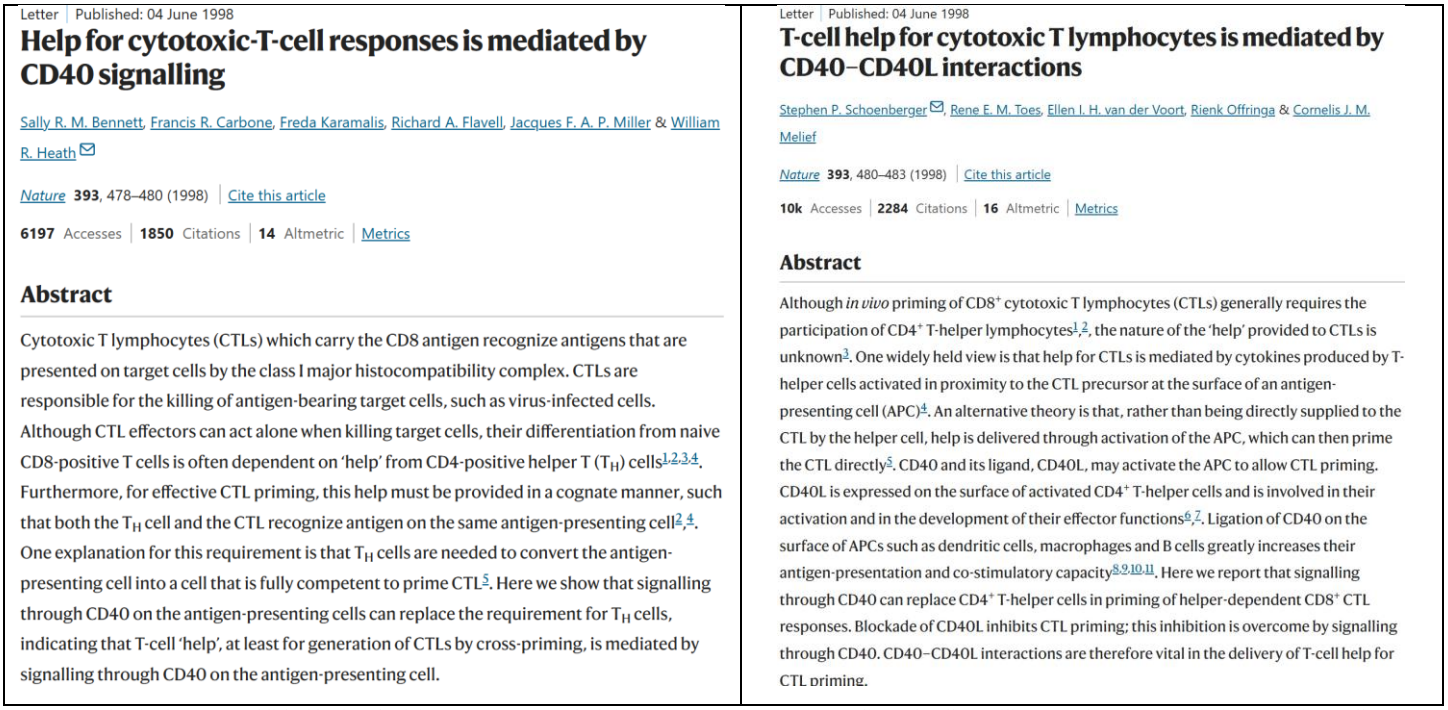
Online Appendix

Figure A.1 Example of twin papers



Letter | Published: 04 June 1998

Help for cytotoxic-T-cell responses is mediated by CD40 signalling

Sally R. M. Bennett, Francis R. Carbone, Freda Karamalis, Richard A. Flavell, Jacques F. A. P. Miller & William R. Heath

Nature **393**, 478–480 (1998) | Cite this article

6197 Accesses | **1850** Citations | **14** Altmetric | Metrics

Abstract

Cytotoxic T lymphocytes (CTLs) which carry the CD8 antigen recognize antigens that are presented on target cells by the class I major histocompatibility complex. CTLs are responsible for the killing of antigen-bearing target cells, such as virus-infected cells. Although CTL effectors can act alone when killing target cells, their differentiation from naive CD8-positive T cells is often dependent on 'help' from CD4-positive helper T (T_H) cells[1,2,3,4]. Furthermore, for effective CTL priming, this help must be provided in a cognate manner, such that both the T_H cell and the CTL recognize antigen on the same antigen-presenting cell[2,4]. One explanation for this requirement is that T_H cells are needed to convert the antigen-presenting cell into a cell that is fully competent to prime CTL[5]. Here we show that signalling through CD40 on the antigen-presenting cells can replace the requirement for T_H cells, indicating that T-cell 'help', at least for generation of CTLs by cross-priming, is mediated by signalling through CD40 on the antigen-presenting cell.

Letter | Published: 04 June 1998

T-cell help for cytotoxic T lymphocytes is mediated by CD40–CD40L interactions

Stephen P. Schoenberger, Rene E. M. Toes, Ellen I. H. van der Voort, Rienk Offringa & Cornelis J. M. Melief

Nature **393**, 480–483 (1998) | Cite this article

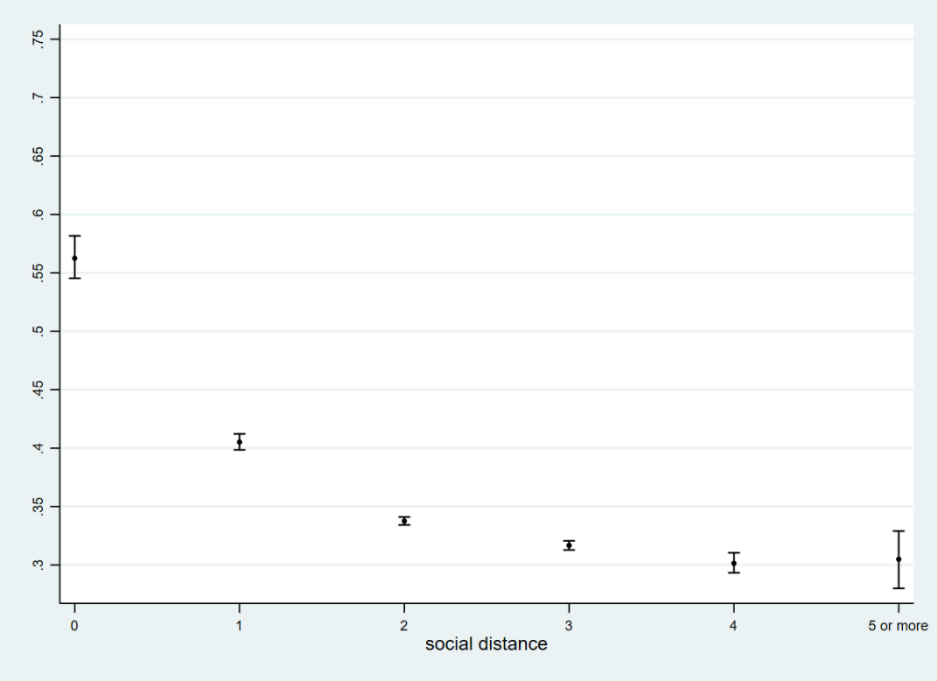
10k Accesses | **2284** Citations | **16** Altmetric | Metrics

Abstract

Although *in vivo* priming of $CD8^+$ cytotoxic T lymphocytes (CTLs) generally requires the participation of $CD4^+$ T-helper lymphocytes[1,2], the nature of the 'help' provided to CTLs is unknown[3]. One widely held view is that help for CTLs is mediated by cytokines produced by T-helper cells activated in proximity to the CTL precursor at the surface of an antigen-presenting cell (APC)[4]. An alternative theory is that, rather than being directly supplied to the CTL by the helper cell, help is delivered through activation of the APC, which can then prime the CTL directly[5]. CD40 and its ligand, CD40L, may activate the APC to allow CTL priming. CD40L is expressed on the surface of activated $CD4^+$ T-helper cells and is involved in their activation and in the development of their effector functions[6,7]. Ligation of CD40 on the surface of APCs such as dendritic cells, macrophages and B cells greatly increases their antigen-presentation and co-stimulatory capacity[8,9,10,11]. Here we report that signalling through CD40 can replace $CD4^+$ T-helper cells in priming of helper-dependent $CD8^+$ CTL responses. Blockade of CD40L inhibits CTL priming; this inhibition is overcome by signalling through CD40. CD40–CD40L interactions are therefore vital in the delivery of T-cell help for CTL priming.

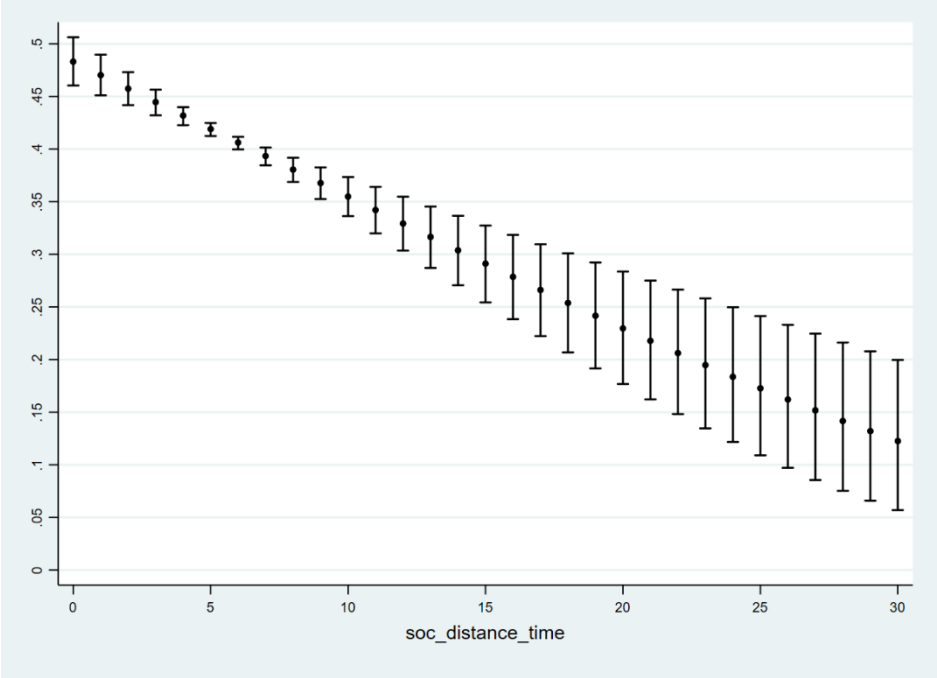
Notes: Example of twin papers published back-to-back in *Nature*. Paper on the left: Bennett S, Carbone F, Karamalis F, Flavell R, Miller J, Heath W (1998) Help for cytotoxic-T-cell responses is mediated by CD40 signalling. Nature 393(6684):478–480. Paper on the right: Schoenberger S, Toes R, van der Voort E, Offringa R, Melief C (1998) T-cell help for cytotoxic T lymphocytes is mediated by CD40–CD40L interactions. Nature 393(6684):480–483.

Figure A.2 Predicted probability of paper-to-paper citations across different levels of social distance between citing and cited academic authors



Notes: Figure A.2 depicts the predicted probability of citation, along with 95% confidence intervals, across different levels of social distance between authors in paper-to-paper dyads. The predicted probabilities and corresponding confidence intervals are obtained using the methodology proposed by King et al. (2000). Detailed regression results are available from the authors.

Figure A.3 Predicted probability of corporate patent citations to academic papers across different levels of *soc_distance_time* between academic authors and corporate inventors



Notes: Figure A.3 depicts the predicted probability of citation along with the 95% confidence interval across different levels of soc_distance_time (shortest path weighted by recency of ties) between academic authors and corporate inventors in paper-patent dyads. The predicted probabilities and corresponding confidence intervals are obtained using the methodology proposed by King et al. (2000). Detailed regression results are presented in Model 5-10 of Table 5.

Table A.1 Universities and academic institutions most commonly listed as affiliations on twin papers

	N. (1)	Share (2)
Harvard University	35	0.070
Massachusetts Institute of Technology	15	0.030
University of California, San Diego	15	0.030
University of California, San Francisco	13	0.026
Yale University	13	0.026
Johns Hopkins University	11	0.022
National Cancer Institute	11	0.022
University of Toronto	10	0.020
Cold Spring Harbor Laboratory	10	0.020
Massachusetts General Hospital	10	0.020
University of Washington	10	0.020
CNRS	10	0.020
University of California at Berkeley	10	0.020
Stanford University	9	0.018
European Molecular Biology Laboratory	9	0.018
Dana Farber Cancer Institute	9	0.018
University of Chicago	8	0.016
University of Michigan	8	0.016
Baylor College of Medicine	8	0.016
California Institute of Technology	7	0.014

Notes: Top 20 universities and institutions most commonly listed as affiliation on twin papers. The number in column 1 refers to the number of papers and the share in column 2 is computed based on the total number of twin papers in our sample (n=501). Papers with multiple affiliations can appear in multiple rows.

Table A.2 Companies most commonly listed as assignees on citing patents

	N. (1)	Share (2)
Genentech	219	0.112
Millennium Pharmaceuticals	209	0.107
ZymoGenetics	59	0.030
BASF Plant Science	45	0.023
BASF AG	35	0.018
CuraGen Corporation	34	0.017
Eli Lilly & Company	32	0.016
Kyowa Hakko Kogyo Co.	26	0.013
Amgen	25	0.013
Kyowa Hakko Kirin Co.	24	0.012
Schering Corporation	23	0.012
Virogenetics Corporation	23	0.012
Abbott Laboratories	22	0.011
Agrinomics	22	0.011
Biogen IDEC	20	0.010
Human Genome Sciences	19	0.010
Bio Merieux	17	0.009
Bristol-Myers Squibb Company	18	0.009
Pioneer Hi-Bred International	16	0.008
Geron Corporation	15	0.008

Notes: Top 20 companies most commonly listed as assignee on citing patents. The number in column 1 refers to the number of patents and the share in column 2 is computed based on the total number of citing patents in our sample (n=1,951). Patents with multiple assignees can appear in multiple rows.

Table A.3 Results by company type

	Main effect	U.S. public firms	Firm age <=3 years	Firm age <=5 years	Firm age <=10 years	Firm patent count<p50	Firm patent count>=p50
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
social_dist	-0.957***	-0.813***	-0.952*	-0.656*	-0.661***	-0.948***	-1.011***
	(0.205)	(0.151)	(0.526)	(0.342)	(0.221)	(0.320)	(0.225)
N of observations (paper-patent dyads)	4,169	2,124	237	405	841	765	1,288
N of twin-patent dyads	2,068	1,050	115	198	411	370	644
N of citing patents	1,951	1,011	103	181	387	353	626
N of papers	501	245	92	130	202	175	142
N of twins	247	121	44	63	99	86	71
Log-Likelihood	-873.8	-366.0	-45.5	-72.3	-162.2	-147.4	-155.1
R2	0.396	0.504	0.449	0.487	0.446	0.447	0.653
Mean(cited)	0.497	0.494	0.489	0.491	0.490	0.485	0.500

Notes: The unit of observation is a paper-patent dyad (n=4,169), including cited paper-patent dyads (n=2,070) and uncited paper-patent dyads (n=2,099). The sample includes 247 paper twins (7 are triplets), 501 papers, and 1,951 citing patents. We identify all cases where a patent cites only one (or a subset) of the twin papers. The binary outcome variable *cited* is equal to 1 in case the paper is cited by the patent, and 0 in case of an uncited twin paper patent dyad, i.e., the uncited twin paper matched to the same patent. The table reports coefficient estimates from conditional logit models with fixed effects at the twin patent-level. Coefficients are in log-odds units. All models include control variables: *degree3_size, no. authors, author-inventor team, publication stock, patent stock, institutional prestige, U.S. based, min geographic distance, journal impact factor, journal commercial impact factor, paper citations, patent citations, time lag, published first, text length and readability*. All control variables measured in counts are logarithmically transformed after adding one to variables with zero values. Robust standard errors clustered at the twin level and reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level. Mean(cited) is the observed citation probability in each column's estimation sample.

Table A.4 Robustness checks: Company inventors active in scientific research as cross-institutional boundary-spanners

	Conditional logit	Conditional logit	Conditional logit	Conditional logit	Conditional logit	Conditional logit	Conditional logit	Conditional logit	Linear probability	Conditional logit	Conditional logit
			Exclude social distance 0	Exclude social distance 0 or 1	Exclude social unreachable	Social distance measured before academic twin	Exclude corporate funding	Exclude front page citations		Social distance time (shortest path weighted by recency of ties)	Count measure for *active_inventor_author*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
social_dist		-0.068	-0.061	-0.001	-0.067	-0.189	-0.162	0.003	-0.009	-0.027	-0.764***
		(0.187)	(0.195)	(0.229)	(0.185)	(0.253)	(0.165)	(0.233)	(0.040)	(0.098)	(0.173)
social_dist*active_inventor_author		-1.081***	-1.032***	-0.849**	-1.072***	-0.718**	-1.018***	-1.185***	-0.223***	-0.218	-0.034**
		(0.289)	(0.299)	(0.356)	(0.287)	(0.295)	(0.318)	(0.326)	(0.056)	(0.134)	(0.014)
N of observations (paper-patent dyads)	4,169	4,169	4,039	3,706	4,099	4,169	3,591	3,422	5,745	4,169	4,169
N of twin-patent dyads	2,068	2,068	2,003	1,837	2,033	2,068	1,782	1,698	2,850	2,068	2,068
N of citing patents	1,951	1,951	1,886	1,732	1,920	1,951	1,704	1,603	2,641	1,951	1,951
N of papers	501	501	493	462	491	501	405	362	634	501	501
N of twins	247	247	243	228	242	247	200	178	312	247	247
Log-Likelihood	-945.8	-859.0	-826.9	-722.9	-846.6	-882.4	-721.1	-679.6		-924.6	-868.0
R2	0.346	0.406	0.410	0.438	0.405	0.390	0.421	0.428	0.390	0.361	0.400
Mean(cited)	0.497	0.497	0.496	0.496	0.496	0.497	0.497	0.497	0.635	0.497	0.497

Notes: The unit of observation is a paper-patent dyad (n=4,169), including cited paper-patent dyads (n=2,070) and uncited paper-patent dyads (n=2,099). The sample includes 247 paper twins (7 are triplets), 501 papers, and 1,951 citing patents. We identify all cases where a patent cites only one (or a subset) of the twin papers. The binary outcome variable *cited* is equal to 1 in case the paper is cited by the patent, and 0 in case of an uncited twin paper patent dyad, i.e., the uncited twin paper matched to the same patent. The table reports coefficient estimates from conditional logit models with fixed effects at the twin patent-level. Coefficients are in log-odds units (except in column 9). Column 9 also includes cases where a patent cites all twin papers, estimated using a linear probability model with twin patent-level fixed effects. *Active_inventor_author* is a dummy equal to 1 if any of the patent's inventors have at least 1 PubMed paper published in the year before patent filing. In column 11, *Active_inventor_author* is a total count of PubMed indexed papers authored by patent's inventors in the year before patent filing. All models include control variables: *degree3_size, no. authors, author-inventor team, publication stock, patent stock, institutional prestige, U.S. based, min geographic distance, journal impact factor, journal commercial impact factor, paper citations, patent citations, time lag, published first, text length and readability*. All control variables measured in counts are logarithmically transformed after adding one to variables with zero values. Robust standard errors clustered at the twin level and reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level. Mean(cited) is the observed citation probability in each column's estimation sample.

Table A.5 Robustness checks: Company inventors with closely related scientific expertise as cross-institutional boundary-spanners

	Conditional logit	Conditional logit	Conditional logit	Conditional logit	Conditional logit	Conditional logit	Conditional logit	Conditional logit	Linear probability	Conditional logit	Conditional logit
			Exclude social distance 0	Exclude social distance 0 or 1	Exclude social unreachable	Social distance measured before academic twin	Exclude corporate funding	Exclude front page citations		Social distance time (shortest path weighted by recency of ties)	Continuous measure for *close_paper*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
social_dist		-0.376***	-0.382***	-0.250	-0.367***	-0.206	-0.398***	-0.311*	-0.084**	0.093	0.573
		(0.140)	(0.141)	(0.187)	(0.140)	(0.207)	(0.153)	(0.164)	(0.033)	(0.069)	(0.556)
social_dist*close_paper		-0.920***	-0.915***	-0.880*	-0.922***	-0.894*	-0.935**	-1.101***	-0.165**	-0.476***	-1.948**
		(0.325)	(0.328)	(0.455)	(0.322)	(0.490)	(0.368)	(0.322)	(0.075)	(0.116)	(0.805)
N of observations (paper-patent dyads)	4,169	4,115	3,985	3,652	4,053	4,115	3,551	3,380	5,675	4,115	4,115
N of twin-patent dyads	2,068	2,041	1,976	1,810	2,010	2,041	1,762	1,677	2,815	2,041	2,041
N of citing patents	1,951	1,924	1,859	1,705	1,897	1,924	1,684	1,582	2,606	1,924	1,924
N of papers	501	493	485	454	485	493	401	356	628	493	493
N of twins	247	243	239	224	239	243	198	175	309	243	243
Log-Likelihood	-945.8	-848.4	-815.4	-712.0	-836.4	-864.5	-711.9	-666.3		-895.3	-856.7
R2	0.346	0.406	0.410	0.438	0.405	0.395	0.422	0.432	0.387	0.373	0.400
Mean(cited)	0.497	0.496	0.496	0.496	0.496	0.496	0.497	0.497	0.635	0.496	0.496

Notes: The unit of observation is a paper-patent dyad (n=4,169), including cited paper-patent dyads (n=2,070) and uncited paper-patent dyads (n=2,099). The sample includes 247 paper twins (7 are triplets), 501 papers, and 1,951 citing patents. We identify all cases where a patent cites only one (or a subset) of the twin papers. The binary outcome variable *cited* is equal to 1 in case the paper is cited by the patent, and 0 in case of an uncited twin paper patent dyad, i.e., the uncited twin paper matched to the same patent. The table reports coefficient estimates from conditional logit models with fixed effects at the twin patent-level. Coefficients are in log-odds units (except in column 9). Column 9 also includes cases where a patent cites all twin papers, estimated using a linear probability model with twin patent-level fixed effects. *Close_paper* is a dummy equal to 1, if the max similarity between inventors' prior papers (published before the patent filing year) and the twin papers (cited or not cited) is higher than sample median (0.772). In column 11, *Close_paper* is a continuous measure, i.e., max cosine similarity between inventors' prior papers (published before the patent filing year) and the twin papers (cited or not cited). Paper-patent dyads where inventors have 0 prior papers have a value of 0 for cosine similarity. Paper-patent dyads for which cosine similarity could not be calculated due to missing text data are dropped in all models. We tested alternative models where patent-twin dyads with 0 prior papers were also put to missing observations, the effect size and significance hold. All models include control variables: *degree3_size, no. authors, author-inventor team, publication stock, patent stock, institutional prestige, U.S. based, min geographic distance, journal impact factor, journal commercial impact factor, paper citations, patent citations, time lag, published first, text length and readability*. All control variables measured in counts are logarithmically transformed after adding one to variables with zero values. Robust standard errors clustered at the twin level and reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level. Mean(cited) is the observed citation probability in each column's estimation sample.

Table A.6 Summary statistics at the paper-paper dyad level

	All		Cited		Not Cited		T test
	Mean	SD	Mean	SD	Mean	SD	p value
cited	0.499	0.500	1.000	0.000	0.000	0.000	
social_dist	2.325	1.051	2.212	1.120	2.438	0.965	0.000***
social_dist_0	0.047	0.212	0.079	0.270	0.015	0.123	0.000***
social_dist_1	0.124	0.330	0.138	0.345	0.110	0.313	0.000***
social_dist_2	0.392	0.488	0.375	0.484	0.410	0.492	0.000***
social_dist_3	0.358	0.479	0.335	0.472	0.380	0.485	0.000***
social_dist_4	0.067	0.250	0.062	0.242	0.072	0.258	0.000***
social_dist_5_or_more	0.012	0.107	0.011	0.103	0.013	0.111	0.069*
min geographic distance	6.641	2.284	6.348	2.509	6.933	1.994	0.000***
time lag	7.417	1.110	7.433	1.053	7.401	1.164	0.002**

Notes: The unit of observation is a paper–paper dyad (n = 44,220), including cited dyads (n = 22,069) and uncited dyads (n = 22,151). The sample comprises 221 paper twins, 448 twin papers, and 21,571 citing papers. We identify all PubMed papers published between 1975 and 2009 that cite at least one of the twin papers in our sample. Following the patent-to-paper identification strategy, we retain only cases where a citing paper references one (or a subset, in the case of triplets) of the twin papers, using the uncited twin paper matched to the same citing paper as a counterfactual or "pseudo-citation." Minimum geographic distance and time lag are logarithmically transformed after adding one. T-tests assume unequal variances. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.7 Influence of social distance on paper-to-paper citation

	Conditional logit	Conditional logit	Conditional logit Exclude social distance 0	Conditional logit Exclude social distance 0 or 1	Conditional logit Exclude social unreachable	Linear Probability
	(1)	(2)	(3)	(4)	(5)	(6)
social_dist		-0.406***	-0.298***	-0.153***	-0.407***	-0.169***
		(0.026)	(0.027)	(0.032)	(0.026)	(0.011)
N of observations (paper-paper dyads)	44,220	44,220	40,058	32,030	44,032	44,220
N of twin-citing paper dyads	22,005	22,005	19,939	15,945	21,911	22,005
N of citing papers	21,571	21,571	19,562	15,703	21,479	21,571
N of twin papers	448	448	448	448	448	448
N of twins	221	221	221	221	221	221
Log-Likelihood	-13890.6	-13620.1	-12699.6	-10241.8	-13560.6	
R2	0.094	0.112	0.086	0.078	0.112	0.125
Mean(cited)	0.499	0.499	0.499	0.499	0.499	0.499

Notes: The unit of observation is a paper–paper dyad (n = 44,220), including cited (n = 22,069) and uncited (n = 22,151) dyads. The sample includes 221 paper twins, 448 twin papers, and 21,571 citing papers. We identify all cases where a citing paper references only one (or a subset, in the case of triplets) of the twin papers. The binary outcome variable *cited* equals 1 if the focal twin paper is cited by the citing paper and 0 if it is the uncited twin paper matched to the same citing paper. The table reports log-odds coefficient estimates from conditional logit models with fixed effects at the twin–citing-paper level (except in column 6). Column 6 is estimated using a linear probability model with twin–citing-paper fixed effects. All models include the same control variables as in the patent-to-paper analysis: *degree3_size, no. authors, author-inventor team, publication stock, patent stock, institutional prestige, U.S. based, min geographic distance, journal impact factor, journal commercial impact factor, paper citations, patent citations, time lag, published first, text length and readability*. All control variables measured in counts are logarithmically transformed after adding one to variables with zero values. Robust standard errors clustered at the twin level and reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level. Mean(cited) is the observed citation probability in each column's estimation sample.

Table A.8 Robustness checks: Binary indicators for social distance

	Exclude higher order twins	Exclude twins published > 3 months apart	Exclude twins text similarity <0.90	Twins with share of adjacent citations =100%	Exclude twins with papers from large teams (>5)	Exclude twins with top cited papers	Exclude patents of Genentech & Millennium Pharmaceuticals
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
social_dist_1	-2.568***	-2.494**	-3.331***	-1.742**	-3.138***	-2.532***	-2.277***
	(0.665)	(1.031)	(0.717)	(0.686)	(0.604)	(0.691)	(0.626)
social_dist_2	-4.104***	-4.089***	-5.232***	-3.510***	-5.655***	-3.985***	-3.631***
	(0.992)	(1.307)	(1.053)	(0.744)	(0.866)	(1.053)	(0.853)
social_dist_3	-4.953***	-5.152***	-6.152***	-4.234***	-6.687***	-4.884***	-4.423***
	(1.075)	(1.433)	(1.188)	(0.668)	(0.981)	(1.130)	(0.895)
social_dist_4	-4.918***	-5.309***	-6.109***	-3.701***	-6.861***	-4.944***	-4.395***
	(1.182)	(1.515)	(1.327)	(0.839)	(1.095)	(1.204)	(0.963)
social_dist_5_or_more	-5.741***	-6.160***	-6.836***	-9.905***	-7.183***	-5.676***	-5.042***
	(1.134)	(1.293)	(1.153)	(1.823)	(1.102)	(1.145)	(1.058)
N of observations (paper-patent dyads)	3,996	3,368	3,310	935	2,953	4,007	3,417
N of twin-patent dyads	1,998	1,668	1,642	467	1,464	1,987	1,692
N of citing patents	1,888	1,583	1,587	450	1,410	1,870	1,584
N of papers	480	354	308	103	246	493	493
N of twins	240	174	152	51	121	243	243
Log-Likelihood	-810.2	-663.5	-632.8	-187.5	-512.4	-799.5	-783.7
R2	0.415	0.433	0.449	0.422	0.500	0.425	0.339
Mean(cited)	0.500	0.496	0.496	0.501	0.496	0.496	0.496

Notes: The unit of observation is a paper-patent dyad (n=4,169), including cited paper-patent dyads (n=2,070) and uncited paper-patent dyads (n=2,099). The sample includes 247 paper twins (7 are triplets), 501 papers, and 1,951 citing patents. We identify all cases where a patent cites only one (or a subset) of the twin papers. The binary outcome variable *cited* is equal to 1 in case the paper is cited by the patent, and 0 in case of an uncited twin paper patent dyad, i.e., the uncited twin paper matched to the same patent. The table reports coefficient estimates from conditional logit models with fixed effects at the twin patent-level. Coefficients are in log-odds units. Distance 0 is excluded category in all models. All models include control variables: *degree3_size, no. authors, author-inventor team, publication stock, patent stock, institutional prestige, U.S. based, min geographic distance, journal impact factor, journal commercial impact factor, paper citations, patent citations, time lag, published first, text length and readability*. Please notice that estimations for the subsample of twins with PubMed rank order=1 or published back-to-back did not converge because of insufficient observations for some of the different binary categories of social distance. All control variables measured in counts are logarithmically transformed after adding one to variables with zero values. Robust standard errors clustered at the twin level and reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level. Mean(cited) is the observed citation probability in each column's estimation sample.

Table A.9 Binary social-distance estimates with and without the geographic proximity control

	Conditional logit Include min geographic distance (1)	Conditional logit Exclude min geographic distance (2)
social_dist_1	-2.571***	-1.357**
	(0.653)	(0.631)
social_dist_2	-4.101***	-2.865***
	(0.974)	(0.803)
social_dist_3	-4.910***	-3.488***
	(1.068)	(0.801)
social_dist_4	-4.906***	-3.468***
	(1.149)	(0.885)
social_dist_5_or_more	-5.585***	-4.492***
	(1.111)	(0.997)
N of observations (paper-patent dyads)	4,169	4,169
N of twin-patent dyads	2,068	2,068
N of citing patents	1,951	1,951
N of papers	501	501
N of twins	247	247
Log-Likelihood	-847.4	-895.4
R2	0.414	0.381
Mean(cited)	0.497	0.497

Notes: The unit of observation is a paper-patent dyad (n=4,169), including cited paper-patent dyads (n=2,070) and uncited paper-patent dyads (n=2,099). The sample includes 247 paper twins (7 are triplets), 501 papers, and 1,951 citing patents. We identify all cases where a patent cites only one (or a subset) of the twin papers. The binary outcome variable *cited* is equal to 1 in case the paper is cited by the patent, and 0 in case of an uncited twin paper patent dyad, i.e., the uncited twin paper matched to the same patent. The table reports coefficient estimates from conditional logit models with fixed effects at the twin patent-level, social_dist_0 is the excluded category in columns 1 and 2. Both models include control variables: *degree3_size, no. authors, author-inventor team, publication stock, patent stock, institutional prestige, U.S. based, journal impact factor, journal commercial impact factor, paper citations, patent citations, time lag, published first, text length and readability*. First model additionally includes *min geographic distance* as a control variable. All control variables measured in counts are logarithmically transformed after adding one to variables with zero values. Robust standard errors clustered at the twin level and reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level.

Table A.10 Count social-distance estimates with and without the geographic proximity control

	Conditional logit	Conditional logit	Conditional logit	Conditional logit	Conditional logit	Conditional logit	Conditional logit	Conditional logit
	Include min geographic distance				Exclude min geographic distance			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
social_dist	-0.957***	-0.068	-0.376***	0.105	-0.775***	0.010	-0.293*	0.173
	(0.205)	(0.187)	(0.140)	(0.213)	(0.153)	(0.214)	(0.166)	(0.263)
social_dist*active_inventor_author		-1.081***		-0.744***		-0.963***		-0.723***
		(0.289)		(0.216)		(0.276)		(0.219)
social_dist*close_paper			-0.920***	-0.728**			-0.759**	-0.572*
			(0.325)	(0.321)			(0.340)	(0.333)
N of observations (paper-patent dyads)	4,169	4,169	4,115	4,115	4,169	4,169	4,115	4,115
N of twin-patent dyads	2,068	2,068	2,041	2,041	2,068	2,068	2,041	2,041
N of citing patents	1,951	1,951	1,924	1,924	1,951	1,951	1,924	1,924
N of papers	501	501	493	493	501	501	493	493
N of twins	247	247	243	243	247	247	243	243
Log-Likelihood	-873.8	-859.0	-848.4	-842.1	-913.2	-900.1	-890.9	-884.5
R2	0.396	0.406	0.406	0.410	0.369	0.378	0.376	0.381
Mean(cited)	0.497	0.497	0.496	0.496	0.497	0.497	0.496	0.496

Notes: The unit of observation is a paper-patent dyad (n=4,169), including cited paper-patent dyads (n=2,070) and uncited paper-patent dyads (n=2,099). The sample includes 247 paper twins (7 are triplets), 501 papers, and 1,951 citing patents. We identify all cases where a patent cites only one (or a subset) of the twin papers. The binary outcome variable *cited* is equal to 1 in case the paper is cited by the patent, and 0 in case of an uncited twin paper patent dyad, i.e., the uncited twin paper matched to the same patent. The table reports coefficient estimates from conditional logit models with fixed effects at the twin patent-level. Models 1-4 include control variables: *degree3_size, no. authors, author-inventor team, publication stock, patent stock, institutional prestige, U.S. based, min geographic distance, journal impact factor, journal commercial impact factor, paper citations, patent citations, time lag, published first, text length and readability*. Models 5-8 exclude *min geographic distance* from control variables list. *Active_inventor_author* is a dummy equal to 1 if any of the patent's inventors have at least 1 PubMed paper published in the year before patent filing. *Close_paper* is a dummy equal to 1, if the max similarity between inventors' prior papers (published before the patent filing year) and the twin papers (cited or not cited) is higher than sample median. Paper-patent dyads where inventors have 0 prior papers have a value of 0 for cosine similarity. Paper-patent dyads for which cosine similarity could not be calculated due to missing text data are dropped in models 3, 4, 7, 8. All control variables measured in counts are logarithmically transformed after adding one to variables with zero values. Robust standard errors clustered at the twin level and reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level.

Table A.11 Descriptive Statistics Comparing Corporate Inventors With and Without a Shortest-Path Connection to Cited Academic Scientists

	Corporate inventor with shortest path		Corporate inventor without shortest path		T test
	Mean	SD	Mean	SD	p value
Years since first patent/paper with same firm	4.302	4.647	3.260	4.296	0.000***
Recent public research affiliation	0.728	0.445	0.539	0.499	0.000***
Past public research affiliation	0.874	0.332	0.725	0.446	0.000***
Shared recent public research affiliation	0.251	0.434	0.150	0.357	0.000***
Shared past public research affiliation	0.373	0.484	0.200	0.400	0.000***
Number of prior patents	16.922	31.539	16.236	39.499	0.396
Number of prior patents in t-1	1.911	3.864	1.563	3.997	0.000***
Patent experience (years)	6.731	5.979	4.665	5.488	0.000***
Number of prior publications	36.733	60.572	11.189	24.820	0.000***
Number of prior publications in t-1	2.599	4.232	0.919	2.046	0.000***
Publication experience (years)	14.275	9.669	9.641	8.893	0.000***
Total network size (up to 3 degrees, log)	11.626	1.767	10.506	1.658	0.000***
Inventor network size (up to 3 degrees, log)	7.141	1.597	6.112	1.619	0.000***
Author network size (up to 3 degrees, log)	11.450	1.813	10.291	1.710	0.000***
Inventor-author network size (up to 3 degrees, log)	9.657	1.549	8.702	1.513	0.000***

Notes: The unit of observation is a corporate inventor at the time of patent filing linked to a cited twin paper. *Corporate inventor with shortest path* indicates the inventor has the shortest social distance to the authors of the cited twin paper, while *without shortest path* refers to other inventors on the same patent. *Recent public research affiliation* equals 1 if the inventor has a non-corporate (university, hospital, government, or other public research) affiliation on a paper or patent in the previous five years (t–5 to t–1); *past public research affiliation* refers to such affiliations held more than five years before filing (t–6 or earlier). *Shared recent (past) public research affiliation* equals 1 if the inventor and any author of the cited paper were affiliated with the same non-corporate organization within the recent (past) period. *Network size* variables measure the logarithm of the number of unique individuals within three degrees of separation at the time of patent filing: *total* includes all individuals, while *inventor*, *author*, and *inventor–author* network sizes restrict the network to inventors only, authors only, or those active in both domains. Reported p-values are based on unequal-variance t-tests; *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.12 Descriptive Statistics Comparing Corporate Inventors With and Without Recent Scientific Publications

	Inventors with recent publications		Inventors without recent publications		T test
	Mean	SD	Mean	SD	p value
Years since first patent/paper with same firm	4.371	4.560	3.239	4.387	0.000***
Recent public research affiliation	0.810	0.393	0.446	0.497	0.000***
Past public research affiliation	0.885	0.320	0.719	0.450	0.000***
Shared recent public research affiliation	0.287	0.453	0.109	0.312	0.000***
Shared past public research affiliation	0.361	0.480	0.227	0.419	0.000***
Number of prior patents	20.800	43.528	12.657	25.831	0.000***
Number of prior patents in t-1	1.864	3.148	1.621	4.542	0.005***
Patent experience (years)	6.484	5.968	4.983	5.609	0.000***
Number of prior publications	42.360	62.579	7.042	15.154	0.000***
Number of prior publications in t-1	3.664	4.189	0.000	0.000	0.000***
Publication experience (years)	14.122	9.687	9.765	8.936	0.000***
Total network size (up to 3 degrees, log)	11.932	1.144	10.270	1.930	0.000***
Inventor network size (up to 3 degrees, log)	7.352	1.351	5.959	1.694	0.000***
Author network size (up to 3 degrees, log)	11.768	1.171	10.044	1.985	0.000***
Inventor-author network size (up to 3 degrees, log)	9.903	1.028	8.515	1.748	0.000***

Notes: The unit of observation is a corporate inventor at the time of patent filing linked to a cited twin paper. *Corporate inventor with recent publications* indicates the inventor has at least one scientific publication in the year before patent filing (t-1), while *without recent publications* refers to inventors without recently published papers. *Recent public research affiliation* equals 1 if the inventor has a non-corporate (university, hospital, government, or other public research) affiliation on a paper or patent in the previous five years (t–5 to t–1); *past public research affiliation* refers to such affiliations held more than five years before filing (t–6 or earlier). *Shared recent (past) public research affiliation* equals 1 if the inventor and any author of the cited paper were affiliated with the same non-corporate organization within the recent (past) period. *Network size* variables measure the logarithm of the number of unique individuals within three degrees of separation at the time of patent filing: *total* includes all individuals, while *inventor*, *author*, and *inventor–author* network sizes restrict the network to inventors only, authors only, or those active in both domains. Reported p-values are based on unequal-variance t-tests; *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.13 Placebo test for differential measurement precision: Moderation of the social-distance effect by inventor network size, by active inventor-author status

	Conditional logit *Active_inventor_author* = 0 (1)	Conditional logit *Active_inventor_author* = 1 (2)
social_dist	-0.279	-0.921***
	(0.310)	(0.157)
social_dist*above_median_patent_network	0.117	-0.413
	(0.464)	(0.355)
N of observations (paper-patent dyads)	891	3,278
N of twin-patent dyads	443	1,625
N of citing patents	410	1,541
N of papers	214	425
N of twins	105	210
Log-Likelihood	-132.1	-705.6
R2	0.573	0.380
Mean(cited)	0.498	0.496

Notes: The unit of observation is a paper-patent dyad. Column 1 restricts the sample to patents where no inventor published a PubMed paper in the year before filing (*active_inventor_author* = 0); Column 2 restricts to patents with at least one such inventor (*active_inventor_author* = 1). *Above_median_patent_network* is a dummy equal to 1 if the patent's inventor network size exceeds the median. *Patent inventor network size* is the number of unique individuals within three degrees of separation of any of the patent's inventors in the collaboration network, measured in the year before patent filing. The table reports coefficient estimates from conditional logit models with fixed effects at the twin patent-level. Coefficients are in log-odds units. All models include the same control variables: *degree3_size, no. authors, author-inventor team, publication stock, patent stock, institutional prestige, U.S. based, min geographic distance, journal impact factor, journal commercial impact factor, paper citations, patent citations, time lag, published first, text length and readability.* All control variables measured in counts are logarithmically transformed after adding one to variables with zero values. Robust standard errors clustered at the twin level and reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level. Mean(cited) is the observed citation probability in each column's estimation sample.

Table A.14 Influence of social distance between academic scientists and corporate inventors on company patent citation to academic paper: Binary indicators for social distance (controlling for shared PI)

	Conditional logit Only controls (1)	Conditional logit (2)	Conditional logit Exclude social distance 0 (3)	Conditional logit Exclude social distance 0 or 1 (4)	Conditional logit Exclude social unreachable (5)	Conditional logit Social distance measured before academic paper twin (6)	Conditional logit Exclude corporate funding (7)	Conditional logit Exclude front page citations (8)	Linear probability (9)
social_dist_1		-2.361***			-2.379***	-2.327***	-2.483***	-2.580***	-0.696***
		(0.685)			(0.683)	(0.722)	(0.751)	(0.724)	(0.145)
social_dist_2		-3.667***	-1.440***		-3.691***	-1.842**	-3.837***	-4.041***	-1.139***
		(0.931)	(0.407)		(0.929)	(0.902)	(1.045)	(0.967)	(0.203)
social_dist_3		-4.445***	-2.243***	-0.845***	-4.452***	-2.681***	-4.636***	-4.920***	-1.344***
		(0.924)	(0.456)	(0.248)	(0.920)	(0.897)	(1.074)	(1.015)	(0.189)
social_dist_4		-4.445***	-2.337***	-0.959**	-4.436***	-3.275***	-4.823***	-4.886***	-1.297***
		(0.994)	(0.550)	(0.382)	(0.987)	(0.991)	(1.136)	(1.174)	(0.203)
social_dist_5_or_more		-5.129***	-3.085***	-1.754***	-5.070***	-3.972***	-5.476***	-4.753***	-1.676***
		(1.055)	(0.689)	(0.608)	(1.059)	(0.984)	(1.210)	(1.316)	(0.301)
shared PI	2.352**	0.733	1.257	0.346	0.728	1.602*	0.664	0.623	0.188
	(0.933)	(0.754)	(0.879)	(1.000)	(0.756)	(0.823)	(0.893)	(0.822)	(0.220)
N of observations (paper-patent dyads)	4,169	4,169	4,039	3,706	4,099	4,169	3,591	3,422	4,169
N of twin-patent dyads	2,068	2,068	2,003	1,837	2,033	2,068	1,782	1,698	2,068
N of citing patents	1,951	1,951	1,886	1,732	1,920	1,951	1,704	1,603	1,951
N of papers	501	501	493	462	491	501	405	362	501
N of twins	247	247	243	228	242	247	200	178	247
Log-Likelihood	-904.3	-844.9	-815.6	-724.4	-831.6	-859.5	-711.8	-665.0	
R2	0.375	0.416	0.418	0.437	0.415	0.406	0.429	0.440	0.427
Mean(cited)	0.497	0.497	0.496	0.496	0.496	0.497	0.497	0.497	0.497

Notes: The unit of observation is a paper-patent dyad (n=4,169), including cited paper-patent dyads (n=2,070) and uncited paper-patent dyads (n=2,099). The sample includes 247 paper twins (7 are triplets), 501 papers, and 1,951 citing patents. We identify all cases where a patent cites only one (or a subset) of the twin papers. The binary outcome variable *cited* is equal to 1 in case the paper is cited by the patent, and 0 in case of an uncited twin paper patent dyad, i.e. the uncited twin paper matched to the same patent. The table reports coefficient estimates from conditional logit models with fixed effects at the twin patent-level. Coefficients are in log-odds units. social_dist_0 is the excluded category in columns 2, 5, 6, 7, 8, 9 and social_dist_1 is the excluded category in column 3, social_dist_2 is the excluded category in column 4. Column 9 presents estimates from a linear probability model with twin patent-level fixed effects; coefficients in this column are directly interpretable as changes in probability. All models include control variables: *degree3_size, no. authors, author-inventor team, publication stock, patent stock, institutional prestige, U.S. based, min geographic distance, journal impact factor, journal commercial impact factor, paper citations, patent citations, time lag, published first, text length and readability, shared PI*. *Shared PI* is an indicator equal to one if any patent inventor and any author of the focal twin paper share an early-career mentor, identified as a last author on their publications during the first five years of their respective publishing careers. All control variables measured in counts are logarithmically transformed after adding one to variables with zero values. Robust standard errors clustered at the twin level and reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level. Mean(cited) is the observed citation probability in each column's estimation sample.

Table A.15 Influence of social distance between academic scientists and corporate inventors on company patent citation to academic paper: Count measure for social distance (Controlling for shared PI)

	Conditional logit	Conditional logit	Conditional logit Exclude social distance 0	Conditional logit Exclude social distance 0 or 1	Conditional logit Exclude social unreachable	Conditional logit Social distance measured before academic paper twin	Conditional logit Exclude corporate funding	Conditional logit Exclude front page citations	Linear probability	Conditional logit Social distance time (shortest path weighted by recency of ties)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
social_dist		-0.799***	-0.755***	-0.626***	-0.789***	-0.679***	-0.877***	-0.881***	-0.221***	-0.165**
		(0.153)	(0.149)	(0.186)	(0.152)	(0.183)	(0.155)	(0.180)	(0.037)	(0.066)
shared PI	2.352**	1.330*	1.727*	0.426	1.336*	1.638**	1.188	1.199	0.447*	2.159**
	(0.933)	(0.770)	(0.944)	(0.971)	(0.773)	(0.792)	(0.920)	(0.823)	(0.227)	(0.983)
N of observations (paper-patent dyads)	4,169	4,169	4,039	3,706	4,099	4,169	3,591	3,422	4,169	4,169
N of twin-patent dyads	2,068	2,068	2,003	1,837	2,033	2,068	1,782	1,698	2,068	2,068
N of citing patents	1,951	1,951	1,886	1,732	1,920	1,951	1,704	1,603	1,951	1,951
N of papers	501	501	493	462	491	501	405	362	501	501
N of twins	247	247	243	228	242	247	200	178	247	247
Log-Likelihood	-904.3	-863.4	-825.8	-729.4	-850.6	-870.8	-724.8	-683.8		-893.9
R2	0.375	0.403	0.411	0.433	0.402	0.398	0.418	0.424	0.427	0.382
Mean(cited)	0.497	0.497	0.496	0.496	0.496	0.497	0.497	0.497	0.497	0.497

Notes: The unit of observation is a paper-patent dyad (n=4,169), including cited paper-patent dyads (n=2,070) and uncited paper-patent dyads (n=2,099). The sample includes 247 paper twins (7 are triplets), 501 papers, and 1,951 citing patents. We identify all cases where a patent cites only one (or a subset) of the twin papers. The binary outcome variable *cited* is equal to 1 in case the paper is cited by the patent, and 0 in case of an uncited twin paper patent dyad, i.e. the uncited twin paper matched to the same patent. The table reports coefficient estimates from conditional logit models with fixed effects at the twin patent-level. Coefficients are in log-odds units. Column 9 presents estimates from a linear probability model with twin patent-level fixed effects; coefficients in this column are directly interpretable as changes in probability. All models include control variables: *degree3_size, no. authors, author-inventor team, publication stock, patent stock, institutional prestige, U.S. based, min geographic distance, journal impact factor, journal commercial impact factor, paper citations, patent citations, time lag, published first, text length and readability, shared PI*. *Shared PI* is an indicator equal to one if any patent inventor and any author of the focal twin paper share an early-career mentor, identified as a last author on their publications during the first five years of their respective publishing careers. All control variables measured in counts are logarithmically transformed after adding one to variables with zero values. Robust standard errors clustered at the twin level and reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level. Mean(cited) is the observed citation probability in each column's estimation sample.

Table A.16 Mechanisms (Controlling for shared PI)

	Conditional logit (1)	Conditional logit (2)	Conditional logit (3)	Conditional logit (4)
social_dist	-0.799***	-0.001	-0.330**	0.154
	(0.153)	(0.192)	(0.152)	(0.237)
social_dist*active_inventor_author		-1.000***		-0.745***
		(0.260)		(0.213)
social_dist*close_paper			-0.813***	-0.625**
			(0.307)	(0.310)
shared PI	1.330*	1.237	1.031	1.044
	(0.770)	(0.774)	(0.759)	(0.746)
N of observations (paper-patent dyads)	4,169	4,169	4,115	4,115
N of twin-patent dyads	2,068	2,068	2,041	2,041
N of citing patents	1,951	1,951	1,924	1,924
N of papers	501	501	493	493
N of twins	247	247	243	243
Log-Likelihood	-863.4	-850.6	-842.7	-836.3
R2	0.403	0.412	0.410	0.414
Mean(cited)	0.497	0.497	0.496	0.496

Notes: The unit of observation is a paper-patent dyad (n=4,169), including cited paper-patent dyads (n=2,070) and uncited paper-patent dyads (n=2,099). The sample includes 247 paper twins (7 are triplets), 501 papers, and 1,951 citing patents. We identify all cases where a patent cites only one (or a subset) of the twin papers. The binary outcome variable *cited* is equal to 1 in case the paper is cited by the patent, and 0 in case of an uncited twin paper patent dyad, i.e. the uncited twin paper matched to the same patent. The table reports coefficient estimates from conditional logit models with fixed effects at the twin patent-level. Coefficients are in log-odds units. All models include control variables: *degree3_size, no. authors, author-inventor team, publication stock, patent stock, institutional prestige, U.S. based, min geographic distance, journal impact factor, journal commercial impact factor, paper citations, patent citations, time lag, published first, text length and readability, shared PI*. *Shared PI* is an indicator equal to one if any patent inventor and any author of the focal twin paper share an early-career mentor, identified as a last author on their publications during the first five years of their respective publishing careers. *Active_inventor_author* is a dummy equal to 1 if any of the patent's inventors have at least 1 PubMed paper published in the year before patent filing. *Close_paper* is a dummy equal to 1, if the max similarity between inventors' prior papers (published before the patent filing year) and the twin papers (cited or not cited) is higher than sample median. Paper-patent dyads where inventors have 0 prior papers have a value of 0 for cosine similarity. Paper-patent dyads for which cosine similarity could not be calculated due to missing text data are dropped in models 3 and 4. In alternative models, we treated paper–patent dyads involving inventors with no prior papers as missing; the effect sizes and significance levels remained substantively unchanged. All control variables measured in counts are logarithmically transformed after adding one to variables with zero values. Robust standard errors clustered at the twin level and reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level. Mean(cited) is the observed citation probability in each column's estimation sample.