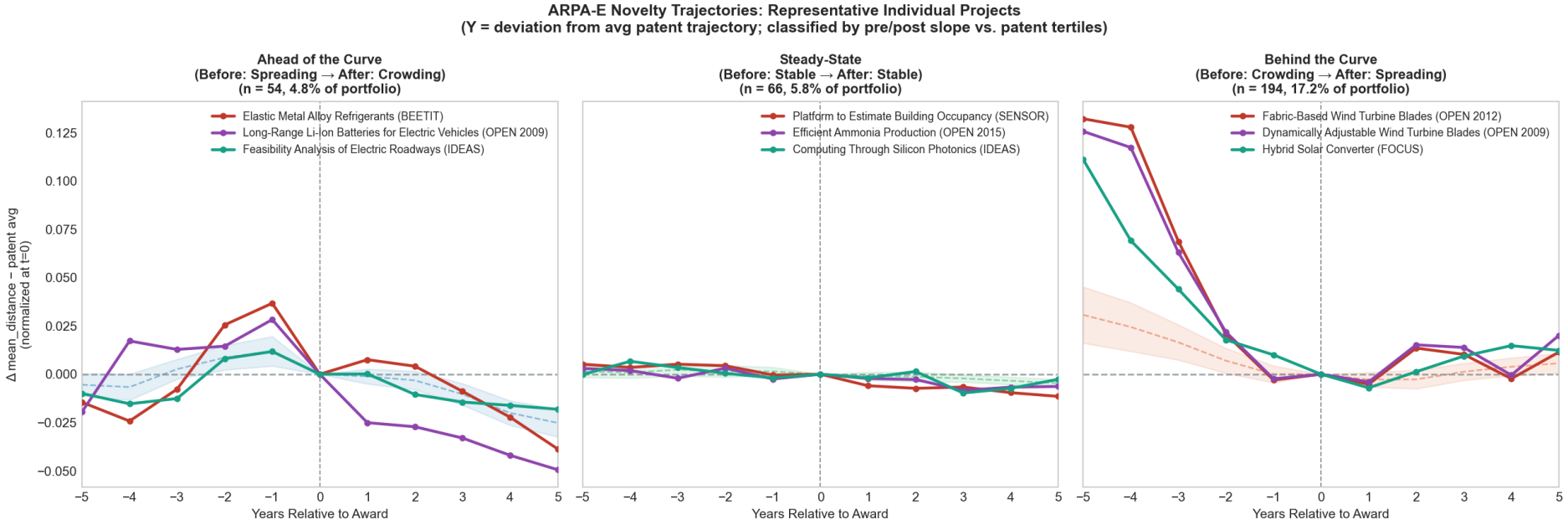

Figure 6 shows the histograms of the technology areas slope in the years $t = -5$ to $t = -1$ prior to ARPA-E grant award or patent issue. The orange series represents the ARPA-E awards and the blue series corresponds to the patent control group, which is comprised of patents from the same broad technology areas (mostly energy-related) as the ARPA-E projects.

**Figure 7. Patents and Publications, by Award Novelty**



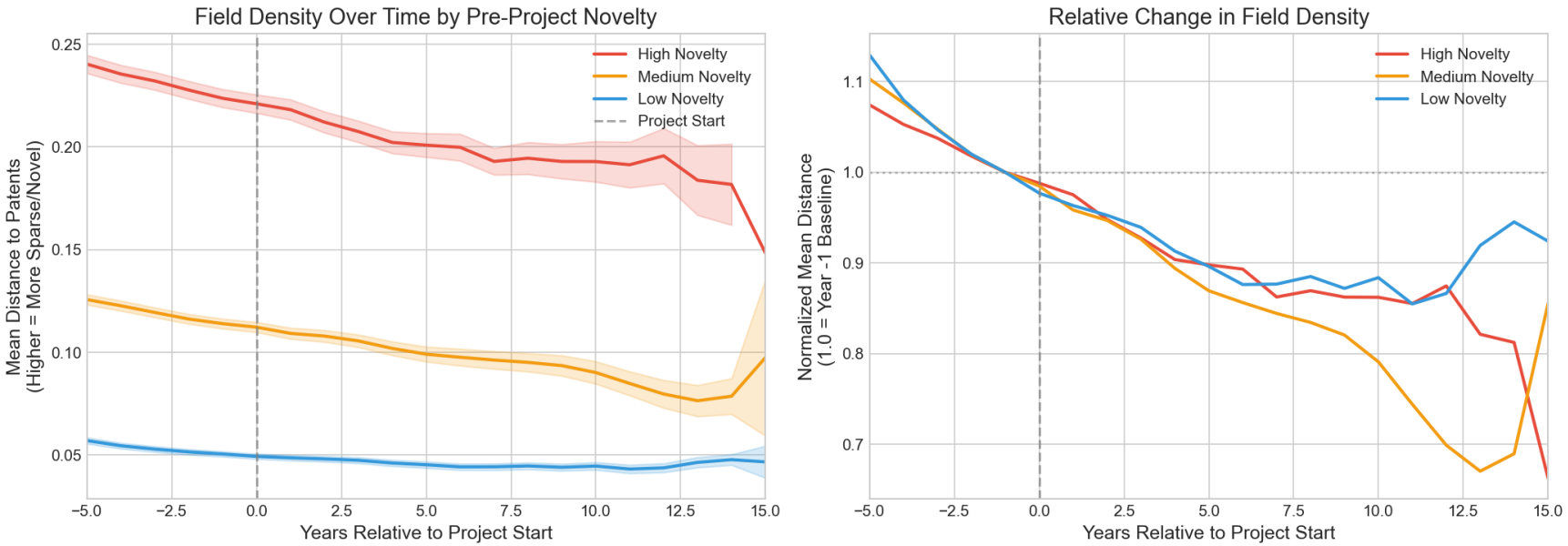

Note: Binscatters control for award start year, organization type (commercial or academic) and broad technology areas.

## Figure 8. Tech Neighborhood Trajectories: Three Types Illustrated



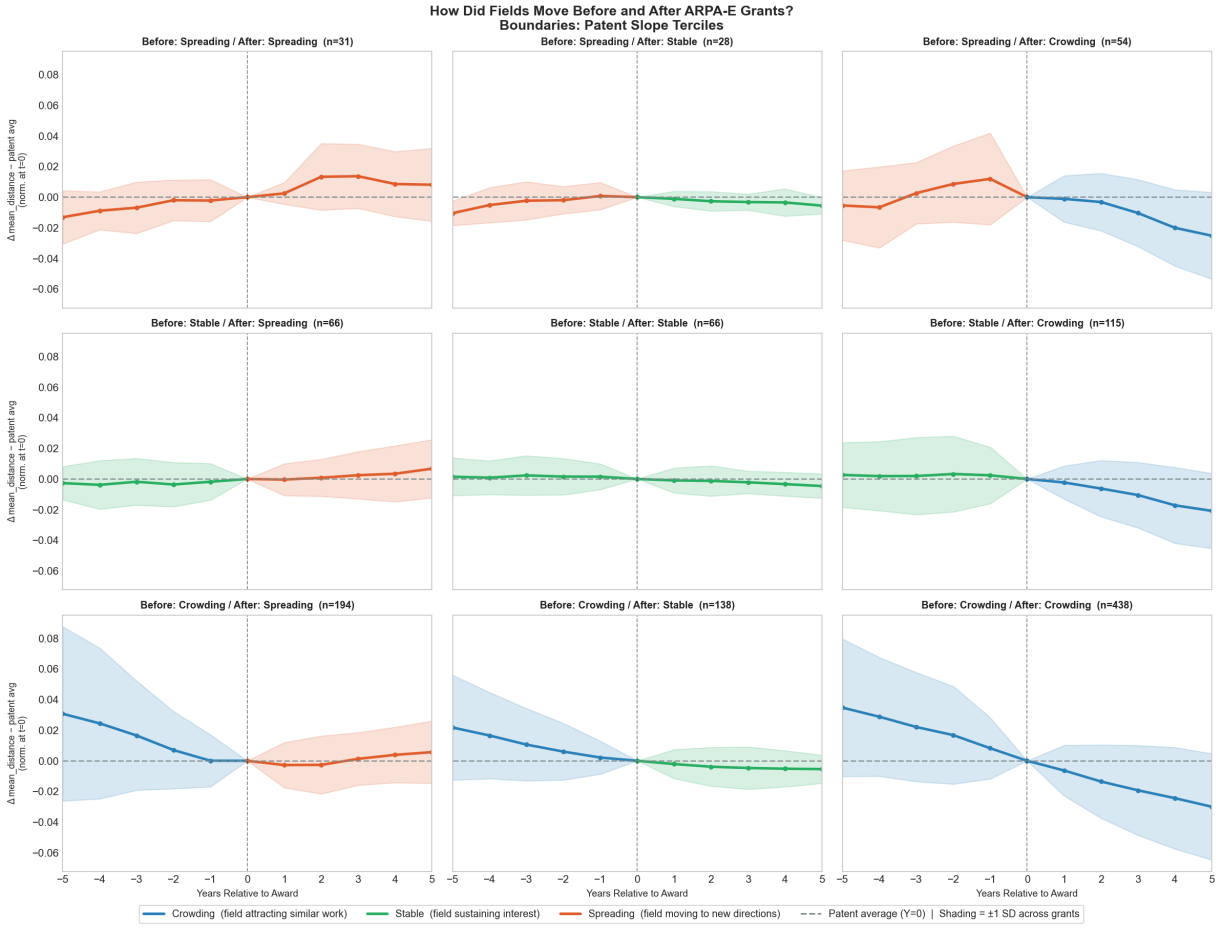

Note: Figure 8 illustrates distinct technology neighborhood trajectories. Each panel displays the patent corpus similarity trajectories for three different ARPA-E projects, after subtracting the average similarity for the patent control group. The left panel shows three projects that, relative to the patent control group, had "spreading" patent similarity—i.e., patents in the years leading up to the awards were getting more distant to the projects' topics—followed by a pattern of "crowding"—patents become more similar to the focal technology—after the award. The middle panel shows three projects with relatively stable before and after trajectories, after demeaning by the control group's trends. The right panel shows three projects that had substantial crowding in prior to the award, and then a reversal in activity trend after the award.

**Figure 9. Tech Neighborhood Trajectories, by Novelty Tercile**




Note: Panel A (left) shows the average distance between the award's technology area and the corpus of patents in each year before/after the ARPA-E award, split by novelty tercile. Novelty is calculated at the time of award, using all prior patents. Panel B (right) normalizes each tercile's mean distance to 1 at year $t = -1$ in order to show the relative change in technology areas density before and after awards, by novelty tercile. Novelty and temporal distance measures both use the CPC embedding methodology that accounts for both topic co-occurance and taxonomic similarity.

## Figure 10. Tech Neighborhood Trajectories, by Novelty Tercile




Note: The 9 panels divide the sample by tercile of technology area trend prior to the award (vertical axis) and post-award (horizontal axis). All panels show the technology areas patent activity in the years before/after an award, based on the CPC embedding distance between the focal award and the corpus of energy technology relevant patents in that year, after de-meaning based on the control patent's average trends.

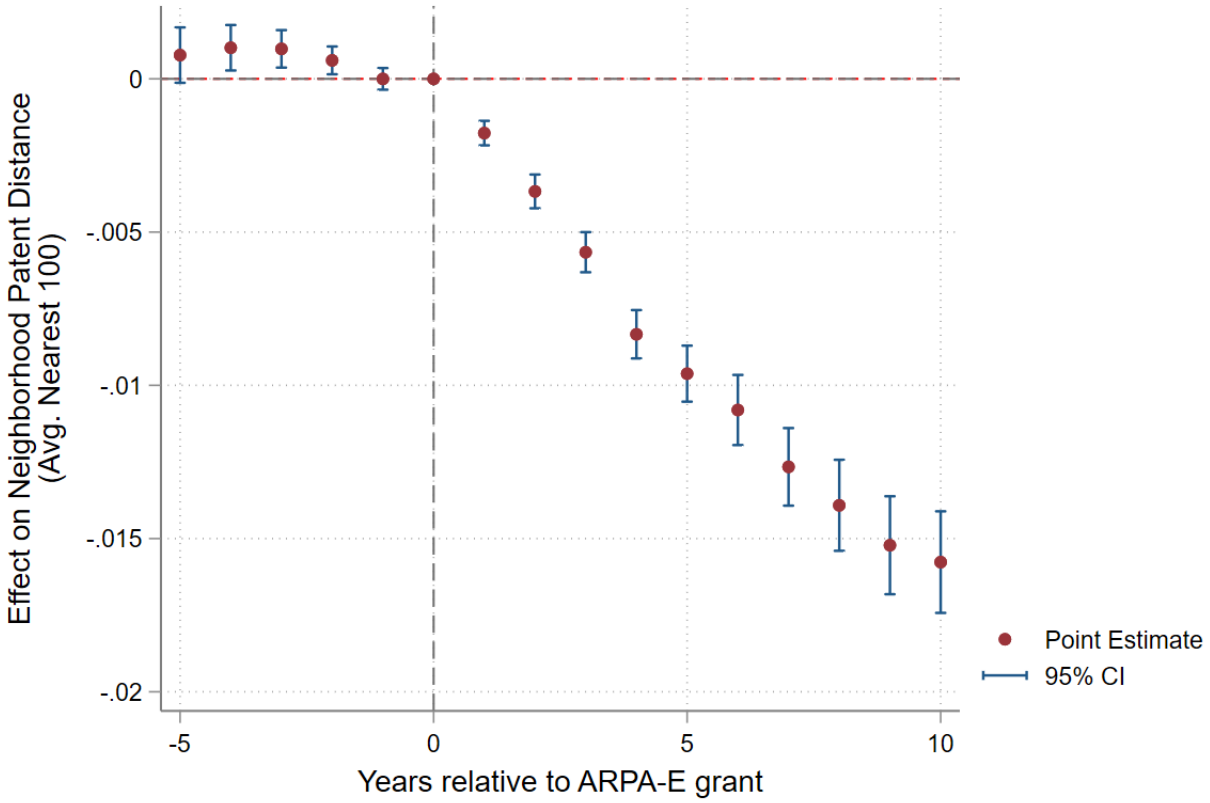
**Table 1.** Post ARPA-E Award Technology Neighborhood Average Distance, Compared to Energy Patent Controls

| | (1) | (2) | (3) |
|---|---|---|---|
| VARIABLES | Tech. Area Distance | Tech. Area Distance | Tech. Area Distance |
| ARPA-E Award | 0.0410*** | -0.0075*** | |
| | (0.002) | (0.001) | |
| Post-Award/Patent | 0.0064*** | -0.0109*** | -0.0049*** |
| | (0.001) | (0.000) | (0.000) |
| Post × ARPA-E | -0.0194*** | -0.0059*** | -0.0095*** |
| | (0.002) | (0.001) | (0.001) |
| Tech. Area Novelty | | 1.1796*** | |
| | | (0.005) | |
| Constant | 0.1025*** | 0.0337*** | 0.1098*** |
| | (0.001) | (0.000) | (0.000) |
| Observations | 1,397,628 | 1,346,803 | 1,346,803 |
| $R^2$ | 0.010 | 0.755 | 0.959 |
| Year FE | Yes | Yes | Yes |
| Award/Patent FE | No | No | Yes |

Robust standard errors in parentheses, clusted at the patent/award level, *** p<0.01, ** p<0.05, * p<0.1

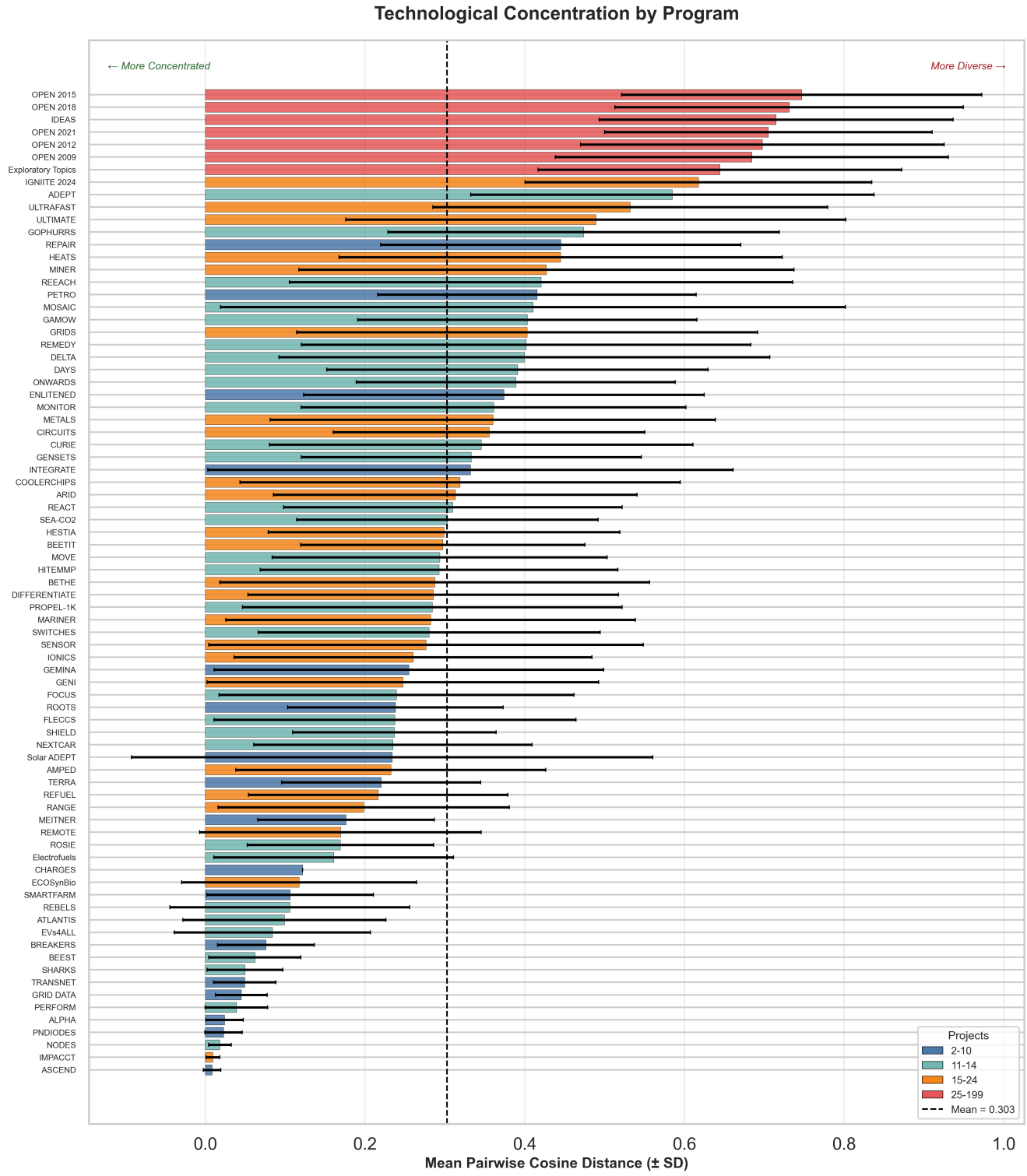
Note: Table 1 uses OLS on an unbalanced panel of technology areas associated ARPA-E awards (n=1,471) and control group patents (n=50,000). The outcome variable is the average distance between the focal technology area (from the ARPA-E project award or control patent) to the 100 most similar patents in that year. Each panel starts at $t = -5$ years until the award start or patent grant, and extends post-award until the end of the sample (max: $t = +25$).

**Figure 11. Tech Neighborhood Trajectory Event Study**



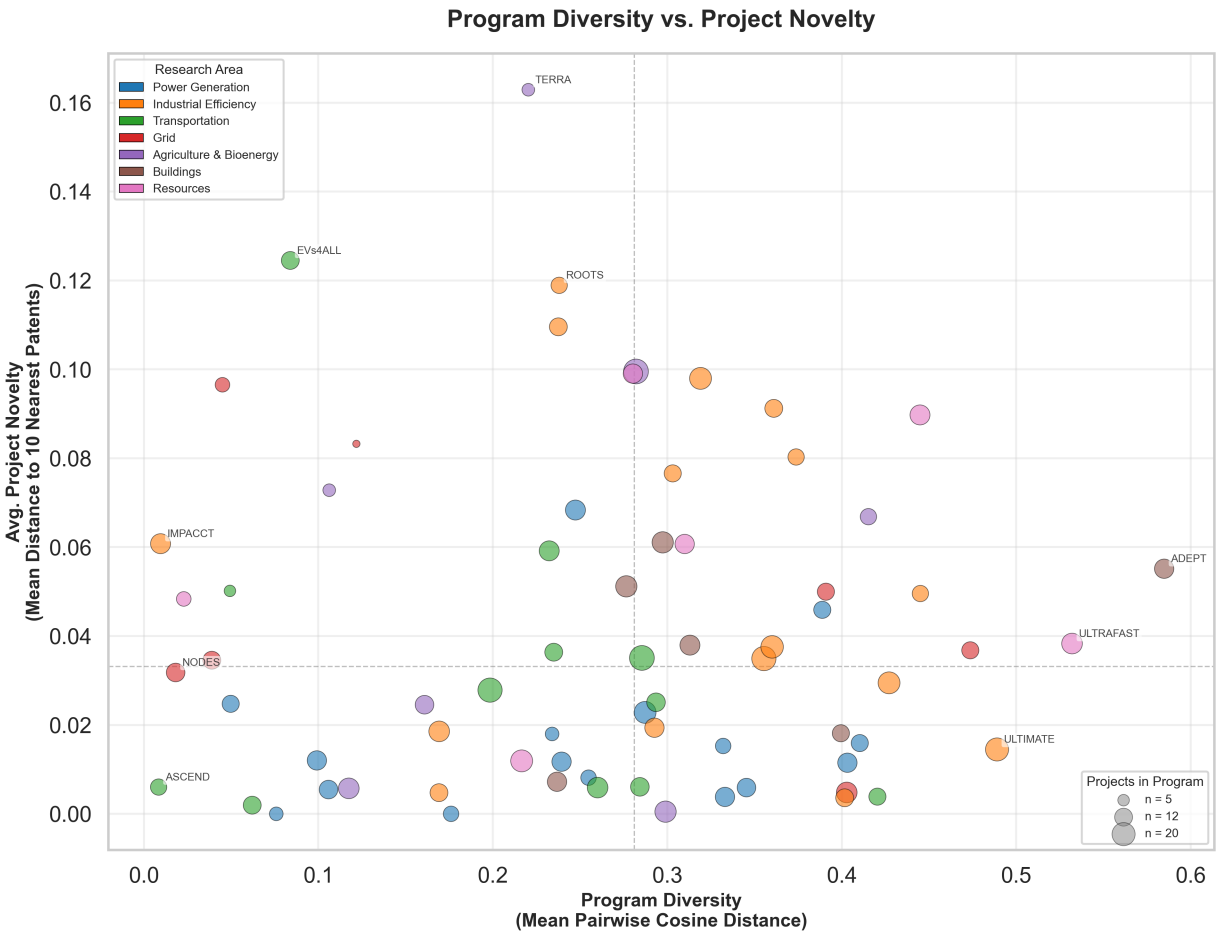

Note: Figure 11 uses OLS on an unbalanced panel of technology areas associated ARPA-E awards (n=1,471) and control group patents (n=50,000). The outcome variable is the average distance between the focal technology area (from the ARPA-E project award or control patent) to the 100 most similar patents in that year. The event study graphs each year to/since the relevant grant year interacted with whether the focal technology area is associated with an ARPA-E award.

## Figure 12. Program Diversity/Concentration




Note: Figure 12 shows the level of diversity for each ARPA-E program, as calculated by the Shannon entropy measure of diversity for CPC embedding technology areas for projects in that program.

**Figure 13. Variation in Program Diversity and Avg. Novelty**




Note: Figure 13 depicts each program's location in diversity (measured as the Shannon Entropy measure of the program's CPC embeddings) and avg. project novelty. Colors reflect ARPA-E's 7 high-level technology classes.

# References

Azoulay, P., J. S. Graff Zivin, D. Li, and B. N. Sampat (2019). Public r&d investments and private-sector patenting: evidence from nih funding rules. The Review of economic studies 86(1), 117–152.

Hähnke, V. D., A. Wéry, M. Wirth, and A. Klenner-Bajaja (2025). Encoder models at the european patent office: Pre-training and use cases. World Patent Information 81, 102360.

Howell, S. T. (2017). Financing innovation: Evidence from r&d grants. American economic review 107(4), 1136–1164.

Kelly, B., D. Papanikolaou, A. Seru, and M. Taddy (2021). Measuring technological innovation over the long run. American Economic Review: Insights 3(3), 303–320.

Myers, K. R. and L. Lanahan (2022). Estimating spillovers from publicly funded r&d: Evidence from the us department of energy. American Economic Review 112(7), 2393–2423.

# Appendix

## 1 Technology Embedding Space Construction

### 1.1 Motivation and Overview

Measuring the novelty of a research project requires a representation that is comparable across all projects, regardless of their downstream outputs. Traditional approaches such as analyzing the novelty of resulting patents or publications are inherently limited. Many projects produce no publications, no patents, or neither. We therefore require a method that enables comparison between projects on parameters that are constant and consistent across the full portfolio.

This leads us to the award submission text itself, which describes each project's proposed technological innovation. By analyzing this text, we compare projects on a common basis rather than on heterogeneous outputs. However, we do not care whether the *written text* is semantically novel—we care whether the *technology* it describes is novel. To bridge this gap, we use the European Patent Office's (EPO) CPC classifier (Hähnke et al., 2025) to assign the most relevant Cooperative Patent Classification codes to each project's text. This yields a set of discrete, detailed technology categories for every award.

To assess whether a given combination of technology categories is unique, we must compare it against a baseline. The patent corpus provides this baseline: by measuring the relative rarity of a project's assigned CPCs against the broader patent landscape, we can quantify how unusual the technology is in aggregate. However, a naïve frequency-based approach (e.g., straight TF-IDF over CPC codes) is insufficient. The hierarchical, taxonomic structure of the CPC system means that two codes may individually be rare in the corpus, yet a closely related code that differs only at the subgroup level may be very common and frequently co-assigned with the other code in the pair. A simple frequency measure would overstate the novelty of such combinations.

This is the motivation for the embedding space. By constructing a vector space that accounts for both CPC co-occurrence frequency *and* taxonomic proximity, we

obtain a robust measure of how unique a given project is compared to existing technologies. Additionally, we restrict the patent corpus to patents sharing at least one two-digit CPC class with any ARPA-E award, so that the dimensions of the embedding space are not consumed by technologies that would never be spatially proximate to any project in our sample.

The pipeline proceeds in four stages: (1) proposal text is fed to the EPO CPC classifier, which returns CPC codes with importance scores; (2) CPC codes are embedded into a hybrid vector space using co-occurrence patterns (70% weight) and taxonomic structure (30% weight); (3) award- and patent-level vectors are computed as (weighted) averages of their constituent CPC embeddings; and (4) novelty is measured as cosine distance between award vectors and the prior patent corpus.

### 1.2 CPC Classification of Awards

Each award's proposal text is classified using the EPO CPC classifier, which returns a ranked list of CPC subgroups with associated importance scores. After applying cutoff rules based on score and rank, each award is assigned a set of relevant CPC codes. Across 1,027 analyzed awards, the mean number of assigned CPC subgroups is 3.48 (SD = 1.00), the median is 3, and the range is 1–6. The importance scores returned by the classifier are retained and used as weights when aggregating CPC-level embeddings to the award level (Section 1.6).

### 1.3 Co-occurrence Embeddings (70% Weight)

**Corpus construction.** The patent corpus consists of all patents from the year 2000 onward that share at least one two-digit CPC code with any ARPA-E award. Each patent is represented as a "document" of space-separated CPC group codes. After filtering, the corpus contains approximately 750,000+ patents.

**TF-IDF.** We construct a term-frequency inverse-document-frequency (TF-IDF) matrix from the patent-CPC documents. The vectorizer uses `min_df=1,max_features=200,000`, and raw CPC tokens (no lowercasing, whitespace tokenization). For a CPC token $c$ in patent document $p$, the TF-IDF weight is:

$$\text{tfidf}(c, p) = \text{tf}(c, p) \times \log\left(\frac{|\mathcal{P}|}{|\{p' \in \mathcal{P} : c \in p'\}|}\right) \tag{1}$$

where $\mathcal{P}$ is the set of all patent documents and $\text{tf}(c, p)$ is the frequency of token $c$ in document $p$.

**Dimensionality reduction.** We apply truncated SVD to reduce the TF-IDF matrix to 300 dimensions (`random_state=42`):

$$X_{\text{reduced}} = \text{TruncatedSVD}_{300}(X_{\text{tfidf}}) \tag{2}$$

Each patent is now represented by a 300-dimensional vector encoding its CPC co-occurrence patterns.

**CPC-level embeddings.** We construct CPC-level co-occurrence embeddings by averaging the SVD vectors of all patents containing each CPC code:

$$e_c^{\text{cooccur}} = \frac{1}{|\mathcal{P}_c|} \sum_{p \in \mathcal{P}_c} x_p^{\text{svd}} \tag{3}$$

where $\mathcal{P}_c$ is the set of patents containing CPC code $c$ and $x_p^{\text{svd}}$ is the 300-dimensional SVD vector for patent $p$.

### 1.4 Hierarchical Taxonomy Embeddings (30% Weight)

The hierarchical component encodes each CPC code's position in the CPC taxonomy. For each CPC group code, we construct a one-hot vector across three concatenated levels: section, class, and subclass. Formally, for CPC code $c$ with section $s(c)$, class $\kappa(c)$, and subclass $\sigma(c)$:

$$h_c = \left[ 1_{s(c)} \; ; \; 1_{\kappa(c)} \; ; \; 1_{\sigma(c)} \right] \tag{4}$$

where $1_{s(c)}$ is a one-hot vector over all observed sections, and similarly for classes and subclasses. The resulting binary matrix is reduced via truncated SVD to $\min(100, \, n_{\text{features}} - 1)$ dimensions (`random_state=42`):

$$e_c^{\text{hier}} = \text{TruncatedSVD}_{\min(100, \, n_f - 1)}(h_c) \tag{5}$$

This component captures taxonomic structure: CPC codes sharing section, class, or subclass receive similar embeddings even if they rarely co-occur in patents.

### 1.5 Hybrid Combination

The two embedding components are combined by first $L_2$-normalizing each component separately, then concatenating with weights 0.7 (co-occurrence) and 0.3 (hierarchical):

$$e_c = \left[ \; 0.7 \cdot \frac{e_c^{\text{cooccur}}}{e_c^{\text{cooccur}}} \;\; ; \;\; 0.3 \cdot \frac{e_c^{\text{hier}}}{e_c^{\text{hier}}} \; \right] \tag{6}$$

The resulting hybrid embedding is typically $\sim$400-dimensional (300 from co-occurrence + 100 from hierarchy), though the exact dimensionality of the hierarchical component depends on the number of unique sections, classes, and subclasses observed.

### 1.6 Award-Level and Patent-Level Vectors

**Award vectors.** Each award's vector is the weighted average of its CPC embeddings, using the EPO classifier importance scores as weights, followed by unit normalization:

$$v_g = \frac{\tilde{v}_g}{\tilde{v}_g}, \qquad \tilde{v}_g = \frac{\sum_{c \in \mathcal{C}_g} w_c \, e_c}{\sum_{c \in \mathcal{C}_g} w_c} \tag{7}$$

where $\mathcal{C}_g$ is the set of CPC codes assigned to award $g$ and $w_c$ is the classifier importance score for code $c$.

**Patent vectors.** Each patent's vector is the simple (unweighted) average of its CPC embeddings, followed by unit normalization:

$$v_p = \frac{\tilde{v}_p}{\tilde{v}_p}, \qquad \tilde{v}_p = \frac{1}{|\mathcal{C}_p|} \sum_{c \in \mathcal{C}_p} e_c \tag{8}$$

**Distance metric.** Because all vectors are unit-normalized, the cosine distance between any two vectors simplifies to:

$$d(a, b) = 1 - a \cdot b \tag{9}$$

This distance has a theoretical range of $[0, 2]$. In practice, distances between awards and their nearest patent neighbors range from approximately 0.00 to 0.33, while distances to the corpus centroid range from approximately 0.42 to 0.94.

### 1.7 Limitations of the Embedding Space

1. **Classifier accuracy.** The EPO CPC classifier is designed for patent text, not award proposals. Proposal language may differ systematically from patent claims, potentially introducing classification noise.

2. **CPC granularity.** Embeddings are constructed at the CPC group level. Finer granularity (subgroup level) could capture more specific technological distinctions but would also increase noise due to sparser co-occurrence patterns.

3. **Weighting heuristic.** The 70/30 split between co-occurrence and hierarchical components was chosen based on domain judgment, not optimized empirically.

4. **Static embeddings.** The co-occurrence embeddings are computed from the full patent corpus and do not evolve over time. CPC usage patterns may shift as technology fields develop, but our embeddings do not capture this drift.

5. **Unit normalization.** Normalizing vectors to unit length discards information about the number and diversity of CPC codes assigned to each award or patent.

## 2 Static Novelty Measure

### 2.1 Definition and Motivation

The static novelty measure quantifies the degree of technological precedent for each award at the time of its funding. For each award, we compute distances to the prior patent corpus—that is, only patents published before the award's award date are included. This temporal filtering prevents forward-looking bias: an award that catalyzed a new technology area is not penalized by the subsequent patents it inspired.

### 2.2 Distance-Based Metrics

Let $\mathcal{P}_g^{<t}$ denote the set of patents published before award $g$'s award date $t$. We sort patents by distance to the award vector $v_g$ and denote the $k$-th closest patent's distance as $d_{(k)}$.

The corpus centroid is computed as:

$$\mu = \frac{\bar{v}}{\bar{v}}, \qquad \bar{v} = \frac{1}{|\mathcal{P}_g^{<t}|} \sum_{p \in \mathcal{P}_g^{<t}} v_p \tag{10}$$

**Table 2.** Distance-based novelty metrics

| Metric | Formula | Interpretation |
|---|---|---|
| `closest_1_dist` | $d_{(1)}$ | Minimum distance to nearest prior patent |
| `closest_N_avg_dist` | $\frac{1}{N}\sum_{k=1}^{N} d_{(k)}$ $N \in \{5, 10, 50, 100, 500\}$ | Mean distance to $N$ nearest prior patents |

For each metric, higher values indicate greater isolation from existing technology (higher novelty). FAISS (Facebook AI Similarity Search) with inner-product indexing is used for efficient nearest-neighbor computation; up to 501 neighbors are retrieved per query to support the `closest_500` metrics.

### 2.3 Neighborhood Metrics

In addition to distance-based metrics, we compute neighborhood characteristics for the $k$-nearest patent neighbors at $k = 100$.

**Neighborhood density.** The average pairwise cosine distance among the $k$ nearest neighbors:

$$\rho_k = \frac{2}{k(k-1)} \sum_{\substack{i,j \in N_k \\ i<j}} d(v_i, v_j) \tag{11}$$

High values indicate a sparse neighborhood (the nearby patents are themselves dissimilar to each other); low values indicate a dense, homogeneous cluster.

**Within-neighborhood percentile.** The award's rank by distance from the neighborhood centroid, expressed as a percentile:

$$\text{percentile}_g = 100 \times \frac{|\{p \in N_k : d(v_p, \mu_k) < d(v_g, \mu_k)\}|}{k} \tag{12}$$

where $\mu_k$ is the unit-normalized mean of the $k$ neighbor vectors. A high percentile indicates a peripheral position (the award is farther from the neighborhood center than most of its neighbors).

**CPC diversity.** The number of unique CPC classes (3-character codes, e.g., H01) among the $k$ nearest patent neighbors.

**Size at threshold.** The count of patents within fixed distance thresholds:

$$\text{size}_{d^*} = |\{p \in N_k : d(v_g, v_p) < d^*\}|, \qquad d^* \in \{0.1, 0.2\} \tag{13}$$

### 2.4 Interpretation and Caveats

**Two types of novelty.** Distance-based metrics capture two distinct forms of novelty: (1) peripheral novelty, where an award lies at the edge of a well-populated cluster, far from the cluster center but still near specific patents; and (2) sparse-region novelty, where an award occupies a region of the embedding space with few nearby patents of any kind. The `closest_N_avg_dist` metrics (particularly $N = 10$) are recommended as the primary novelty measure, as they are robust to individual patent outliers while remaining sensitive to genuine technological isolation.

**Awards vs. patents.** ARPA-E awards are systematically more novel than the typical patent: mean award `closest_1_dist` is 0.057 versus 0.019 for patents; mean `closest_10_avg_dist` is 0.075 versus 0.033. This is consistent with ARPA-E's mission to fund high-risk, transformative research.

## 3 Temporal Novelty Measure

### 3.1 Definition and Motivation

The temporal novelty measure tracks how the patent landscape around each entity (award or patent) evolves from five years before to five years after its reference year ($t_0$). While static novelty captures *ex-ante* technological isolation, temporal novelty addresses a different question: *how did the surrounding technology field evolve after the entity appeared?* An award that catalyzed a new field should see its neighborhood become denser (more similar patents appearing) faster than the natural background rate.

The key analytic metric is **neighborhood density**: the average pairwise cosine distance among an entity's $k = 100$ nearest patent neighbors in each year ($\rho_k$, Equation 11). High neighborhood density indicates a sparse, heterogeneous neighborhood; low density indicates a crowded, homogeneous cluster. A decreasing trajectory means the technology field around the entity is converging—new patents are filling in nearby.

### 3.2 Computing Temporal Metrics

For each entity with reference year $t_0$, we compute novelty metrics against the subset of patents published in each calendar year $y \in [t_0 - 5,\, t_0 + 5]$. The relative year is $\tau = y - t_0$. In each year, the patent corpus consists only of patents from that specific calendar year (not cumulative), providing a snapshot of contemporaneous patenting activity.

We apply this procedure identically to two populations:

1. **Awards** ($n = 1{,}471$). For each ARPA-E award, the reference year $t_0$ is the project start year. The award's embedding vector $v_g$ is compared against each year's patent corpus.

2. **Patents** ($n = 50{,}000$). A stratified random sample of patents from the same corpus used to construct the embedding space. For each sampled patent, $t_0$ is its publication year. The patent's embedding vector $v_p$ is compared against each year's patent corpus (excluding itself in year $t_0$).

The patent sample serves as an external benchmark: it establishes the baseline rate at which technology neighborhoods naturally evolve, independent of any ARPA-E funding. The same distance and neighborhood metrics from Section 2.2 are computed per year for both populations: nearest-neighbor distances, centroid distance, neighborhood density, within-neighborhood percentile, CPC diversity, and size-at-threshold counts.

### 3.3 Pre/Post Slope Computation

To summarize each entity's temporal trajectory, we fit OLS regressions of neighborhood density on relative year separately for the pre-award and post-award windows:

$$\rho_k(\tau) = \alpha + \beta_{\text{pre}} \cdot \tau, \qquad \tau \in [-5, -1] \tag{14}$$

$$\rho_k(\tau) = \alpha + \beta_{\text{post}} \cdot \tau, \qquad \tau \in [+1, +5] \tag{15}$$

A negative slope indicates the neighborhood is becoming denser (field converging); a positive slope indicates it is becoming sparser (field diverging). The pre-slope captures how the technology area was evolving before the entity appeared; the post-slope captures evolution after.

Slopes are computed for all entities with at least two valid observations in the respective window. This yields slopes for 1,471 awards (pre) and 1,263 awards (post), and 47,459 patents (pre) and 44,723 patents (post).

### 3.4 Patent-Benchmarked Normalization

A naïve approach would normalize each award's slope relative to the average award slope. However, this is circular: the mean of the award distribution is mechanically the reference point, preventing any assessment of whether awards as a group differ from the natural baseline. Instead, we use the average patent slope as the external benchmark.

**Table 3.** Mean and standard deviation of temporal slopes by population

| | Pre-slope ($\tau = -5$ to $-1$) | | Post-slope ($\tau = +1$ to $+5$) | |
|---|---|---|---|---|
| Population | Mean | SD | Mean | SD |
| Patents ($n \approx 47{,}000$) | $-0.00294$ | 0.00979 | $-0.00001$ | 0.00842 |
| Awards ($n \approx 1{,}300$) | $-0.00276$ | 0.00734 | $-0.00028$ | 0.00821 |

**Benchmark values.** Two patterns are notable. First, the pre-slopes are nearly identical: both patents and awards experience similar rates of field convergence before their reference year. Second, the post-slopes diverge: the average patent neighborhood is essentially static after publication ($\bar{\beta}^{\text{pat}}_{\text{post}} \approx 0$), while the average award neighborhood continues converging at $\bar{\beta}^{\text{award}}_{\text{post}} = -0.00028$, approximately 34 times faster than the patent baseline. This suggests that ARPA-E awards are, on average, associated with accelerated field densification in the years following the award.

**Subtraction normalization.** We normalize each award's slope by subtracting the corresponding patent mean:

$$\hat{\beta}_{\text{pre},g} = \beta_{\text{pre},g} - \bar{\beta}^{\text{pat}}_{\text{pre}}, \qquad \hat{\beta}_{\text{post},g} = \beta_{\text{post},g} - \bar{\beta}^{\text{pat}}_{\text{post}} \tag{16}$$

A normalized slope of zero means the award's field evolved at exactly the patent-average rate. A negative normalized post-slope means the award's neighborhood converged faster than the typical patent's neighborhood.

### 3.5 Novelty Terciles

To examine whether temporal dynamics differ by initial novelty level, we classify entities into terciles based on neighborhood density at $\tau = 0$. The tercile boundaries are computed from the patent distribution at $\tau = 0$:

**Table 4.** Novelty tercile definitions (from patent distribution at $\tau = 0$)

| Tercile | Neighborhood density range | Interpretation |
|---|---|---|
| T1 (Low novelty) | $\rho_k < 0.044$ | Dense, well-established field |
| T2 (Mid novelty) | $0.044 \leq \rho_k < 0.120$ | Moderately sparse field |
| T3 (High novelty) | $\rho_k \geq 0.120$ | Sparse, frontier field |

Awards are assigned to terciles using the same patent-derived cut points, allowing direct comparison of slope distributions within each novelty stratum.

### 3.6 Award Classification by Trajectory Shape

We classify awards into trajectory types based on how their pre- and post-slopes compare to the patent benchmark. Two classification schemes are used.

**Binary (2×2) classification.** Awards are split at the patent mean for each window:

$$\text{Pre category} = \begin{cases} \text{High} & \text{if } \beta_{\text{pre},g} \geq \bar{\beta}^{\text{pat}}_{\text{pre}} \\ \text{Low} & \text{otherwise} \end{cases} \tag{17}$$

and analogously for the post-slope. This yields four trajectory types (Table 5).

**Table 5.** 2×2 trajectory classification ($n = 1{,}263$ awards)

| Type | Pre | Post | Interpretation |
|---|---|---|---|
| High / High | Above avg | Above avg | Field was diverging and continued diverging |
| High / Low | Above avg | Below avg | Field was stable/diverging, then converged |
| Low / High | Below avg | Above avg | Field was converging, then stabilized |
| Low / Low | Below avg | Below avg | Field was converging and continued converging |

**Three-category (3×3) classification.** For finer resolution, awards are classified as High ($> \bar{\beta}^{\text{pat}} + 1\sigma$), Low ($< \bar{\beta}^{\text{pat}} - 1\sigma$), or Average (within $\pm 1\sigma$ of the patent mean).

This yields nine cells but concentrates most awards in the Average/Average cell, since $\pm 1\sigma$ from the patent mean captures the bulk of the award distribution.

### 3.7 Limitations

1. **Causality.** Temporal changes in the patent landscape cannot be attributed to the ARPA-E award. Broader technology trends, other funding sources, and market forces all drive patenting activity in the same technology areas.

2. **Year-level granularity.** Metrics are computed at the calendar-year level. In years with few patents in an award's technology area, metrics may be noisy or undefined.

3. **Single-year corpus.** Each year's metrics use only patents from that calendar year, not the cumulative corpus. This provides a cleaner temporal signal but means that metrics for individual years may be volatile, particularly in sparse technology areas.

4. **Patent sample representativeness.** The 50,000-patent benchmark sample is drawn from the same corpus used to construct the embedding space (patents sharing CPC codes with ARPA-E awards). This corpus is not representative of all patenting activity but is the appropriate comparison set for ARPA-E's technology domains.

5. **Post-award data requirements.** Computing post-slopes requires at least two years of post-award patent data, restricting the post-slope analysis to awards started early enough to have data through $\tau = +5$.

## 4 Program Concentration and Diversity

### 4.1 Definition and Motivation

Each ARPA-E program funds a portfolio of projects. Technology concentration measures how technologically similar or diverse those projects are. A highly concentrated program funds projects that all work on similar technology; a diverse program spans many different technology areas. Metrics are computed at the program level across 79 programs.

### 4.2 Pairwise Cosine Distance

Pairwise cosine distance measures how different projects are from each other within a program. Let $V \in \mathbb{R}^{n \times d}$ be the matrix of unit-normalized project vectors for a program with $n$ projects. The similarity matrix is:

$$S = VV^\top \tag{18}$$

Pairwise distances are extracted from the upper triangle (excluding the diagonal):

$$\{d_{ij}\}_{i<j} = \{1 - S_{ij}\}_{i<j} \tag{19}$$

Summary statistics (mean, median, standard deviation, minimum, maximum) are reported across all $\binom{n}{2}$ unique pairs. Low mean distance indicates concentrated programs; high mean distance indicates technological diversity.

### 4.3 Centroid-Based Dispersion

Centroid-based dispersion measures how far projects scatter from the program's technological center. The program centroid is the normalized mean of all project vectors:

$$c = \frac{\bar{v}}{\bar{v}}, \qquad \bar{v} = \frac{1}{n} \sum_{i=1}^{n} v_i \tag{20}$$

Each project's distance to the centroid is $d_i = 1 - v_i \cdot c$, and summary statistics are reported.

**Difference from pairwise.** Centroid distance is sensitive to cluster shape. Two programs with the same average pairwise distance can differ in centroid dispersion: a single diffuse cluster yields low centroid dispersion (projects near center), while two tight but distant sub-clusters yield high centroid dispersion (centroid sits between the clusters, far from all projects). Per-project centroid distances also enable identification of which individual projects are central versus peripheral within their program.

### 4.4 Shannon Entropy

Shannon entropy measures categorical diversity across the CPC taxonomy. For each of four hierarchy levels—section, class, subclass, and group—we compute score-

weighted shares and entropy:

$$p_i = \frac{\sum_{c \in \text{category}_i} w_c}{\sum_c w_c}, \qquad H = -\sum_i p_i \ln p_i \tag{21}$$

where $w_c$ is the EPO importance score. Entropy equals zero when all projects fall under a single category (perfectly concentrated) and reaches its maximum of $\ln(k)$ when scores are uniformly distributed across $k$ categories.

**Difference from embedding metrics.** Entropy operates on discrete technology categories, not continuous vector space. It captures categorical diversity—whether a program spans many CPC classes—but ignores how semantically similar those classes are. A program with projects in two very similar CPC classes receives higher entropy than a program entirely in one class, even if the actual technologies are nearly identical.