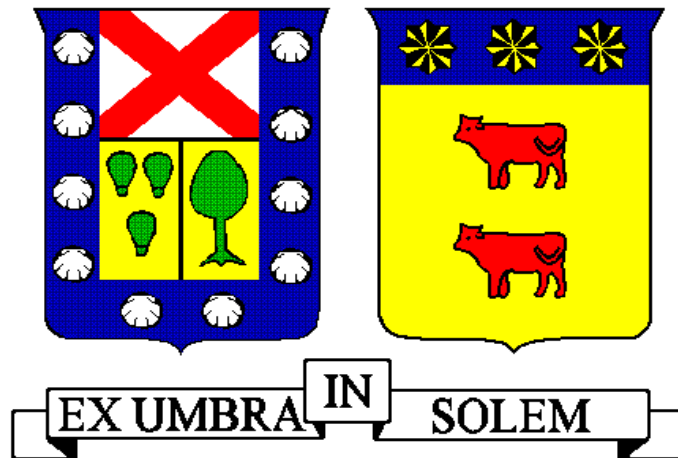


UNIVERSIDAD TECNICA FEDERICO SANTA MARIA
DEPARTAMENTO DE INDUSTRIAS
VALPARAISO - CHILE



El vínculo de las patentes con el éxito de ensayos clínicos: El papel del conocimiento en el desarrollo de fármacos

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL

Autor: Patricio Fernando Vargas Paredes

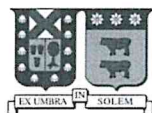
Rol USM: 201960044-3

Profesor guía USM: Pablo Viveros

Profesor guía Politécnico de Milán: Daniela Silvestri

Correferente Politécnico de Milán: Daniel Fernando De Souza

Año Académico: 2025



CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título; ☐ Tesis de Postgrado;

Título del trabajo: El vínculo de las patentes con el éxito de ensayos clínicos: El papel del conocimiento en el desarrollo de fármacos

Nombre del candidato(a): Patricio Fernando Vargas-Paredes

Carrera / Grado: Ingeniería Civil Industrial

Campus: Casa Central Valparaíso ; **Departamento:** Industrias

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Pablo Viveros Gunckel

Yo, _____, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente
DEJO CONSTANCIA que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL

☒ El trabajo **NO contiene información que amerite confidencialidad** y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (embargo) por:

☐ 6 meses; ☐ 12 meses; ☐ 2 años; ☐ 3 años; ☐ 5 años; ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 26.08.25 ; Firma: _____

Estudiante o Candidato(a):

Fecha: 12/08/2025 ; Firma: _____

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Resumen

La innovación es un motor clave del crecimiento económico y el progreso tecnológico, en particular en el sector farmacéutico, donde la intersección de la ciencia y la tecnología es fundamental en el desarrollo de nuevas terapias. Esta tesis examina cómo los conocimientos tecnológicos incorporados y recombinados en las patentes afectan a la probabilidad de que los ensayos clínicos alcancen fases elevadas de desarrollo. Centrándome en el campo de las vacunas de ARNm, desarrollé un método novedoso para emparejar semánticamente patentes y ensayos clínicos utilizando el modelo codificador de lenguaje PAT-SPECTER y la similitud coseno. Esto permitió construir un conjunto de datos emparejados que vinculan el contenido textual de ambas fuentes. Varios indicadores basados en patentes -como el recuento de patentes relacionadas, el tamaño acumulativo de la familia, el ámbito tecnológico (códigos IPC) y el conocimiento previo citado- se evalúan a continuación mediante un modelo de regresión logística multifactorial para determinar su efecto en la progresión de los ensayos. Los resultados muestran una asociación negativa constante entre estas variables relacionadas con las patentes y el avance del desarrollo clínico. Aunque la amplia actividad de patentamiento y la diversidad tecnológica suelen ser indicios de innovación, también pueden reflejar retrasos estratégicos, complejidad de la coordinación o una orientación exploratoria que dificulta la aplicación real en fases posteriores. Esto es particularmente frecuente en campos como el ARNm, donde la densa maraña de patentes y la propiedad fragmentada pueden complicar el desarrollo. Por el contrario, se observan efectos positivos y significativos en las variables institucionales y relacionadas con los ensayos. Las colaboraciones, la madurez institucional y los entornos competitivos de I+D se asocian a una mejor progresión de los ensayos, lo que pone de relieve el valor de la experiencia organizativa, la coordinación multidisciplinar y la urgencia competitiva.

En general, estos resultados sugieren que, si bien el acceso a grandes y amplios repositorios de conocimiento es esencial, el éxito de la innovación clínica puede depender en última instancia de la capacidad de absorber y aplicar ese conocimiento en marcos especializados y orientados a la implementación. Además, esta tesis contribuye a nuestra comprensión de la interacción entre las áreas de conocimiento y las capacidades organizativas, y los efectos que éstas pueden tener en los resultados de la innovación en dominios tecnológicamente complejos.

Contenido

Resumen.....	iii
Contenido.....	iv
Introducción	7
1 Problemática.....	10
2 Objetivo y alcance.....	12
2.1. Objetivo general.....	12
2.2. Objetivos específicos	12
2.3. Alcance del estudio	13
3 Revisión de la literatura.....	14
3.1. El papel del conocimiento previo para la innovación.....	15
3.2. El papel del conocimiento y la recombinação de conocimientos en la comercialización y la innovación disruptiva	19
3.3. Las patentes como proxy del conocimiento	24
3.4. Investigación previa sobre conocimiento e innovación en la industria farmacéutica: Enfoques y conclusiones	25
3.5. Los ensayos clínicos y su importancia como fuente de conocimiento para la innovación.....	32
3.6. PLN, redes neuronales y codificación de documentos	34
3.7. Incrustación de palabras.....	37
3.7.1. Incrustación no contextual de palabras	38
3.7.2. Incrustación contextual de palabras	40
3.7.3. Transformers y BERT	42
3.7.4. SPECTER.....	46
3.7.5. PAT-SPECTER.....	50
3.7.6. El papel fundamental de la selección del modelo.....	51
3.8. Pregunta de investigación.....	52
4 Datos y métodos	53
4.1. Selección del campo del conjunto de datos	53
4.2. Estructura de los ensayos clínicos, recuperación y limpieza del conjunto de datos	54
4.3. Estructura de las patentes, recuperación y limpieza del conjunto de datos	56
4.4. Código para la similitud de documentos.....	60

4.5.	Establecimiento del umbral de similitud del coseno	63
4.6.	Vinculación de ensayos clínicos y patentes.....	66
2.7	Variables.....	67
2.7.1	Variable dependiente	67
2.7.2	Variables independientes.....	67
2.7.3	Variables de control.....	68
2.8	Modelo empírico	69
5	Resultados y discusión	71
3.1	Análisis exploratorio.....	71
3.2	Resultados de la regresión logística	74
3.3	Discusión e implicaciones	78
6	Conclusiones y limitaciones	83
	Bibliografía.....	85
A	Apéndice.....	93
A.1.	Regresión logística con variables base	93
A.2.	Regresión logística con variables independientes log-transformadas .	94
A.3.	Regresión logística con variables de control categóricas activas	95
	Lista de figuras.....	96
	Lista de tablas.....	98
	Agradecimientos.....	99

Introducción

La innovación está ampliamente reconocida como piedra angular del crecimiento económico sostenido y progreso tecnológico. Es la base de las nuevas empresas, la creación de empleo y el aumento de la productividad, haciendo que las economías sean más adaptables, competitivas y resistentes (OCDE y Banco Mundial, 2009). Desde este punto de vista, los académicos y las instituciones políticas suelen hacer hincapié en que la innovación se basa cada vez más en la generación y difusión de conocimientos: los sistemas de innovación sólidos requieren mecanismos sólidos de investigación, capital humano e intercambio de información (OCDE, 2015).

La creciente importancia del conocimiento en la innovación puede observarse en todas las industrias y es especialmente clara en las ciencias de la salud. El sector de ciencias de la salud, que incluye en particular las industrias farmacéutica, biotecnológica y de dispositivos médicos, es bien conocido por tener un alto grado de vinculación con la ciencia y la tecnología (Xu et al., 2025). En el sector de ciencias de la salud, la innovación suele surgir de la recombinación de conocimientos, un proceso en el que los elementos de conocimiento existentes se combinan, modifican o amplían para crear soluciones novedosas. En este contexto, las patentes y los ensayos clínicos, que representan la tecnología y la aplicación o validación, respectivamente, son componentes del conocimiento distintos pero interconectados en el ámbito de la comunicación académica de las ciencias de la salud. Cada uno de ellos sirve a un propósito único y sigue una estructura específica, aunque comparten el objetivo común de difundir conocimientos y hallazgos. Las patentes ofrecen ventajas significativas en la protección de la innovación como documentos legales que establecen derechos de propiedad intelectual para nuevas invenciones y procesos. Pueden revelar las tendencias tecnológicas y las direcciones de la investigación industrial, proporcionando información sobre la comercialización de los hallazgos. Sin embargo, las patentes suelen carecer de los amplios datos experimentales característicos de los resultados de los estudios de ensayos clínicos, lo que puede dar lugar a una laguna en la comprensión detallada de los avances tecnológicos del sector. Por otro lado, los registros de ensayos clínicos proporcionan una visión completa de todos los aspectos de los ensayos, desvelando importantes datos que incluyen discusiones y conclusiones que aportan un mayor contexto. Esos resultados son cruciales para la aprobación regulatoria, la adopción clínica y, por lo tanto, podrían representar un indicio importante sobre las posibilidades reales de comercialización de la tecnología patentada.

Aunque las patentes y los ensayos clínicos cumplen una función única dentro del panorama científico, no se aprovechan plenamente las sinergias potenciales entre ellos. En el pasado, algunos estudios han obtenido conclusiones interesantes utilizando datos de diferentes tipos de literatura, incluidas las patentes y los ensayos clínicos. Por ejemplo, un estudio sobre el uso del polen de abeja ha utilizado artículos académicos, patentes y ensayos clínicos para presentar una visión holística de sus beneficios para la salud (Algethami et al., 2022); y un análisis de citas sobre ensayos clínicos y artículos concluye que los artículos citados por ensayos clínicos tienen más probabilidades de tener impacto a largo plazo (Meyer, 2006). A pesar de estos interesantes descubrimientos, el uso directo de ambas fuentes de información combinadas no se ha utilizado en una gama de aplicaciones significativamente más amplia.

Uno de los retos fundamentales del sector de las ciencias de la salud es determinar lo cerca que está una tecnología o un producto innovador de su comercialización. Aunque cada año se producen numerosos descubrimientos científicos, sólo una fracción de ellos pasa de la investigación a las aplicaciones en el mundo real. En este ámbito, el presente estudio propone una medida sustitutiva de la comercialización basada en la vinculación entre los ensayos clínicos y las patentes, que ayuda a identificar los productos que están más cerca de su introducción en el mercado en la industria farmacéutica. Para investigar el método y con el objetivo de establecer este vínculo, se revisaron varios estudios anteriores. Se han desarrollado enfoques para medir el vínculo entre la ciencia (artículos) y la tecnología (patentes) basados en la información de las citas de patentes con literatura no relacionada con patentes, principalmente artículos de investigación, y en última instancia utilizando diferentes métodos de procesamiento del lenguaje natural (PLN) que analizan la semántica y el contenido de ciertas partes de los documentos para establecer una similitud y posteriormente un vínculo. Estos últimos enfoques también se han utilizado para relacionar patentes con patentes con el fin de realizar tareas de clasificación, recuperación de información concreta y paisajes de patentes. Sin embargo, en el caso de la relación ensayo clínico-patente, aunque existe alguna información sobre ensayos clínicos citados en patentes, tal escenario es particularmente raro, y otros métodos de vinculación no han investigado esta relación de conocimiento. En consecuencia, necesitamos métodos o herramientas para procesar y analizar los contenidos textuales de textos de gran volumen y establecer medidas de relación entre ellos. El problema que se plantea es cómo aplicar el concepto de similitud entre documentos. Para ello es necesario representar los documentos en una forma numérica, con la que un ordenador pueda procesar para generar representaciones similares para documentos similares a gran escala. En el campo del procesamiento del lenguaje natural (PLN), este proceso se conoce como codificación de documentos. Aunque existen varios métodos para representar numéricamente los documentos, como el tradicional TD-IDF (Sparck Jones, 1972) ampliamente utilizado en la literatura, estos métodos tradicionales no son escalables, ya que una ampliación del vocabulario requiere un nuevo cálculo de todas

las representaciones del texto. Además, estos enfoques no capturan la semántica de los documentos; es decir, se pierde el significado de las palabras o la interpretación de las frases en su contexto. Por lo tanto, se eligió PAR-SPECTER (Ghosh et al., 2024), un modelo de lenguaje para codificar publicaciones científicas y patentes. Siendo el lenguaje de última generación para la codificación y el cálculo de similitud entre documentos, este lenguaje aprovecha la arquitectura de transformador bidireccional (BERT) (Devlin et al., 2019). Por lo tanto, mediante la utilización de este modelo de codificador de lenguaje y el desarrollo de un algoritmo novedoso para procesar y tratar los corpus de texto de patentes y ensayos clínicos, esta tesis tiene como objetivo realizar la coincidencia, descubrir el vínculo y analizar a través de indicadores basados en patentes, ensayos y firmas- , cómo el conocimiento incorporado en las patentes influye en alcanzar fases avanzadas de desarrollo de ensayos clínicos en la industria farmacéutica.

Esta tesis está organizada de la siguiente manera. En el capítulo 1, la revisión bibliográfica presenta, en primer lugar, una visión general y una contextualización de la importancia del conocimiento y la recombinación de conocimientos para la innovación, en términos de investigación, desarrollo y comercialización de productos. Posteriormente, cómo se ha emparejado la interacción entre conocimiento e innovación en la industria farmacéutica, y cómo las patentes y los ensayos clínicos sirven como fuente de conocimiento. En la segunda parte, explico el modo de funcionamiento y discuto los métodos pasados y actuales de PLN, sus limitaciones y aplicaciones relacionadas con el campo y la selección del modelo que permitirá obtener las correspondencias patente-ensayo clínico mediante el cálculo de la similitud coseno entre el cuerpo de texto del ensayo clínico y la patente. En el Capítulo 2, ilustro el campo de estudio dentro de la industria farmacéutica, la metodología para obtener, extraer, codificar y limpiar los datos recogidos de PATSTAT online y clinicaltrial.gov, y el proceso seguido para desarrollar el código de similitud. Posteriormente, cómo se utilizaron estos datos con el código de similitud para obtener las similitudes entre los documentos y establecer las variables con el fin de ser analizadas. El capítulo 3 muestra los resultados de la investigación y del modelo utilizado para su interpretación. En el último capítulo, reflexiono sobre los resultados del trabajo, señalando sus limitaciones y aportando algunas reflexiones finales.

1 Problemática

La innovación en la industria farmacéutica ha demostrado ser un factor crítico para el desarrollo económico, el avance científico y la mejora en la calidad de vida. Dada su naturaleza intensiva en conocimiento, esta industria depende fuertemente de la generación, acumulación y aplicación de información tecnológica y clínica, muchas veces representada en documentos estructurados como las patentes y los registros de ensayos clínicos. A pesar de que ambos instrumentos cumplen funciones distintas en el ciclo de innovación, su potencial de complementariedad es evidente. Sin embargo, la relación entre el conocimiento tecnológico protegido mediante patentes y el éxito en el desarrollo clínico de nuevos productos no ha sido suficientemente explorada desde una perspectiva integradora.

Tradicionalmente, el análisis de innovación ha vinculado el conocimiento científico y tecnológico mediante el uso de citas entre publicaciones académicas y patentes. Estas aproximaciones han sido útiles para entender flujos de conocimiento, pero no resultan aplicables de forma directa en el caso de los ensayos clínicos, que por su estructura y propósito no suelen establecer vínculos explícitos con documentos de propiedad intelectual. Esta ausencia de referencias cruzadas limita la capacidad de trazar conexiones directas entre los desarrollos tecnológicos y su validación clínica, dificultando la evaluación temprana del potencial de comercialización de nuevas terapias o productos.

Por otra parte, el campo de las tecnologías ARNm ha presentado un crecimiento exponencial en cuanto a producción de conocimiento y actividad patentada, muchas veces en paralelo a desarrollos clínicos. Esta aceleración ha generado estructuras complejas de propiedad intelectual, donde se observa alta densidad de patentes y fragmentación de titulares. Estos factores pueden generar barreras al desarrollo clínico, ya sea por dificultades de coordinación, incertidumbre legal o simplemente por redundancia tecnológica. A pesar de esto, no existe aún un consenso sobre cómo estas dinámicas influyen realmente en la progresión de las tecnologías hacia fases más avanzadas de desarrollo clínico.

En este escenario, el problema central se ubica en la falta de mecanismos robustos para cuantificar y analizar la relación entre el conocimiento tecnológico expresado en patentes y el avance real de las innovaciones farmacéuticas en su etapa de validación clínica. Esta brecha de conocimiento impide no solo una mejor comprensión de las dinámicas de innovación en el sector, sino también la identificación temprana de

factores que podrían favorecer o dificultar la transición desde la invención hasta una eventual introducción al mercado.

Por lo tanto, se reconoce la necesidad de una aproximación que permita evaluar dicha relación desde un enfoque distinto al tradicional, considerando el contenido semántico de los documentos como vía para establecer posibles correspondencias. Esto permitiría avanzar hacia una visión más integrada de la innovación farmacéutica, donde el conocimiento técnico y el proceso clínico se analicen como elementos interdependientes dentro de un mismo sistema de desarrollo tecnológico.

2 Objetivo y alcance

2.1. Objetivo general

Analizar la relación entre el conocimiento tecnológico contenido en patentes y la probabilidad de avance de los ensayos clínicos hacia fases más cercanas a la comercialización, utilizando como caso de estudio el campo de las tecnologías ARNm y aplicando técnicas de procesamiento de lenguaje natural para vincular semánticamente ambos tipos de documentos.

2.2. Objetivos específicos

- Desarrollar un procedimiento de vinculación semántica entre patentes y ensayos clínicos mediante el uso del modelo de codificación PAT-SPECTER y la métrica de similitud coseno.
- Construir un conjunto de datos emparejado que permita relacionar información textual entre patentes y ensayos clínicos de manera sistemática.
- Identificar y extraer indicadores relevantes basados en patentes, tales como el número de documentos relacionados, el tamaño acumulado de la familias, la diversidad tecnológica y las citas de conocimiento previo.
- Evaluar, mediante un modelo de regresión logística multifactorial, el efecto de estos indicadores sobre la probabilidad de que un ensayo clínico alcance fases avanzadas de desarrollo.
- Interpretar los resultados obtenidos a la luz de las dinámicas de innovación en el sector farmacéutico, destacando patrones que puedan contribuir a entender los factores que influyen en el éxito clínico de tecnologías emergentes.

2.3. Alcance del estudio

Este trabajo se enmarca dentro del ámbito de la innovación farmacéutica, con especial énfasis en tecnologías basadas en ARNm, un campo altamente dinámico y caracterizado por la intensa actividad patentada y clínica. El estudio considera únicamente información textual extraída de fuentes públicas: ClinicalTrials.gov para ensayos clínicos, y PATSTAT online para documentos de patente, lo que implica un enfoque limitado al contenido accesible mediante estos sistemas.

El análisis se restringe a la relación entre características tecnológicas (proxy mediante patentes) y el avance de ensayos clínicos, sin evaluar de forma directa resultados clínicos, eficacia terapéutica ni decisiones regulatorias. Asimismo, se centra exclusivamente en la fase de desarrollo clínico, por lo que no considera etapas posteriores como la comercialización efectiva o el desempeño del producto en el mercado.

Desde el punto de vista metodológico, el enfoque de emparejamiento semántico propuesto representa una aproximación exploratoria, basada en codificadores lingüísticos de última generación, que permite aproximarse al concepto de recombinación de conocimiento desde una perspectiva cuantificable y replicable. No obstante, los resultados obtenidos deben ser interpretados en función de las limitaciones inherentes al método de vinculación y a la calidad del contenido textual disponible.

3 Revisión de la literatura

En este capítulo se analiza la importancia del conocimiento y la recombinación de conocimientos en la innovación, un vínculo importante que, según se ha comprobado, aumenta la ventaja competitiva de las empresas y los beneficios derivados del lanzamiento de nuevos productos (Wagner y Wakeman, 2016; Zhong et al., 2024; Kaplan y Vakili, 2015). La revisión comienza estableciendo cómo las empresas están utilizando el conocimiento en forma de patentes y otras fuentes, como el conocimiento científico, cómo este proviene de dentro y fuera de las organizaciones, y cómo este conocimiento acelera la tasa de innovación y allana el camino para obtener innovaciones más impactantes. Posteriormente, se estudia el efecto de recombinación del conocimiento, debido a que es una de las fuentes más importantes para obtener innovaciones disruptivas y que cada vez es más importante para las empresas (Baker et al., 2014). Este análisis se lleva a cabo para descubrir el peso y la variación de los diferentes tipos de conocimientos utilizados en los ensayos clínicos en fase 3 o más (que representan la proximidad al mercado) en forma de tecnología patentada y detectar de qué manera estos conocimientos favorecen la comercialización de nuevos productos. Este último propósito nos lleva a la segunda parte de esta revisión, que se centra en el procesamiento del lenguaje natural (PLN) y la codificación de documentos. En esta parte, se analizan y discuten los principales métodos y modelos utilizados para generar incrustaciones de palabras, landscaping, similitudes y recuperación de documentos que incluyen literatura científica, para llegar finalmente al modelo utilizado para calcular esta similitud entre ensayos clínicos y patentes, para así obtener una aproximación a esta recombinación de conocimiento. A través de las secciones de la segunda parte, funcionamiento de las redes neuronales por su importancia para el entrenamiento de los modelos, seguido de la literatura de los métodos de incrustación de palabras no contextuales y contextuales para la posterior codificación completa de textos, junto con sus aplicaciones pasadas más importantes que me dejaron importantes datos. Para terminar finalmente con el estado del arte y la sección del modelo SPECTER, cómo fue entrenado, y cómo condujo al modelo más avanzado para el cálculo de similitud de documentos, PAT-SPECTER (Ghosh et al., 2024), que se utilizó para realizar este trabajo. También se indaga en la aplicación previa de PAT-SPECTER con "Logic Mill". Debido a que el proceso de codificación del texto y cálculo de la similitud se basó en el método utilizado por los autores de "Logic Mill".

3.1. El papel del conocimiento previo para la innovación

La innovación es una de las principales vías por las que las empresas nuevas y existentes crean y mantienen una ventaja sobre sus competidores, y es un factor clave para establecer un crecimiento económico sostenido y el progreso tecnológico. El comité asesor estadounidense para la medición de la innovación la define abiertamente como el diseño, la invención, el desarrollo y/o la aplicación de productos, servicios, procesos, sistemas, estructuras organizativas o modelos de negocio nuevos o modificados para crear nuevo valor para los clientes y rendimientos financieros para la empresa (Advisory Committee on Measuring Innovation in the 21st Century Economy, 2008). Sin embargo, el éxito de las distintas innovaciones depende de los conocimientos -científicos, tecnológicos, estratégicos y relacionados con el mercado (Roper & Hewitt-Dundas, 2015)-, por la razón de que las innovaciones se construyen sobre conocimientos y tecnologías anteriores y sirven en sí mismas como arte previo para las generaciones futuras (Weitzman, 1998). Centrándonos en el conocimiento científico y tecnológico, se ha estudiado ampliamente que ambas fuentes de conocimiento, la ciencia en forma de artículos (entre otros) que expresan un tema de investigación concreto y la tecnología en forma de patentes que protegen una propiedad intelectual concreta con interés económico para las empresas, son motores importantes y a la vez interconectados para la innovación y el crecimiento económico (Romer, 1989). De este modo, los investigadores han llegado a la conclusión, utilizando diferentes formas de relacionar patentes y documentos, como el método tradicional de utilizar citas y referencias, entre otros, de que las invenciones patentadas que se basan directamente en avances científicos tuvieron un impacto especial y un valor económico esperado superior en comparación con otras patentes. Por otro lado, los trabajos citados directamente por patentes eran también los de mayor impacto dentro del ámbito científico (Ahmadpoor & Jones, 2017). O más aún, que existe una relación robusta y fuerte entre la alta calidad científica de una publicación referenciada en una patente (medida en términos de citas científicas e impacto en revistas) y el impacto y valor comercial de la patente (medido en términos de citas de patentes, alcance y retornos bursátiles) (Poege et al., 2019). A pesar de estas conclusiones, la importancia del conocimiento no depende únicamente de la calidad y la aceptación pública. Como ya se ha señalado, las innovaciones se construyen sobre conocimientos previos, y la difusión de los hallazgos desempeña un papel fundamental en el ritmo de desarrollo de la innovación tecnológica, abriendo la información recién adquirida a un público mucho más amplio de inversores que potencialmente podrían basarse en ella, permitiendo a la comunidad desarrollar tecnologías más rápidamente reduciendo también la probabilidad de duplicación de esfuerzos (Dasgupta y David, 1994). Sorenson y Fleming (2004) estudiaron el papel de la difusión del conocimiento a través de la publicación de patentes, literatura científica y no científica. El estudio analizó las citas que las patentes reciben de patentes futuras y las citas de estas patentes futuras a un conjunto de patentes focales que hacen referencia o no a varios tipos de materiales

publicados. En concreto, los autores compararon tres grupos de patentes focales: las que se basan en investigaciones científicas publicadas (revisadas por pares), las que hacen referencia a otros tipos de conocimientos publicados (no científicos) y las que no hacen referencia a ninguna publicación no relacionada con patentes, como puede verse en la figura 1.

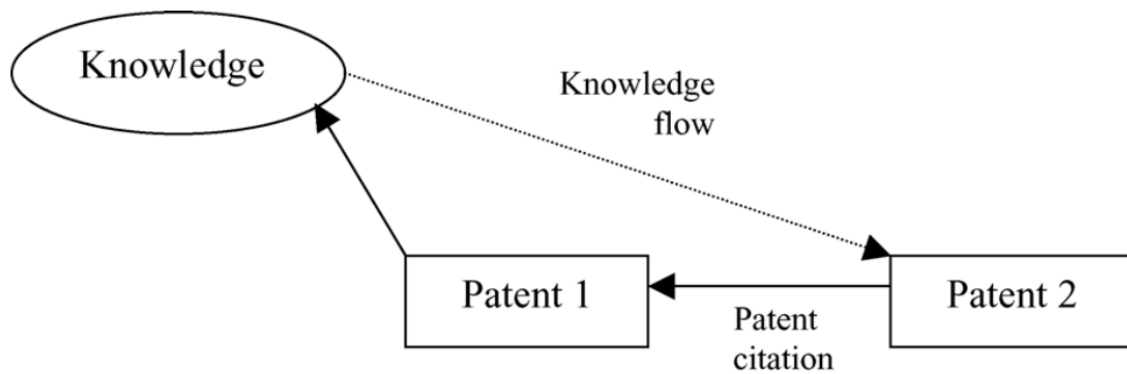


Figura 1: Modelo de flujos de conocimiento (Fuente: Sorenson & Fleming, 2004)

Sus resultados mostraron que las patentes que hacen referencia a materiales publicados, ya sean académicos o no (como revistas populares y materiales de marketing), reciben más citas, antes y de inventores más alejados geográfica y socialmente que las que no hacen referencia a material no patentado, lo que concuerda con la idea de que la publicación acelera la difusión de la información. Aunque estos resultados aportaron pruebas empíricas útiles de la importancia de la difusión de la información como mecanismo subyacente al vínculo entre la investigación científica y las tasas de innovación, y de que la rápida difusión del conocimiento puede mejorar la eficiencia de las inversiones investigativas al reducir el grado de duplicación de esfuerzos. En la medida en que esta reducción del solapamiento de las actividades de investigación estimula las tasas de innovación (y potencialmente el crecimiento económico) (Sorenson & Fleming, 2004).

Para sacar partido de estos flujos de conocimiento, las empresas adquieren estos importantes recursos de conocimiento internamente a través de la investigación corporativa en laboratorios y externamente a través de la colaboración con instituciones académicas, adquisiciones de start-ups intensivas en ciencia, y mediante licencias y contratos con otras empresas (Arora, Belenzon, & Pataconi, 2018). Incluso teniendo en cuenta la amplia gama de estudios que apoyan la idea de que el conocimiento científico afecta positivamente a las tasas de innovación, que las invenciones empresariales siguen recurriendo a la ciencia y que la cosecha de conocimientos científicos utilizados en la invención no ha cambiado con el tiempo (Arora et al., 2018), la disposición de las grandes empresas a participar en la investigación científica interna ha disminuido, véanse las figuras 2 y 3 a continuación.

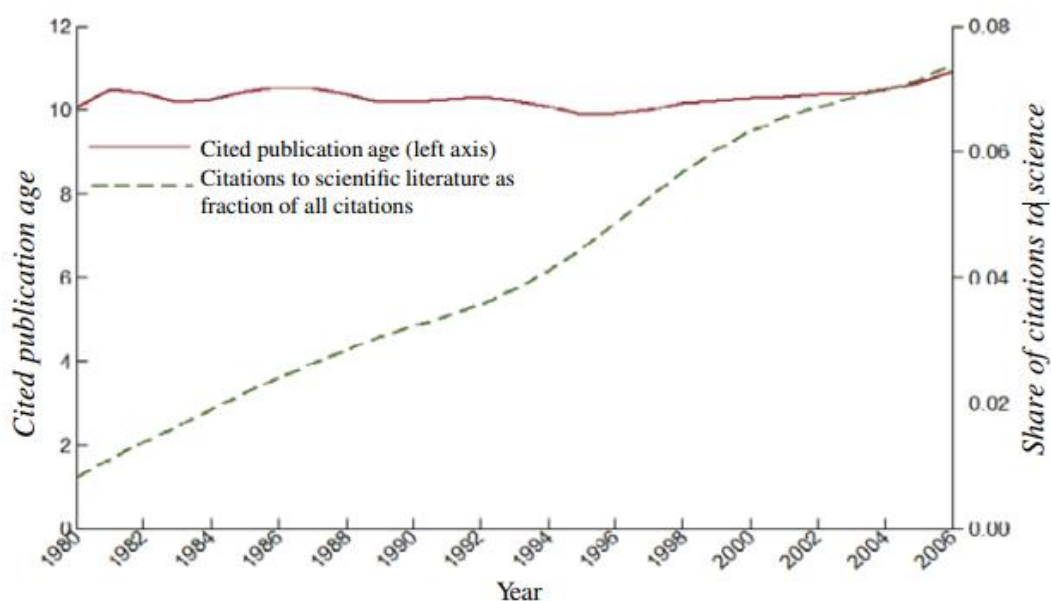


Figura 2: Uso de la ciencia en la innovación: edad media de la ciencia citada en patentes y media de citas por patente, 1980-2006. (Fuente: Arora et al., 2018)

Nota: Esta figura presenta la edad media de publicación de los artículos citados por las patentes y la proporción de citas a la ciencia por empresa-año para nuestras empresas de la muestra. La edad de publicación citada es la diferencia entre el año de concesión de la patente y el año de publicación del artículo citado. La proporción de citas científicas es la relación entre las citas de patentes en las principales revistas científicas y el número total de referencias que hace la patente.

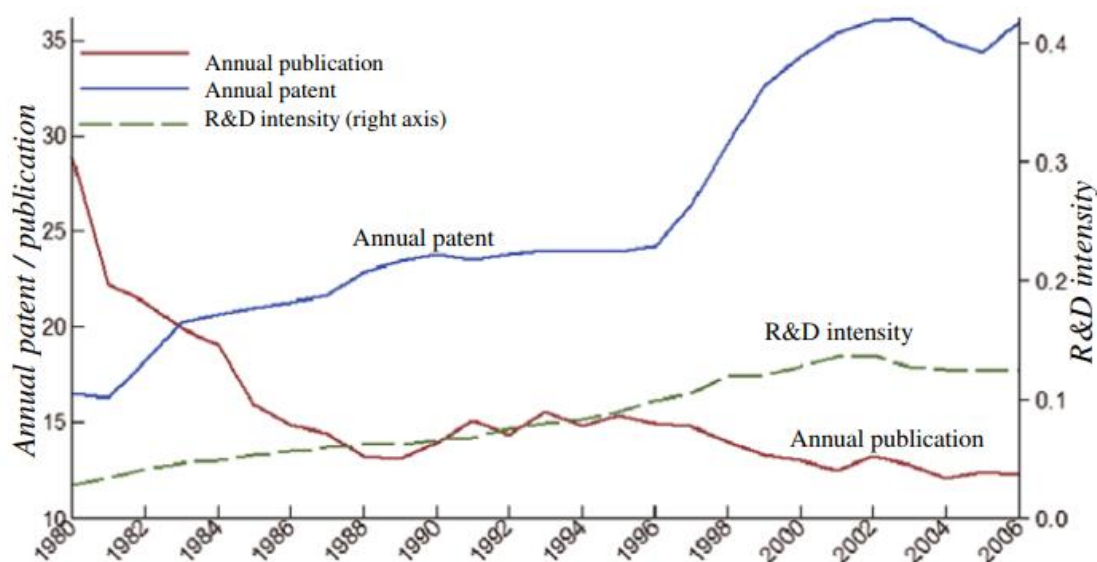


Figura 3: Publicación, patentes e investigación: Empresas estadounidenses, 1980-2006. (Fuente: Arora et al., 2018)

Nota: Esta figura presenta el promedio de publicaciones y patentes de la empresa a lo largo del tiempo para las empresas de Compustat con al menos 1 año de gastos positivos en I+D y al menos una patente. La intensidad de I+D es el gasto en I+D sobre las ventas. La publicación anual está condicionada a al menos una publicación científica.

Arora et al. (2018) analizaron en su estudio el declive de la investigación científica en la I+D empresarial y clasificaron las posibles razones en tres grandes categorías (a) una disminución del valor privado de la investigación interna, (b) un aumento del coste de la investigación y (c) cambios en las prácticas de publicación. Las dos primeras categorías engloban una serie de fuerzas institucionales, tecnológicas y económicas que pueden haber inducido a las grandes empresas a reducir sus inversiones en ciencia. La tercera categoría, cambios en las prácticas de publicación, sugiere que el descenso en la producción de publicaciones puede reflejar no cambios en la composición de los esfuerzos de I+D de las grandes empresas, sino más bien un rechazo de la "ciencia abierta" en favor de un mayor secretismo. Además, la mayor afirmación del artículo es que las empresas están pasando a un gasto mínimo de investigación interna que les permite aprovechar los conocimientos generados externamente para impulsar su crecimiento. En apoyo de estas conclusiones, Arora et al. (2016) documentaron que casi el 50% de las innovaciones de productos introducidas por las empresas manufactureras estadounidenses se basaban en invenciones externas. Aunque las universidades, las start-ups y otros especialistas en tecnología aportan colectivamente solo el 17 % del total de invenciones, estas son las de mayor valor.

Volviendo a la tercera razón que explica el declive de la investigación en la I+D empresarial, los "cambios en las prácticas de publicación" se explican en profundidad por la compensación de dos fuerzas opuestas para las empresas: por un lado, los beneficios derivados del uso de la ciencia en las invenciones propias en sentido descendente y, por otro, los costes de los derrames de conocimiento para los rivales. Los nuevos productos y procesos desarrollados pueden protegerse de posibles imitadores mediante patentes, derechos de autor y secretos comerciales, mientras que la investigación suele divulgarse en publicaciones científicas, incluso cuando la llevan a cabo científicos de la empresa. Por lo tanto, es más probable que la investigación genere derrames de conocimiento. Cuando estos derrames llegan a los rivales, no son meras externalidades, sino que pueden reducir el rendimiento privado de la investigación. Las empresas invierten cada vez menos en investigación. La disminución de la participación de las empresas en la ciencia, incluso cuando las propias invenciones dependen cada vez más de la ciencia, es contradictoria. Incluso cuando las empresas hacen un mayor uso del conocimiento científico, están menos dispuestas a producir dicho conocimiento, prefiriendo desplazar la atención y los recursos de la investigación previa al desarrollo posterior (Ahora et al., 2021). La figura

4 muestra que la propensión de las patentes a citar la ciencia corporativa (medida como la relación entre las citas de la ciencia corporativa por patente y el número total de publicaciones corporativas disponibles en un año determinado) ha ido aumentando con el tiempo tanto para las citas internas como para las rivales. Pero mientras que las propensiones a citar internamente y por parte de rivales son idénticas en 1990, en 2015 las citas por parte de rivales pasan a ser dos veces más numerosas que las citas internas.

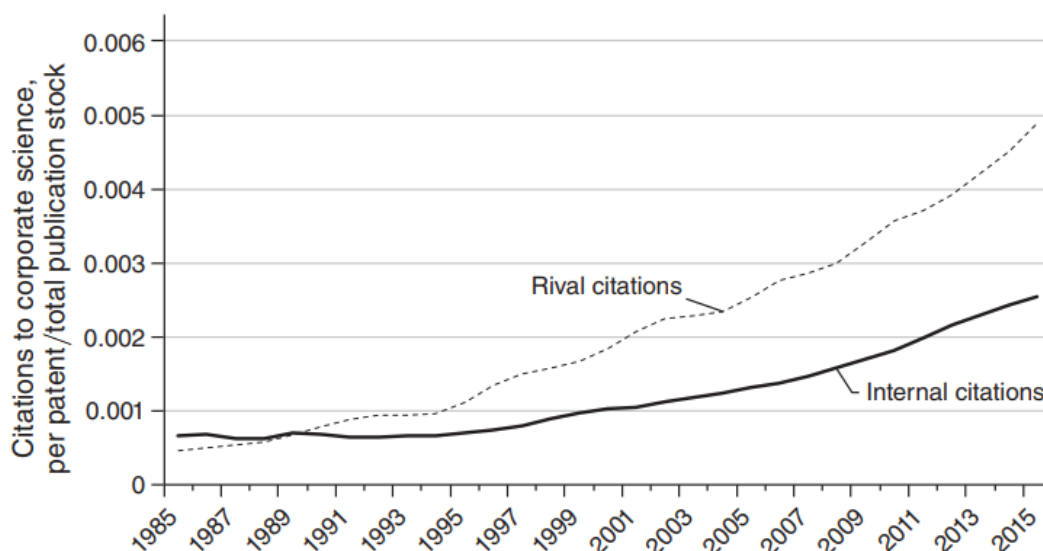


Figura 4: Tendencias en el uso de la ciencia por las patentes corporativas, 1985-2015. (Fuente: Ahora et al., 2021)

Nota: Este gráfico presenta las tendencias en la propensión de las patentes empresariales a citar la ciencia empresarial, medida como el cociente entre las citas a la ciencia por patente empresarial dividido por el número total de publicaciones empresariales en cada año (los valores *del eje y* se multiplican por 1.000). La muestra está condicionada a observaciones de años de empresa con al menos una patente concedida. Las citas rivales ponderan las citas externas por la proximidad del mercado de productos entre la empresa que cita y la citada.

3.2. El papel del conocimiento y la recombinación de conocimientos en la comercialización y la innovación disruptiva

Conociendo la importancia actual de los flujos externos de conocimiento, que mejoran la eficiencia de las inversiones en investigación (Sorenson & Fleming, 2004), y tanto la probabilidad de innovar como el éxito de la innovación (Roper & Hewitt-Dundas, 2015), podemos rastrear el efecto del conocimiento directamente al ámbito de la comercialización. Como es evidente en el resto de la literatura considerada, los outputs de la innovación se miden en base a las patentes, pero al mismo tiempo, las posibilidades de comercialización de los productos podrían estar correlacionadas con ciertos indicadores basados en patentes. Esta posibilidad se explora en el estudio

realizado por Wagner y Wakeman (2016), en el que investigan cómo los indicadores de patentes de uso frecuente asociados al valor de una invención y a la incertidumbre sobre el alcance y la fuerza de la protección de patentes están relacionados con los resultados reales de los procesos de comercialización de productos. Mediante la vinculación de los resultados de la comercialización de productos en la industria farmacéutica con datos detallados sobre patentes, relacionaron los indicadores basados en patentes que captan el valor de una invención o la incertidumbre que rodea el proceso de patentamiento con los resultados del proceso de desarrollo del producto. Los resultados sugieren que las medidas relacionadas con el valor, como el número de patentes diferentes en torno a un proyecto de comercialización (que también puede interpretarse como un indicador que refleja la evaluación por parte de una empresa del valor esperado de un proyecto), el número de países en los que se han solicitado patentes primarias (la ampliación de la protección de patentes a diferentes países es costosa y los solicitantes están dispuestos a incurrir en estos costes adicionales en caso de invenciones valiosas) y diferentes medidas basadas en citas, todas ellas relacionadas con un aumento de la probabilidad de una mayor velocidad de comercialización de los productos (Wagner y Wakeman, 2016).

Tabla 1: Número de productos por etapas del proceso de desarrollo de productos (Fuente: Wagner & Wakeman, 2016)

Development stage started	# Products	Reaching stage (%)	Year of invention (estimated)	Age at reaching stage (years)	# Primary patents/product	# of EP patents	Granted (%)	Opposed (%)
Total	5923	100	1991.4	–	1.12	1.58	68.7	7.05
Stage reached								
Preclinical	5685	96.0	1991.3	4.57	1.12	1.58	69.4	6.83
Phase 1	3653	61.7	1990.8	5.93	1.16	1.66	73.7	7.93
Phase 2	2953	49.9	1990.5	7.13	1.18	1.69	75.1	8.50
Phase 3	1565	26.4	1989.7	8.79	1.24	1.82	78.0	11.43
Launched in EP country	630	10.6	1988.5	11.31	1.30	1.92	82.0	17.66

Otra conclusión importante es el hecho de que para los productos con un valor más elevado y/o una mayor incertidumbre de sus especificaciones finales, es crucial conocer con detalle los derechos de patente (y que el solicitante se forme expectativas sobre la dificultad de la imitación o para inventar en torno a la patente). En estos casos, el solicitante podría posponer las decisiones relativas al desarrollo del producto hasta que se haya resuelto la incertidumbre. En esta línea, el indicador variable en el tiempo formulado por los autores capta el momento exacto en que se concede la patente principal asociada a un proyecto de comercialización e indica que la resolución de la incertidumbre acelera significativamente el proceso de comercialización.

Más allá de alcanzar la comercialización, algunas innovaciones pueden transformar la dirección del desarrollo tecnológico introduciendo nuevos paradigmas y abriendo un nuevo camino para las tecnologías posteriores, lo que tiene un profundo impacto en las empresas, las industrias y los mercados. Este tipo de innovaciones se conocen como innovaciones disruptivas (Zhong et al., 2024). Este tipo de innovaciones pueden tener implicaciones dramáticas, permitiendo a las organizaciones y empresas encontrar nuevas oportunidades de crecimiento, reinventarse, desafiar los órdenes competitivos existentes, e incluso redefinir paisajes industriales enteros (Kirchhoff, 1991), por estas

razones las innovaciones disruptivas son cada vez más valoradas por los directivos y se convierten gradualmente en una dirección significativa para la planificación estratégica organizativa y la implementación (Baker et al., 2014). Los estudiosos han sostenido durante mucho tiempo que la innovación disruptiva, al igual que la innovación incremental, se produce a través de la recombinación de ideas existentes. Pero a diferencia de la innovación incremental, la innovación disruptiva suele caracterizarse por la recombinación de ideas procedentes de ámbitos del conocimiento distantes, heterogéneos y previamente desconectados (Haas y Ham, 2015; Zhang, Wang y Xu, 2021). Esta última afirmación también cuenta con el apoyo de Capponi et al. (2022), que recientemente señalaron que la innovación disruptiva se origina en dos etapas, a saber, la etapa de búsqueda y la etapa de selección, que depende de la recombinación de conocimientos. Por tanto, se concluye que las empresas deben desarrollar capacidades para combinar de forma innovadora los conocimientos existentes y evaluar la nueva combinación de conocimientos, determinando así la cartera de propiedades intelectuales óptima que se ajuste a sus requisitos de innovación.

Sabiendo que está significativa y positivamente correlacionada con la innovación rompedora de las empresas (Zhong et al., 2024; Kaplan & Vakili, 2015), la recombinación de conocimientos se define como el proceso de integración creativa de componentes de conocimiento de diferentes dominios para generar conocimiento tecnológico novedoso. Este enfoque se ha convertido en el principal paradigma de la innovación tecnológica. En el contexto de la gestión de la innovación tecnológica, la recombinación de conocimientos se manifiesta como convergencia tecnológica (también conocida como fusión tecnológica), que implica la fusión de múltiples tecnologías para formar una nueva tecnología híbrida capaz de abordar problemas complejos (Xie et al., 2025).

La teoría de la recombinación de conocimientos tiene su origen en la observación de Schumpeter: "la innovación combina componentes de una manera nueva, o que consiste en llevar a cabo nuevas combinaciones" (Schumpeter, 1934). Siguiendo esta teoría, y como ya se ha mencionado, numerosos estudios sobre innovación han demostrado que las nuevas invenciones surgen esencialmente de la recombinación de componentes del conocimiento ya existentes. Este es el caso del estudio realizado por Xie et al. (2025), que examina cómo la recombinación de conocimientos impulsa la innovación tecnológica en el campo farmacéutico empleando el análisis de redes y el aprendizaje automático explicable para investigar el valor predictivo de la recombinación de conocimientos sobre el rendimiento de la innovación. El estudio llevó a la conclusión de que más del 90% de las patentes de las últimas décadas consisten en híbridos de tecnologías existentes y, lo que es más importante, "la evolución dinámica de las redes demuestra además que los patrones de recombinación de conocimientos han permanecido estacionarios, con pocas combinaciones novedosas introducidas año tras año" (Xie et al., 2025), véase la figura 5 a continuación.

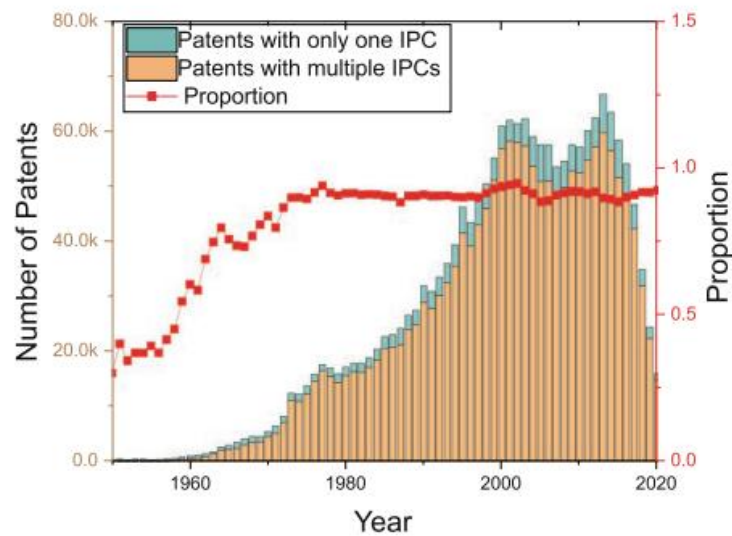


Figura 5: Curva de distribución año-patente en el campo farmacéutico (Fuente: Xie et al., 2025)

Además, los autores también afirman que cuanto más concentrado en las carteras de patentes existentes esté el conocimiento recombinao, más probable será que la patente sea disruptiva. Esta última afirmación es importante porque, al buscar nuevos conocimientos para la recombinación, las empresas no sólo deben prestar atención a la amplitud de la búsqueda de conocimientos (conocimientos diversificados y diferenciados), sino también a la profundidad de los conocimientos (conocimientos de vanguardia en la dirección de la profundidad, que pueden acelerar la transformación y la utilización de los conocimientos existentes en la empresa y de los conocimientos externos). Así, pueden enriquecer el stock de recursos de conocimiento heterogéneo aumentando la amplitud de búsqueda y obtener el conocimiento de vanguardia reforzando la profundidad de búsqueda del conocimiento (Zhang, Wang y Xu, 2021; Kaplan y Vakili, 2015). Esta diversificación del conocimiento y la profundidad de la recombinación pueden considerarse también en diferentes países y entre pequeños y grandes innovadores, junto con el hecho de que las empresas extienden sus actividades innovadoras a través de campos tecnológicos relacionados entre sí como consecuencia de los procesos de aprendizaje (ya sea no intencionado en términos de derrame de información, o intencionado en términos de aprendizaje local) y de características y vínculos del conocimiento (debido al alcance, la complementariedad o la naturaleza genérica del conocimiento) (Breschi et al., 2003). Las figuras 6 y 7 muestran que muy pocas empresas están diversificadas en la mayoría de los campos tecnológicos, y que éstas son las que más innovan.

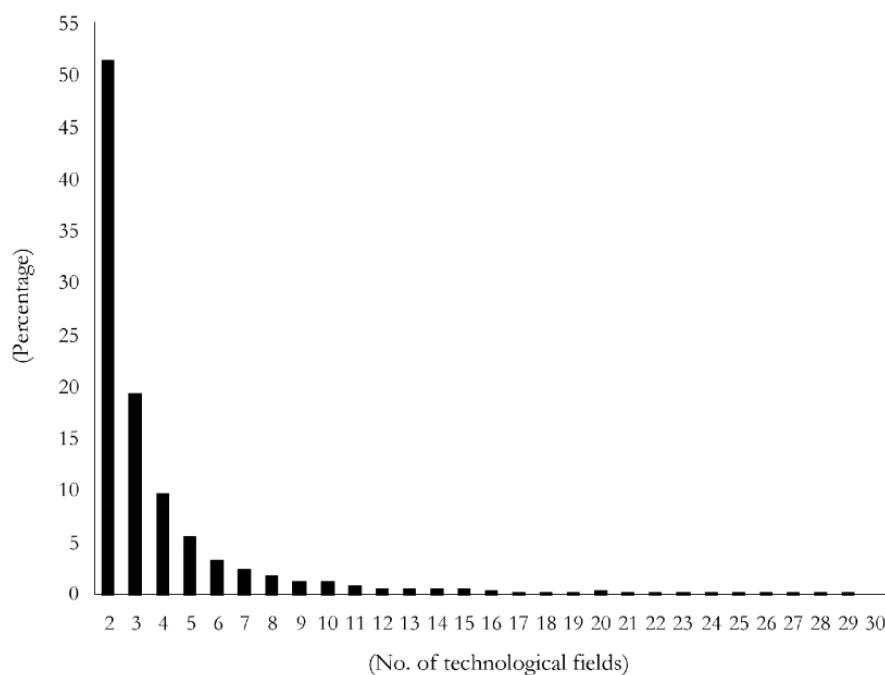


Figura 6: Distribución de los innovadores diversificados según el número de campos tecnológicos en los que solicitaron patentes (valores porcentuales, 1982-1993) (Fuente: Breschi et al., 2003)

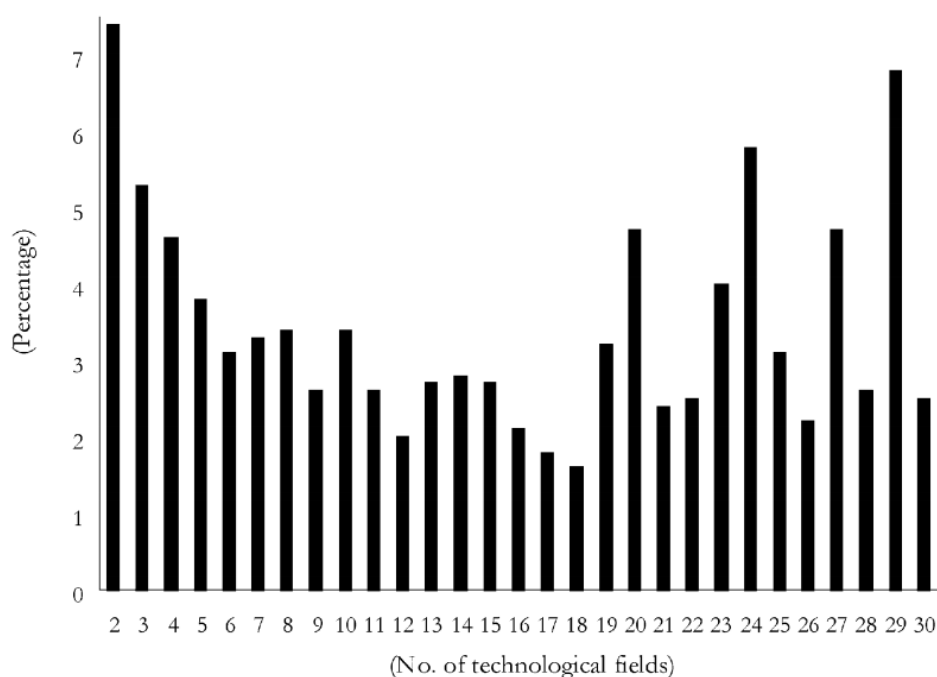


Figura 7: Distribución del total de patentes solicitadas por innovadores diversificados según el número de campos tecnológicos en los que posteriormente solicitaron patentes (valores porcentuales, 1982-1993) (Fuente: Breschi et al., 2003).

Teniendo en cuenta los resultados anteriormente mencionados, es clara la importancia del conocimiento en la generación y el ritmo de la actividad innovadora, y que el uso

de diferentes tipos de conocimiento a través de la recombinación presente en las diferentes patentes y productos consecutivos tiene un efecto sustancial en su comercialización, éxito y posible carácter disruptivo. Como establece la literatura, las patentes y las publicaciones científicas son las fuentes y recursos de conocimiento para la innovación más analizadas. Dependiendo de la industria en la que se inserten las empresas, existen diferentes fuentes útiles, como los ensayos clínicos para la industria farmacéutica (Leten et al., 2022). En este trabajo de tesis, la atención se centrará en el análisis de las patentes y ensayos clínicos. Los detalles de su importancia y su contribución a la comercialización de productos farmacéuticos se abordarán en las siguientes secciones.

3.3. Las patentes como proxy del conocimiento

Las patentes sirven como documentación formal de las invenciones, proporcionando información detallada sobre nuevas tecnologías y procesos. Como fue mencionado en la sección anterior, su naturaleza estructurada permite a los investigadores analizar las tendencias tecnológicas, la difusión del conocimiento y la producción innovadora de empresas o regiones.

Entrando más en detalle, las patentes proporcionan una ventana a la generación y difusión del conocimiento a través de sus referencias a otros materiales. Sin embargo, también deberían proporcionar un ámbito especialmente difícil en el que encontrar beneficios asociados a la divulgación pública del conocimiento. La principal razón para ofrecer patentes y los monopolios de duración limitada que conceden, se basa en la idea de que la sociedad se beneficia de incentivar a los inventores para que divulguen la base de sus descubrimientos. Y, de hecho, las oficinas de patentes ponen a disposición del público los materiales de solicitud cuando conceden una patente; por lo tanto, es probable que la propia patente acelere la difusión de los conocimientos incorporados en ella (Sorenson y Fleming, 2004).

Los documentos de patentes, para cumplir el requisito legal de ofrecer una descripción completa del estado de la técnica, contienen citas de otras patentes y referencias a otras publicaciones diversas, como artículos, resúmenes, actas de congresos, libros, etc. (referencias no relacionadas con patentes) (Brusoni et al., 2005). La disponibilidad de estos datos de citas para fines de investigación motivó a investigadores de diferentes campos a utilizar los datos para identificar las diversas tendencias seguidas en la innovación y la tendencia cambiante de la tecnología, la medición de los efectos indirectos de la tecnología, las difusiones tecnológicas, la originalidad y la generalidad en las patentes. Además, las referencias a patentes se han utilizado ampliamente para aprehender las diversas relaciones entre las fuentes de conocimiento y, por supuesto, entre la tecnología y la ciencia (Sharma et al., 2017).

Para las aplicaciones anteriormente mencionadas, las patentes son una fuente útil de conocimiento por sí mismas y relacionadas con otras para generar nuevos

conocimientos de diferentes tipos. Sin embargo, aunque las patentes proporcionan una visión estructurada de los avances tecnológicos, no son un sustituto perfecto para medir todos los aspectos de la innovación. No todas las invenciones se patentan, y los comportamientos estratégicos de patentamiento pueden distorsionar su papel como indicadores de conocimiento (Arora et al., 2018). Además, las patentes por sí solas no captan plenamente el conocimiento tácito incorporado en los descubrimientos científicos, los resultados experimentales o la experiencia específica de la empresa. Por lo tanto, las patentes deben analizarse junto con otras fuentes de conocimiento, como publicaciones científicas, ensayos clínicos y esfuerzos de I+D a nivel de empresa, para obtener una visión más completa de los flujos de conocimiento y la dinámica de la innovación (Leten et al., 2022). El uso creciente de técnicas avanzadas de Procesamiento del Lenguaje Natural (PLN), como los modelos de similitud patente-documento, mejora aún más nuestra capacidad para vincular las patentes con ecosistemas de conocimiento más amplios, ofreciendo nuevas perspectivas sobre la evolución tecnológica y la recombinación de conocimientos.

En la siguiente sección, exploraré enfoques alternativos que aprovechan las patentes, las publicaciones y otras fuentes de datos estructurados para profundizar en el proceso de innovación y su potencial de comercialización en la industria farmacéutica.

3.4. Investigación previa sobre conocimiento e innovación en la industria farmacéutica: Enfoques y conclusiones

La industria farmacéutica es un sector de rápido crecimiento y rápida evolución cuya contribución al crecimiento económico y visibilidad en el ámbito de las políticas públicas ha aumentado considerablemente en las tres últimas décadas. Dos fenómenos empíricos relacionados con su reciente desarrollo revisten especial interés. En primer lugar, gran parte del crecimiento del sector se debe a procesos endógenos de innovación. Si bien es cierto que los cambios en los marcos normativos y la globalización de los mercados de usuarios desempeñan un papel, la fuerza motriz de la industria es la aparición de nuevos campos de conocimiento científico y tecnológico (Brusoni et al., 2005).

Las empresas luchan por seguir el ritmo de los nuevos avances en una serie de cuerpos especializados de conocimientos científicos y tecnológicos, que les brindan grandes oportunidades innovadoras al tiempo que desafían constantemente sus formas establecidas de hacer las cosas. Los conocimientos cada vez más especializados y la capacidad de generar continuamente nuevos conocimientos científicos se han convertido en la clave de la ventaja competitiva de las principales empresas farmacéuticas del mundo. Para ello, la industria farmacéutica se caracteriza por la coexistencia de cuerpos heterogéneos de conocimientos científicos y tecnológicos que

es necesario desarrollar, coordinar e integrar. Por este motivo, es importante considerar la cuestión de la heterogeneidad de su base de conocimientos y la amplitud cambiante de las capacidades dentro de las empresas farmacéuticas (Xie et al., 2025; Brusoni et al., 2005).

El papel del conocimiento en este campo se ha estudiado de diferentes maneras, pero con un punto de partida común: las patentes. Las patentes desempeñan un papel crucial en el proceso de descubrimiento de fármacos, ya que proporcionan protección jurídica a los descubrimientos e incentivan las inversiones en investigación y desarrollo. Mediante la identificación de patrones en los recursos de datos de patentes, los investigadores pueden conocer mejor las tendencias y prioridades del mercado de las industrias farmacéutica y biotecnológica, además de ofrecer perspectivas adicionales sobre aspectos más fundamentales como la aparición de posibles nuevas dianas farmacológicas. Centrándose en las patentes, Gadiya et al. (2023), como se muestra en la Figura 8, realizaron un estudio del panorama de patentes farmacéuticas utilizando la Herramienta de Enriquecimiento de Patentes (PEMT) para analizar las tendencias de innovación en enfermedades raras y en la enfermedad de Alzheimer. PEMT, en el marco de la minería de datos, permitió la extracción, integración y categorización de patentes basadas en áreas terapéuticas, estructuras químicas, indicaciones de destino y mecanismos de acción. El conjunto de datos incluía solicitudes de patentes de las principales jurisdicciones, como la USPTO, la EPO y la OMPI, recuperadas mediante búsquedas basadas en palabras clave y códigos de la Clasificación Internacional de Patentes (IPC). Los datos se sometieron a procesos de limpieza y normalización para eliminar duplicados y armonizar la terminología, garantizando la coherencia de los compuestos químicos y los nombres de los fármacos. El estudio aplicó técnicas de procesamiento del lenguaje natural (PLN) para clasificar las patentes por categorías terapéuticas y utilizó el análisis de redes de citas para rastrear la difusión del conocimiento, revelando cómo las patentes se basan en innovaciones anteriores. Las principales conclusiones del estudio sugieren un notable aumento de las solicitudes de patentes relacionadas con enfermedades raras y el Alzheimer, lo que subraya el compromiso cada vez mayor de la industria farmacéutica para atender necesidades médicas no cubiertas. En términos más generales, el análisis identificó a los principales titulares de patentes en el tema, una presencia cada vez mayor de universidades y empresas de biotecnología en investigación en fase inicial que contribuye a avances científicos fundamentales que más tarde se transforman en aplicaciones comerciales. Además, a través de las citas científicas en las patentes, como ya se ha visto antes en la literatura, la fuerte vinculación entre el descubrimiento científico y el desarrollo tecnológico. Por último, y como apoyo interesante a la literatura sobre la recombinación de conocimientos, una tendencia al alza en la agrupación de patentes, donde las patentes relacionadas están interconectadas, facilita las estrategias de reutilización de medicamentos y apoya la utilización eficiente de los conocimientos existentes para nuevas aplicaciones terapéuticas.

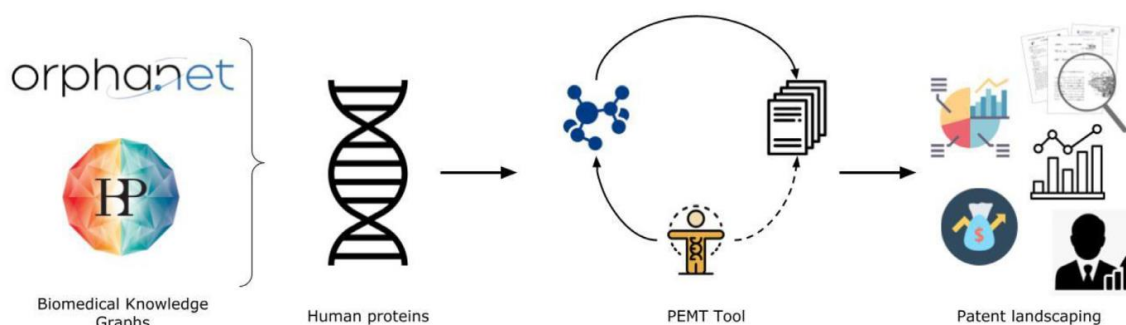


Figura 8: Sinopsis del análisis del panorama de las patentes. (Fuente: Gadiya et al. (2023))

Siguiendo en el análisis del campo de las patentes, otros estudios rastrearon el efecto de esta fuente de conocimiento en la comercialización de determinados productos. Como el ya citado trabajo de Wagner y Wakeman (2016), que analizó el poder predictivo de los indicadores basados en patentes sobre el éxito de la comercialización de productos farmacéuticos mediante la construcción de un amplio conjunto de datos que vincula 5.923 proyectos de desarrollo farmacéutico a sus patentes asociadas registradas en las bases de datos de la USPTO y la EPO. El estudio utiliza el modelo de riesgos proporcionales de Cox como metodología para analizar la probabilidad de que un medicamento farmacéutico pase por distintas fases de desarrollo en función de las características de sus patentes relacionadas. Para llevar a cabo este análisis, los autores combinan datos de desarrollo de fármacos procedentes de bases de datos comerciales de medicamentos, al tiempo que siguen de cerca la trayectoria de cada producto desde los ensayos preclínicos hasta su aprobación en el mercado. Los datos sobre patentes se complementan con un conjunto de medidas del valor y la incertidumbre de las patentes, como el tamaño de la familia de patentes (un indicador de la protección internacional y la importancia económica), las citas prospectivas (un indicador del impacto tecnológico y la difusión del conocimiento), el retraso en la concesión (una medida de la incertidumbre en los plazos de aprobación de las patentes) y el número de reivindicaciones (que evalúa la amplitud de la protección de las patentes). El conjunto de datos se somete a una exhaustiva limpieza y preprocesamiento de datos, incorporando técnicas de minería de textos y métodos de desambiguación para vincular correctamente las patentes con las empresas y los productos farmacéuticos. Además, se utilizan algoritmos de correspondencia para validar las asociaciones entre patentes y proyectos farmacéuticos, teniendo en cuenta factores como los nombres de los inventores, los cesionarios y las fechas de presentación. Los resultados sugieren que las patentes con familias de mayor tamaño y mayor número de citas tienen más probabilidades de estar asociadas a fármacos comercializados con éxito, basando estas conclusiones en la fase de ensayo clínico en la que se encuentran los medicamentos, como puede observarse en la Tabla 1, lo que refuerza la idea de que una fuerte protección de las patentes y el impacto tecnológico aumentan el potencial de comercialización. Por el contrario, un mayor retraso en la

concesión de patentes y unas reivindicaciones más amplias introducen incertidumbre, lo que puede obstaculizar la progresión de un fármaco hacia el mercado.

Teniendo en cuenta otras fuentes de conocimiento para medir su efecto en la progresión de la innovación hacia el mercado, Magazzini et al. (2012) realizaron un estudio sobre el tema del aprendizaje a partir de los fracasos en el contexto de la investigación y el desarrollo de fármacos. La investigación utilizó un conjunto de datos de más de 22.000 proyectos de desarrollo de fármacos y 200.000 patentes de fármacos para analizar la difusión de conocimientos en la industria farmacéutica. Los investigadores llevaron a cabo principalmente el análisis de citas prospectivas con el objetivo de determinar la transferencia de conocimientos en la industria farmacéutica. Las patentes se clasificaron según su vinculación con proyectos de desarrollo de fármacos exitosos, fallidos o no concluyentes, en función de los resultados de los ensayos clínicos del respectivo medicamento. Esto implicaba seguir el progreso del fármaco a través de varias fases de ensayo y su posterior éxito o fracaso a la hora de alcanzar la aprobación del mercado. Para aclarar la relación entre las citas de patentes y los resultados de los proyectos de desarrollo de fármacos, los autores emplearon un análisis de regresión. En concreto, utilizaron modelos de regresión binomial negativa para examinar cómo influía en el número de citas el éxito o el fracaso de la fase de ensayo clínico. Los resultados de la regresión revelaron que las patentes relacionadas con proyectos de fármacos exitosos tenían un número significativamente mayor de citas, lo que sugería que estas innovaciones eran influyentes en la investigación futura. Sin embargo, las patentes resultantes de ensayos clínicos fallidos fueron muy citadas, lo que indica que, a pesar del fracaso, se aprendieron conocimientos relevantes que complementaron los esfuerzos posteriores en la investigación del desarrollo de fármacos. Esto apoya el argumento de que los ensayos clínicos, independientemente de su éxito, proporcionan valiosas oportunidades de aprendizaje que apoyan la innovación posterior. Curiosamente, la información procedente de los ensayos clínicos permitió a los investigadores evaluar hasta qué punto el fracaso en la fase de ensayo clínico repercutió en los conocimientos que, a su vez, afectaron a las fases posteriores de desarrollo de fármacos. Además, la investigación sugirió que las empresas biotecnológicas, a pesar de las elevadas tasas de fracaso en los ensayos clínicos, contribuyeron enormemente a la base de conocimientos a través de los esfuerzos en materia de patentes. Los resultados subrayan la importancia tanto de los ensayos clínicos exitosos como de los fallidos en la configuración de la trayectoria de innovación de la industria farmacéutica, destacando que aprender del fracaso es un componente crucial del avance científico y tecnológico (Magazzini et al., 2012). Este trabajo aportó datos significativos sobre la importancia de los ensayos clínicos, los conocimientos que conllevan y cómo podrían influir decisivamente en la tasa de innovación y comercialización de productos en la industria farmacéutica.

Conociendo el vínculo entre la ciencia y la tecnología en la innovación establecido en la literatura anterior, Leten et al. 2021 investigaron el papel de la investigación básica

en la mejora del rendimiento innovador dentro de la industria farmacéutica, centrándose tanto en la investigación interna como en la externa. Analizaron un conjunto de datos de panel de 21 años (1995-2015) de 50 grandes empresas farmacéuticas, incorporando variables como el gasto en I+D, la diversificación tecnológica y el enfoque biotecnológico. Los autores emplearon modelos de regresión de efectos fijos para tener en cuenta las variaciones específicas de cada empresa y utilizaron el modelo de ecuaciones estructurales (SEM) para evaluar cómo influía la investigación básica interna en los resultados de la innovación. El conjunto de datos incluía solicitudes de patentes, publicaciones científicas y citas de patentes como indicadores de la recombinación de conocimientos. Al examinar los patrones de citas, el estudio demostró que las empresas con una fuerte capacidad de absorción (la capacidad de integrar conocimientos científicos externos) tenían más probabilidades de tener éxito en la innovación, acordando con el estudio realizado por Xie et al. (2025). Los resultados revelaron que las empresas con importantes inversiones internas en investigación básica superaban a las demás, especialmente cuando se combinaban con conocimientos externos, lo que pone de relieve el papel fundamental de las colaboraciones y la investigación externa para impulsar la innovación. A pesar de lo último, el estudio destacó que, con el tiempo, las empresas farmacéuticas dependían cada vez más de la investigación básica externa, lo que refleja un cambio hacia enfoques más estratégicos y rentables de I+D, como afirmaron anteriormente Arora et al. (2018) y como se puede ver en la Figura 9.

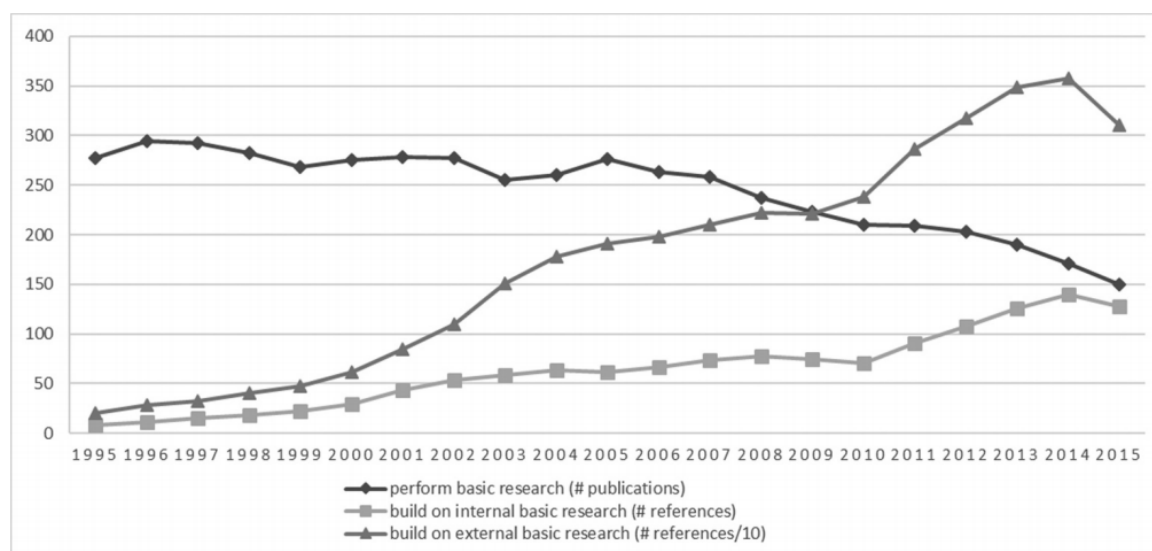


Figura 9: Tendencia temporal de la investigación básica interna y del aprovechamiento de la investigación básica interna y externa (Fuente: Leten et al. 2021)

Nota: Las cifras indicadas son medias anuales de una submuestra equilibrada de 28 empresas. Todas las variables se miden a lo largo de 4 años. En aras de claridad expositiva, la variable basada en la investigación básica externa se ha multiplicado por 10.

Aunque la mayoría de los enfoques de la bibliografía de este campo seguían métodos similares en cuanto al análisis panorámico de las patentes y la vinculación de la ciencia y la tecnología. Otras contribuciones recientes tuvieron en cuenta textos procedentes de patentes, publicaciones y ensayos clínicos. El estudio sobre el polen de abeja realizado por Algethami et al. (2022) estableció un vínculo entre los ensayos clínicos y las solicitudes de patentes, ilustrando cómo el conocimiento científico apoya e impulsa la innovación en salud y bienestar. Los ensayos clínicos considerados en el estudio aportan pruebas sólidas de la eficacia del polen de abeja en el tratamiento de afecciones como la prostatitis, el cáncer y la diabetes, destacando sus beneficios medicinales. Estos ensayos se convierten en la base para el desarrollo de nuevos productos y terapias. Al mismo tiempo, las solicitudes de patentes demuestran la aplicación práctica de estos conocimientos, donde las innovaciones derivadas de los hallazgos clínicos se convierten en productos comerciales. Por ejemplo, se han presentado muchas patentes de fórmulas a base de polen de abeja destinadas al tratamiento de enfermedades de la próstata, afecciones crónicas e incluso para la estética en el cuidado de la piel. El estudio subraya que los ensayos clínicos aportan las pruebas necesarias para respaldar las afirmaciones hechas en las patentes, conectando así la validación clínica del polen de abeja con su comercialización a través de patentes.

Una de las contribuciones más recientes que utiliza métodos vanguardistas para la vinculación de documentos es el grafo de conocimiento 2.0 de PubMed. En este trabajo, Xu et al. (2025) utilizaron técnicas de Procesamiento del Lenguaje Natural (PLN) para construir un grafo de conocimiento integrado que vinculara documentos académicos, patentes y ensayos clínicos en el campo biomédico. Se empleó PLN para extraer entidades biomédicas de artículos académicos e informes de ensayos clínicos, lo que permitió un análisis más granular de las relaciones entre los descubrimientos científicos, sus solicitudes de patentes y sus ensayos clínicos. El estudio se basó en el reconocimiento de entidades, la desambiguación de autores y el análisis de citas para garantizar que la información se relacionara con precisión entre las distintas fuentes (véase la Figura 10). Al vincular estos tres tipos de conocimiento, los autores pudieron rastrear cómo los conocimientos científicos de los artículos de investigación se aplicaban en las patentes y se probaban en los ensayos clínicos. Los resultados demostraron que vincular los ensayos clínicos con las patentes y las publicaciones científicas facilita el flujo de conocimientos y, por tanto, el proceso de innovación en la investigación biomédica. En concreto, descubrieron que la información divulgada como resultado de los ensayos clínicos podía impulsar el proceso de solicitud de patentes al validar el potencial terapéutico de las innovaciones, y las patentes, a su vez, podían orientar los futuros esfuerzos de investigación y las investigaciones clínicas.

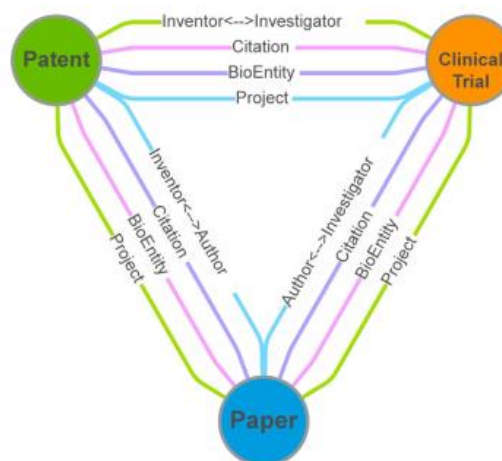


Figura 10: Cuatro formas en que los tres tipos de literatura académica se relacionan entre sí (Fuente: Xu et al. 2025)

Sin embargo, este enfoque presenta algunas limitaciones y puntos débiles. Una dificultad notable es que los ensayos clínicos rara vez citan patentes o publicaciones académicas, lo que limita la conexión entre estas fuentes de conocimiento. Los ensayos clínicos se centran sobre todo en cuestiones prácticas y normativas y pueden pasar por alto detalles importantes de los avances tecnológicos descritos en patentes o investigaciones anteriores. Esta desconexión podría generar dificultades en el método utilizado en el estudio, provocando que algunas patentes vinculadas a ensayos en realidad no se correspondan con el método de los autores. Además, aunque la PLN es una potente herramienta para analizar y procesar grandes volúmenes de texto no estructurado, su precisión y fiabilidad dependen en gran medida de la calidad de los datos y de los modelos utilizados para el reconocimiento y la desambiguación de entidades. Unas bases de datos inadecuadas y una extracción de entidades inexacta podrían dar lugar a conexiones incorrectas, lo que podría afectar negativamente a la comprensión de cómo fluye el conocimiento entre los distintos dominios (Dai et al., 2019). Además, los métodos de citación de patentes a veces pueden ser incoherentes (Sampat, B. N. 2016), especialmente en las primeras etapas de la innovación, cuando la relación entre las patentes y los ensayos clínicos no está bien establecida. A pesar de estos desafíos y de la aplicación limitada en tareas específicas, el documento destaca el considerable potencial de integrar datos de ensayos clínicos, artículos científicos y patentes para acelerar la innovación biomédica proporcionando una visión más conectada y completa del panorama de la investigación.

Teniendo en cuenta los estudios realizados en el campo farmacéutico, es evidente el papel del conocimiento para la generación y comercialización de nuevos productos mediante el uso de diferentes patentes que protegen cierta tecnología, y la ciencia que fomenta la tasa de innovación y aporta conocimientos a las pruebas de campo en forma de ensayos clínicos. Si bien el vínculo entre los ensayos clínicos y la innovación es evidente, resulta igualmente interesante conocer el efecto y la variedad del

conocimiento tecnológico que puede haber dentro de ellos, cómo se recombina, de qué manera se relaciona y cómo contribuye a alcanzar las fases más avanzadas del fármaco.

3.5. Los ensayos clínicos y su importancia como fuente de conocimiento para la innovación

Los ensayos clínicos son estudios de campo en los que los investigadores asignan a los participantes una o varias intervenciones para comprobar el efecto directo, considerando que son los tratamientos más prometedores y que ya han sido probados en el laboratorio y en animales. Esas intervenciones pueden ser un fármaco, un procedimiento o un dispositivo médico (NHLBI, 2021). Antes de entrar en la fase de ensayo clínico, se lleva a cabo un desarrollo preclínico para evaluar tanto la seguridad como la eficacia de la intervención. Si los resultados preclínicos demuestran que la terapia es prometedora, los ensayos clínicos proceden a través de múltiples fases (Fase 0, I, II, III, IV y V) para evaluar si el tratamiento es beneficioso o perjudicial para los seres humanos. Cada fase tiene un objetivo diferente (NHLBI, 2021):

- Ensayos de fase I: Los investigadores prueban por primera vez un medicamento u otro tratamiento en un pequeño grupo de personas. El propósito es conocer la mejor dosis para un medicamento u otro tratamiento y conocer la seguridad y los efectos secundarios.
- Ensayos de fase II: Los investigadores estudian el nuevo medicamento o tratamiento en un grupo más amplio de personas para determinar su eficacia y seguir estudiando su seguridad.
- Ensayos de fase III: Los investigadores administran el nuevo medicamento o tratamiento a un grupo aún mayor de participantes para confirmar su eficacia, controlar los efectos secundarios, compararlo con tratamientos estándar o similares o con un placebo, y recopilar información que permita utilizar el nuevo medicamento o tratamiento de forma segura.
- Ensayos de fase IV: Después de que la Administración de Alimentos y Medicamentos de EE.UU. (FDA) apruebe un medicamento o tratamiento y se ponga a disposición del público, los investigadores realizan un seguimiento de su seguridad en la población general, buscando más información sobre los beneficios y el uso óptimo del medicamento o tratamiento.

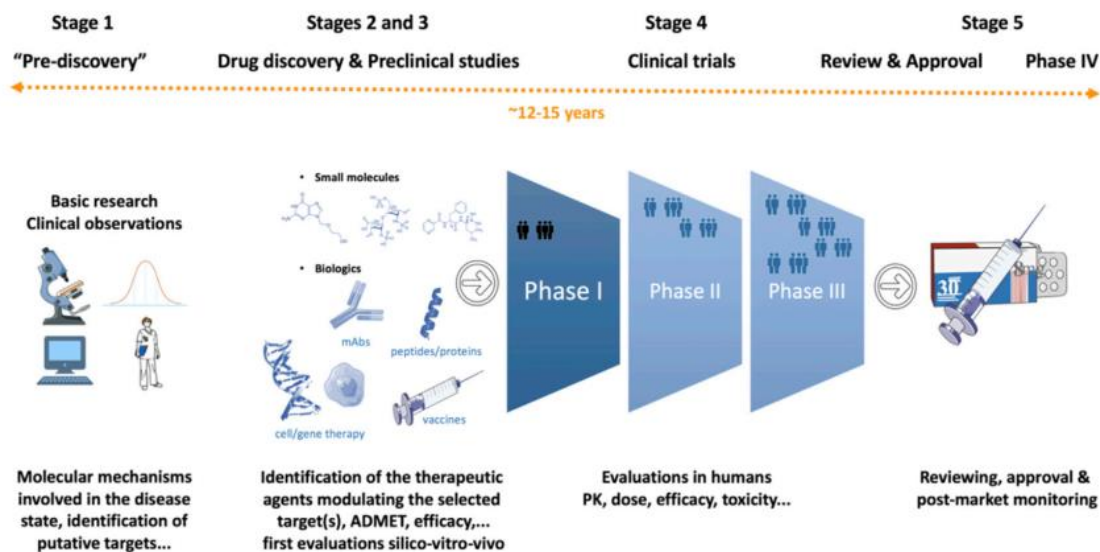


Figura 11: Descubrimiento y desarrollo de fármacos (Fuente: Singh et al., 2023)

Los ensayos clínicos constituyen un aspecto integral del proceso de innovación en la industria farmacéutica, tal y como se representa en la Figura 11, sirviendo como punto focal en la evolución de los descubrimientos científicos hacia productos comercializables. Por lo tanto, es importante reflexionar sobre el proceso de recombinação de conocimientos entre las solicitudes de patentes y los ensayos clínicos para impulsar la innovación. Los ensayos clínicos producen pruebas empíricas para tecnologías o medicamentos que generalmente se han protegido con patentes, y los resultados de estos ensayos desempeñan un papel importante en las posteriores solicitudes de patentes. Se trata de un proceso interdependiente que fomenta la disponibilidad de tecnologías en esos hallazgos, ya que los ensayos clínicos pueden motivar la creación de nuevas patentes o la modificación de las preexistentes. El proceso de recombinação de conocimientos se produce cuando la información científica derivada de los ensayos clínicos se combina con la información de patentes preexistentes y viceversa, lo que da lugar al desarrollo de productos mejorados o totalmente nuevos (Xie et al., 2025). El proceso ayuda a acelerar el ciclo de la innovación al garantizar que los productos se someten a pruebas exhaustivas mientras se acercan a resultados beneficiosos que, a su vez, reciben más protección en forma de patentes (Wagner y Wakeman, 2016). En caso de ensayos clínicos fallidos, se puede obtener información útil para mejorar las patentes existentes o motivar nuevas solicitudes de patentes mediante la identificación de áreas de mejora o nueva metodología (Magazzini et al., 2012). Por lo tanto, la comprensión de la interacción entre los resultados de los ensayos clínicos y la presentación de patentes es importante para impulsar la recombinação de conocimientos que facilite el desarrollo tecnológico, maximice el rendimiento de la investigación y acelere la comercialización.

En el caso de este trabajo de tesis, la recombinação de conocimiento y su posible efecto en la comercialización de productos farmacéuticos se mide a través del empleo de ensayos clínicos que representan la preparación para la comercialización del producto

porque son pruebas de campo legalmente controladas, y patentes que representan la tecnología y el conocimiento en sí mismo que se considera que tiene valor o potencial para ser explotado, debido a estos hechos, se tomarán como variables clave entre otras que también forman parte del modelo, el número de patentes relacionadas con un determinado ensayo clínico (considerando éste como un producto próximo al mercado) y la varianza de patentes vinculadas a ensayos clínicos considerando si son del mismo dominio o interdisciplinarios. Esta decisión se explica debido a los amplios y variados tipos de conocimiento que se utilizan en los ensayos clínicos, la mayoría de ellos mediante el uso de compuestos, dispositivos, procedimientos y software patentados que podrían suponer una importante recombinación de conocimientos.

Para estudiar el valor de estas variables de recombinación de conocimientos fue necesario encontrar, en primer lugar, la forma correcta de medir estas relaciones para saber en qué medida los ensayos clínicos utilizan conocimientos divulgados en patentes. Para este trabajo, se estudiaron diferentes métodos de relacionar el conocimiento científico con el procesamiento del lenguaje natural para obtener un grado de similitud, llegando a la conclusión de que los métodos más avanzados para esta tarea se basan en redes neuronales y en la codificación de documentos para el cálculo de la similitud, temas que se presentarán en el siguiente capítulo.

3.6. PLN, redes neuronales y codificación de documentos

El Procesamiento del Lenguaje Natural (PLN), también llamado lingüística computacional, es un subcampo de la inteligencia artificial cuyo objetivo es realizar tareas como el análisis, la comprensión y la generación del lenguaje humano natural mediante procesos informáticos. Se trata de una rama interdisciplinaria de la lingüística, la informática y la inteligencia artificial que se utiliza para manejar con destreza grandes y diversos volúmenes de datos textuales. En esta línea, es importante reconocer que el lenguaje humano tiene una complejidad y ambigüedad inherentes que lo separan de los lenguajes de programación formales. Esto se explica en parte por varios tipos de ambigüedad, entre ellos

- Ambigüedad léxica: Una misma palabra puede tener múltiples significados (por ejemplo, "banco" puede denotar una institución financiera o la orilla de un río).
- Ambigüedad sintáctica: La estructura de una frase puede permitir múltiples interpretaciones (por ejemplo, "Volar en avión puede ser peligroso" puede

implicar que el acto de volar o los propios aviones son peligrosos) (Daud et al., 2016).

- Ambigüedad semántica: El significado global de una frase puede variar en función de la interpretación (por ejemplo, "Vi al hombre con el telescopio": ¿quién posee el telescopio?) (Khan et al., 2016).
- Ambigüedad anafórica: Un pronombre o frase puede referirse de forma ambigua a un elemento anterior del texto (por ejemplo, "Juan le dijo a Marcos que ganó el premio" - ¿de quién es la victoria?) (Mankolli & Guliashki, 2020).

Para hacer frente a estos retos, la PLN emplea modelos computacionales diseñados para reconocer patrones, extraer información significativa y representar el texto en un formato estructurado.

Aprendizaje automático en PLN

El aprendizaje automático (Machine Learning, ML) es la rama de la IA que se centra en hacer que los ordenadores aprendan de los datos y, con este aprendizaje, hagan predicciones o tomen decisiones sin programación explícita. En PLN, las técnicas de ML son muy importantes debido a la complejidad del lenguaje humano, que es complicado de modelar. En lugar de confiar solo en sistemas basados en reglas, el ML aprovecha los modelos estadísticos y el aprendizaje profundo para analizar y derivar el significado de los datos textuales. Los enfoques de aprendizaje automático en PLN incluyen:

- Aprendizaje supervisado: Los modelos se entrenan en conjuntos de datos etiquetados, donde cada texto se empareja con su salida esperada. Este método prevalece en tareas como la clasificación de textos (por ejemplo, detección de spam, análisis de sentimientos) y reconocimiento de entidades con nombre.
- Aprendizaje no supervisado: Los modelos descubren patrones en datos no etiquetados. Este método se utiliza habitualmente en tareas como la agrupación de documentos y el modelado de temas.
- Aprendizaje profundo: Un subconjunto del ML que utiliza redes neuronales artificiales para aprender estructuras y patrones lingüísticos más complejos. Los modelos basados en transformadores (por ejemplo, BERT, GPT) se utilizan en aplicaciones más avanzadas como la traducción automática, el resumen de textos y la respuesta a preguntas.

Mediante el uso de estas técnicas, los modelos de PLN pueden producir representaciones precisas del lenguaje que soportan diversas aplicaciones, incluida la codificación de documentos, un proceso clave en esta tesis.

La codificación de documentos y el papel del aprendizaje por transferencia

La codificación de documentos consiste en convertir datos textuales en representaciones numéricas que puedan procesarse computacionalmente. Esta transformación es esencial para tareas como

- Calcular la similitud entre documentos
- Clasificación de distintos tipos de texto
- Facilitar la búsqueda semántica

En las siguientes secciones se abordan las redes neuronales y los hitos clave para el desarrollo de la investigación en PLN en el campo de la similitud entre documentos. El objetivo es destacar los trabajos e inventos que allanaron el camino para el aprendizaje por transferencia y los paradigmas de ajuste y modelos de vanguardia como BERT, SPECTER y PAT-SPECTER. Estos últimos métodos y principios son importantes ya que todos ellos se basan en redes neuronales, que representan el funcionamiento del motor de codificación, búsqueda y similitud que permitió el trabajo de esta tesis. Posteriormente, se presentarán casos recientes de aplicación de estos modelos relacionando patentes y patentes o patentes y artículos científicos junto con su aportación al campo. Finalmente, se presentará el modelo seleccionado para el objetivo de emparejar patentes y ensayos clínicos, exponiendo las consideraciones para la decisión.

Redes neuronales en PLN

En esta sección se ofrece una introducción a la PLN y a cómo este campo se basa en ordenadores que pueden aprender funciones a partir de los datos mediante redes neuronales. En las siguientes secciones, las innovaciones y modelos presentados son diferentes tipos de redes neuronales enfocadas a la PLN, donde el conjunto de datos está compuesto generalmente por datos textuales con diferentes características y extensiones. Estos modelos se han utilizado en el pasado para fines similares al que pretendo utilizar en esta tesis.

En el núcleo de estos modelos avanzados de PLN se encuentran las redes neuronales, que se construyen a partir de capas de unidades de procesamiento conectadas, donde cada unidad recibe la entrada de las capas anteriores. Cada unidad procesa la entrada y pasa el resultado a las capas siguientes. Uno de los tipos más comunes de redes neuronales son las redes profundas feedforward. El objetivo de la red es encontrar una función $f(x)$ que asigne la entrada x a la clase objetivo y (Goodfellow et al., 2016). Matemáticamente, el procesamiento en la unidad puede representarse como:

$$y = f\left(\sum_{i=1}^n w_i x_i + b\right) \quad (1)$$

donde y es la salida, x_i son las entradas, w_i son los pesos, b es el sesgo, y f es la función de activación. Cada capa consta de pesos w_i , que se ajustan para minimizar el error en las predicciones. Las predicciones son la salida final de la red. Los pesos se

ajustan para minimizar el error de predicción, un proceso guiado por una función de pérdida como el Error Cuadrático Medio (ECM) para tareas de regresión o la Pérdida de Entropía Cruzada para tareas de clasificación. Por ejemplo, el ECM se define como

$$L = \frac{1}{N} \sum_{i=1}^N (y_i - y'_i)^2 \quad (2)$$

donde L es la pérdida, y_i es el valor real, y'_i es la predicción de la red y N es el número de elementos para los que se calcula la pérdida. El valor real procede del conjunto de datos, mientras que el valor predicho lo genera la red. La diferencia entre el valor real y el previsto se eleva al cuadrado para que la pérdida sea siempre positiva. Las pérdidas individuales se suman y promedian para obtener la pérdida final. Una vez calculada la pérdida, los pesos se ajustan para minimizar el error de predicción mediante un proceso conocido como retropropagación (Hecht-Nielsen, 1989), que calcula el gradiente de la función de pérdida con respecto a los pesos. Los gradientes se utilizan entonces para actualizar los pesos en la dirección que reduce la pérdida mediante algoritmos de optimización. Formalmente, la actualización del gradiente puede presentarse como

$$w_i = w_i - \eta \frac{\partial L}{\partial w_i} \quad (3)$$

donde w_i son los pesos, η es la tasa de aprendizaje, que define cuánto debe ajustar el modelo los pesos, y L es la pérdida, que es el error de la red. Este proceso iterativo continúa a lo largo de múltiples pasadas por los datos de entrenamiento, conocidas como épocas, hasta que el rendimiento del modelo se estabiliza, lo que indica convergencia.

3.7. Incrustación de palabras

Aunque las redes neuronales han hecho avanzar considerablemente la PLN al permitir que los modelos aprendan patrones complejos y relaciones contextuales en el lenguaje, los enfoques anteriores de la representación de textos eran mucho más sencillos y se basaban en métodos estadísticos en lugar de en el aprendizaje profundo. Antes de la aparición de los modelos basados en redes neuronales, los sistemas de PLN utilizaban principalmente la incrustación de palabras (Word embeddings) no contextuales para codificar el texto en representaciones numéricas. Estos métodos, como Bag of Words (BoW) y Term Frequency-Inverse Document Frequency (TF-IDF), proporcionaban una forma sencilla de transformar datos textuales en un formato que pudiera procesarse matemáticamente. Sin embargo, estos enfoques tenían limitaciones críticas: no captaban el significado semántico de las palabras, ignoraban el orden de las palabras y sus relaciones, y a menudo producían representaciones dispersas de alta dimensión. En las siguientes subsecciones se analizan estas técnicas fundamentales de codificación

de textos, las siguientes, sus aplicaciones en investigaciones anteriores y su papel a la hora de sentar las bases de incrustaciones más avanzadas basadas en redes neuronales.

3.7.1. Incrustación no contextual de palabras

Las primeras aproximaciones a la incrustación de palabras se basan en la bolsa de palabras, procedimientos sencillos y rápidos que construyen un vector cuyos elementos binarios indican la presencia de una palabra en un documento ("codificación binaria de términos"), el número de apariciones de una palabra en un documento ("matriz de recuento") o el número ponderado de apariciones de una palabra en un documento ("frecuencia de términos-frecuencia inversa de documentos (TF-IDF)"). Estos enfoques tienen el inconveniente de no captar el significado o el contexto de las palabras en el documento, de ahí el uso del término "bolsa de palabras". Además, no son escalables, ya que hay que volver a entrenar todo el modelo si se añade una nueva palabra al vocabulario, y la longitud de cada vector es igual al tamaño del vocabulario. Desde el punto de vista computacional, son ineficientes, ya que generan matrices dispersas en las que la mayoría de los elementos son iguales (Erhardt et al., 2024).

Este enfoque de bolsa de palabras se ha aplicado recientemente para analizar la relación bidireccional entre ciencia (artículos de investigación) y tecnología (patentes) (Motohashi et al., 2024). En concreto, se crearon vectores de codificación de documentos utilizando los títulos y resúmenes de cada documento y se utilizó la similitud coseno para extraer 200 documentos vecinos mediante el algoritmo del vecino más cercano aproximado. El estudio descubrió que la intensidad científica de las invenciones (medida por el número de documentos de investigación vecinos por patente) ha aumentado con el tiempo, sobre todo en las patentes de universidades e institutos públicos de investigación (IIP), así como en las colaboraciones universidad-industria. Esta tendencia se vio reforzada por las reformas institucionales en Japón, como la incorporación de laboratorios gubernamentales (2001) y la incorporación de universidades nacionales (2004), que reforzaron la interacción entre la investigación científica y la innovación tecnológica. A la inversa, el estudio también reveló que la intensidad tecnológica de la ciencia (medida por el número de patentes vecinas por trabajo de investigación) ha disminuido con el tiempo. Estas conclusiones son muy relevantes para esta tesis, ya que demuestran cómo las técnicas de minería de textos como BoW pueden captar eficazmente las relaciones semánticas entre distintos tipos de documentos como los científicos y tecnológicos, en qué medida las invenciones se están convirtiendo en científicas y cómo podemos utilizar los métodos de PLN para distintos fines.

Después de la Bolsa de palabras, uno de los hitos más importantes de la PLN reciente fue el desarrollo de incrustaciones de palabras preentrenadas. Mikolov et al. (2013) abordaron este campo dando a conocer el método word2vec. El concepto clave de este método es que "una palabra se define por su entorno". Los autores utilizaron una red

neuronal simple compuesta por una capa de entrada, una capa de proyección y una capa de salida para el preentrenamiento de las incrustaciones. Las incrustaciones son vectores reales que representan palabras en un espacio vectorial continuo, capturando sus significados y relaciones semánticas. La red neuronal se entrenó con dos enfoques diferentes: bolsa continua de palabras (CBOW) y skip-gram (Mikolov et al., 2013).

En la primera arquitectura propuesta, se elimina la capa oculta no lineal y la capa de proyección se comparte para todas las palabras (no solo la matriz de proyección); de este modo, todas las palabras se proyectan en la misma posición (sus vectores se promedian). Esta arquitectura se denomina modelo de bolsa de palabras, ya que el orden de las palabras en el historial no influye en la proyección. Además, se utilizaron palabras del futuro; obtuvieron el mejor rendimiento en la tarea introducida en la siguiente sección construyendo un clasificador log-lineal con cuatro palabras del futuro y cuatro del historial en la entrada, donde el criterio de entrenamiento es clasificar correctamente la palabra actual(focal).

La segunda arquitectura es similar a CBOW, pero en lugar de predecir la palabra actual basándose en el contexto, intenta maximizar la clasificación de una palabra basándose en otra palabra de la misma frase. Más concretamente, cada palabra actual se utilizó como entrada para un clasificador log-lineal con una capa de proyección continua y predijo palabras dentro de un cierto rango antes y después de la palabra actual. Se comprobó que aumentar el rango mejora la calidad de los vectores de palabras resultantes, pero también aumenta la complejidad computacional. Dado que las palabras más distantes suelen estar menos relacionadas con la palabra actual que las cercanas a ella, se dio menos peso a las palabras distantes muestreando menos de esas palabras en los ejemplos de entrenamiento (Mikolov et al., 2013).

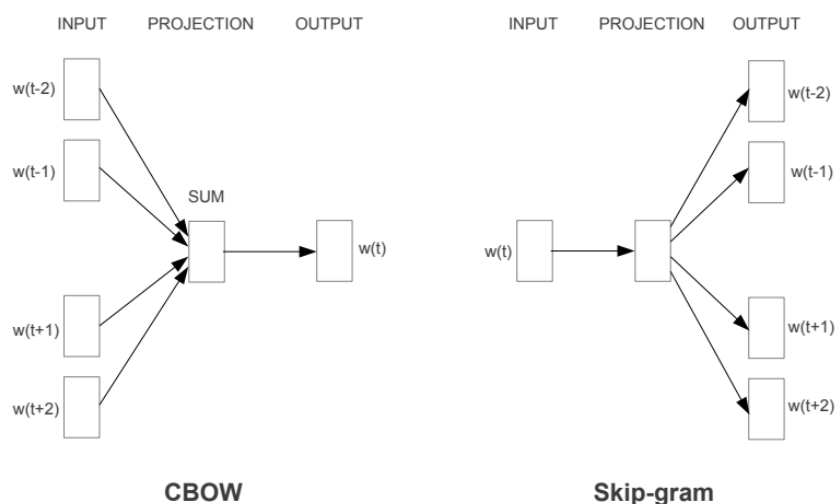


Figura 12: Nueva arquitectura del modelo. La arquitectura CBOW predice la palabra actual basándose en el contexto, y la Skip-gram predice las palabras circundantes a partir de la palabra actual (Fuente: Mikolov et al., 2013).

Ambas soluciones utilizan una arquitectura similar para maximizar la probabilidad de una palabra fuente w_t (donde t representa la posición de la palabra en la secuencia) basada en sus alrededores $w_{t-1}w_{t-2}w_{t+1}$, y w_{t+2} , o los alrededores basados en una palabra fuente w_t . Los alrededores se seleccionan a partir del contexto izquierdo y derecho de la palabra fuente con una ventana de contexto determinada. Aunque el objetivo del modelo es predecir la distribución de probabilidades de las palabras, las incrustaciones reales v_w se aprenden en la capa de proyección, donde v_w representa el vector de incrustación de la palabra w .

Este enfoque ha sido utilizado anteriormente por Hain et al. (2022) para medir la similitud tecnológica patente-patente. Utilizando esta metodología, los autores pudieron calcular las similitudes entre todas las patentes existentes, lo que a su vez les permitió representar todo el universo de patentes como una red tecnológica. Además, utilizando el caso de los vehículos eléctricos demostraron su utilidad para medir los flujos de conocimiento, cartografiar el cambio tecnológico y crear indicadores de calidad de las patentes. A pesar de estos interesantes hallazgos, es bien sabido que este enfoque depende en gran medida del conjunto de entrenamiento, ya que la mayoría de los casos pueden ser limitados y el modelo puede ser fuertemente superado por métodos preentrenados y contextuales (Hain et al., 2022; Noh & Kavuluru, 2021). La mayor limitación de word2vec, al igual que el método de la bolsa de palabras, en comparación con otros modelos de vanguardia, es que las incrustaciones de palabras no son contextuales, lo que significa que cada palabra sólo tiene un vector de representación estático. Sin embargo, el significado de una palabra depende del contexto en el que aparece. Este problema se abordó posteriormente en PLN con modelos de incrustación contextual de palabras que se presentarán en la siguiente subsección.

3.7.2. Incrustación contextual de palabras

Las limitaciones de las incrustaciones de palabras no contextuales, como Bag-of-Words (BoW) y Word2Vec, llevaron al desarrollo de modelos de lenguaje contextualizados más avanzados. En 2018, se introdujeron ELMo (Peters et al., 2018), ULMFit (Howard & Ruder, 2018) y BERT (Devlin et al., 2019), marcando un cambio significativo en cómo se generan las representaciones de palabras. A diferencia de las incrustaciones tradicionales que asignan un vector estático a cada palabra independientemente del contexto, ELMo introdujo una representación sensible al contexto, permitiendo que la misma palabra tenga diferentes significados en función del texto circundante.

La principal innovación de ELMo fue su codificación bidireccional mediante redes de memoria larga a corto plazo (LSTM), redes neuronales secuenciales que permiten que la información persista y sea capaz de procesar el texto tanto hacia delante como hacia atrás. Este enfoque mejora la desambiguación de palabras, como demuestran ejemplos como:

- *"Un ratón está cerca del gato"* (donde "ratón" se refiere a un animal)
- *"Un ratón está cerca del portátil"* (donde "ratón" se refiere a un accesorio informático).

En los ejemplos anteriores, al considerar el contexto adecuado de la palabra "ratón", el significado semántico de la palabra ("ratón") cambia de un animal a una utilidad informática. La calidad de la incrustación de la palabra fuente aumenta a medida que el contexto de la palabra se añade a la propia incrustación. Al capturar el significado dependiente del contexto de las palabras, ELMo mejoró significativamente las técnicas de codificación de documentos, que son fundamentales para tareas de similitud semántica, clasificación de textos y recuperación de información, todas ellas relevantes para el objetivo de esta tesis. Además, ELMo introdujo el concepto de aprendizaje por transferencia basado en características, que permite a los modelos inicializar sus capas más bajas con incrustaciones preentrenadas, lo que reduce los costes de entrenamiento al tiempo que mejora el rendimiento. Desde la perspectiva del ajuste y los problemas específicos de la tarea, una limitación clave de Elmo era que sólo la primera capa podía inicializarse con incrustaciones preentrenadas, mientras que el resto de los parámetros del modelo tenían que entrenarse por completo. Esto se debe a que Elmo utilizaba una arquitectura específica para cada tarea en la que las incrustaciones preentrenadas servían como entradas de características fijas. Por lo tanto, las capas y pesos preentrenados aprendidos en una tarea no podían transferirse completamente a otra.

Sobre esta base, ULMFit (Howard y Ruder, 2018) introdujo un paradigma de formación más generalizado, diseñado para aprovechar al máximo los modelos lingüísticos preformados en diferentes tareas de PLN. En lugar de solo inicializar la primera capa con incrustaciones preentrenadas (como en ELMo), ULMFit propuso un enfoque de tres pasos:

1. Preentrenamiento de un modelo lingüístico en datos de texto a gran escala,
2. Ajuste del modelo para una tarea específica de PLN,
3. Ajuste supervisado para tareas de clasificación.

Este método resultó muy eficaz y demostró que los modelos lingüísticos preentrenados necesitan muchos menos ejemplos etiquetados para lograr un buen rendimiento. Con sólo 100 ejemplos etiquetados, ULMFit igualó la precisión de un modelo completamente entrenado que requería 10 veces más datos.

3.7.3. Transformers y BERT

BERT, lanzado a finales de 2018, introdujo dos avances importantes en el modelado del lenguaje: (1) la codificación bidireccional de tokens de entrada y (2) el objetivo de modelado de lenguaje enmascarado (MLM). Además, su novedosa arquitectura mejoró significativamente la escalabilidad del procesamiento de texto." Mientras que ULMFit y ELMo se basaban en LSTM, BERT utilizaba la arquitectura Transformer introducida por Vaswani et. al. (2017). La arquitectura Transformer se muestra en la Figura 13. La arquitectura general consiste en una estructura de codificador-decodificador, en la que el codificador procesa la secuencia de entrada y el decodificador genera la secuencia de salida a partir de la salida del codificador. El codificador se compone de capas apiladas de redes neuronales multicabezales de autoatención y feed-forward. Dado que el modelo utilizado en esta tesis sólo utiliza el codificador, esta sección se centrará exclusivamente en el componente codificador. El codificador tiene seis capas apiladas similares vinculadas entre sí. Cada capa está formada por unidades de autoatención multicabezal y una red feed-forward totalmente conectada. El mecanismo de autoatención permite al modelo atender conjuntamente a todos los tokens de la secuencia al procesar cada token. Un token puede ser una palabra o una subpalabra, dependiendo de la técnica de conversión de texto a valores numéricos. En la práctica, la autoatención determina qué partes de la secuencia son relevantes para cada token (Jokinen, 2024).

La implementación original de Transformers también incluía la codificación posicional a las secuencias de entrada. La incrustación posicional, que representa el orden de los tokens, se añadió a la incrustación de entrada directamente. Así, el modelo aprende la ubicación y la distancia entre las fichas directamente de la incrustación, y el marco no necesita almacenar fichas anteriores en la memoria para "recordar" el contexto como en las redes neuronales recurrentes. Este enfoque mejora la gestión de las dependencias de largo alcance en las secuencias de entrada.

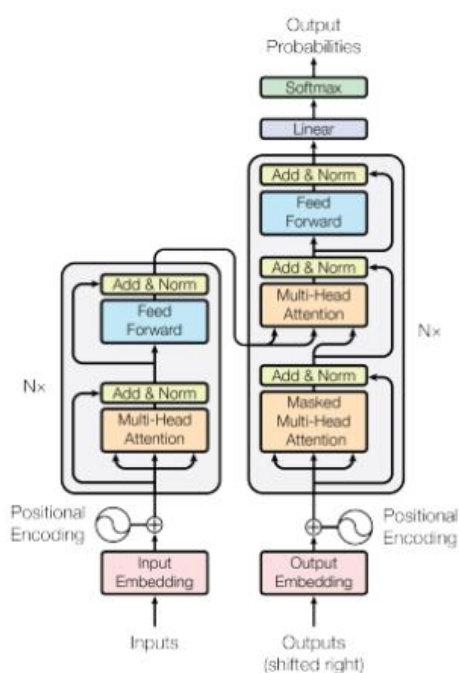


Figura 13: La arquitectura transformer (Fuente: Vaswani et al., 2017)

BERT utiliza solo la parte codificadora de la arquitectura Transformer, que se diseñó originalmente para la traducción automática utilizando una estructura codificador-decodificador. Más allá de resolver problemas específicos de modelado lingüístico, la arquitectura Transformer también abordó múltiples desafíos relacionados con la escalabilidad de los modelos lingüísticos, en particular en comparación con los modelos basados en la memoria a corto plazo larga (LSTM). Esta mayor escalabilidad se consiguió principalmente eliminando las dependencias secuenciales. En lugar de depender de unidades recurrentes, BERT emplea capas de autoatención, lo que permite al modelo procesar secuencias enteras en paralelo (Jokinen, 2024). Además, el orden de los tokens en una secuencia se incorpora al propio vector de entrada a través de la codificación de posición (Vaswani et al., 2017), lo que mejora la capacidad del modelo para manejar secuencias de manera eficiente.

Una de las innovaciones clave de BERT es la codificación bidireccional, que aborda las limitaciones de los modelos lingüísticos anteriores, como ELMo y GPT de OpenAI (Howard y Ruder, 2018). Estos modelos anteriores utilizaban la codificación unidireccional, lo que significa que procesaban el texto de izquierda a derecha o de derecha a izquierda, limitando su capacidad para comprender plenamente el contexto. BERT supera este problema al considerar simultáneamente los contextos izquierdo y derecho, lo que da lugar a incrustaciones contextuales más ricas.

Dado que BERT se basa en la arquitectura Transformer, sus dimensiones de entrada y salida se adhieren a las restricciones de Transformer. En concreto, la longitud máxima de entrada que admite es de 512 tokens (Vaswani et al., 2017). BERT implementa la

autoatención como el trabajo original de Transformer, pero con 12 cabezas de atención paralelas. Dado que $h = 12$, la dimensión de incrustación del modelo (d_{model}) se calcula como:

d_{model} : La dimensión de incrustación total del modelo, este es el tamaño de las incrustaciones de entrada y salida utilizadas en todo el modelo Transformer.

h : Número de Cabezas de Atención, la autoatención multicabeza permite al modelo Transformer centrarse en diferentes aspectos de la secuencia de entrada simultáneamente. Cada cabeza de atención procesa un subespacio más pequeño de la dimensión total de incrustación.

d_k : Dimensión de cada cabeza de atención, cada cabeza de atención opera en un subespacio de menor dimensión en lugar del espacio de incrustación completo. Esto reduce el coste computacional al tiempo que mantiene la expresividad.

$$d_{model} = h * d_k = 12 * 64 = 768 \quad (4)$$

BERT se entrenó utilizando el Modelado de Lenguaje Enmascarado (MLM), un método de aprendizaje no supervisado en el que se enmascaran palabras aleatorias de una frase y el modelo aprende a predecirlas basándose en el contexto circundante. Por ejemplo, dada la frase "El rápido zorro marrón salta sobre el perro perezoso". Si las palabras "rápido" y "perezoso" están enmascaradas, la entrada sería: "El [ENMASCARADO] zorro marrón salta sobre el perro [ENMASCARADO]". (Noh & Kavuluru, 2021). El objetivo de BERT es predecir correctamente las palabras enmascaradas como "rápido" y "perezoso", aprovechando el contexto de las palabras restantes de la frase. Este método de entrenamiento permite a BERT captar relaciones semánticas profundas entre las palabras, mejorando su capacidad para comprender el significado y la estructura del lenguaje.

La naturaleza no supervisada de MLM permite preentrenar BERT en corpus de texto a gran escala de diversos dominios antes de afinarlo para tareas posteriores específicas, como el análisis de sentimientos, la respuesta a preguntas o el reconocimiento de entidades con nombre. Este paradigma de preentrenamiento-ajuste ha hecho de BERT uno de los modelos más influyentes en el procesamiento moderno del lenguaje natural (Erhardt et al., 2024).

BERT se convirtió en el primer modelo de representación basado en el ajuste fino y que superó a las arquitecturas que se diseñaron por separado para problemas específicos de la tarea (Devlin et al., 2019). BERT utilizó la misma arquitectura para todas las tareas y se pudo ajustar con precisión añadiendo una capa de salida adicional. En comparación con ULMFit, el preentrenamiento y el ajuste fino fueron más sencillos, ya que BERT requirió menos cambios en la arquitectura en diferentes fases. Dado que BERT era de código abierto, los modelos preentrenados podían utilizarse eficazmente

a través del ajuste fino en diferentes organizaciones, incluso cuando solo tienen pequeños volúmenes de datos específicos del dominio.

Más adelante, en 2019, se lanzó Sentence BERT. La principal diferencia con el BERT original es que Sentence BERT se entrena para la comprensión a nivel de frase y utiliza una red siamesa y funciones de pérdida de triplete donde las frases semánticamente similares se aprenden para tener representaciones vectoriales cercanas (Reimers & Gurevych, 2019); por otro lado, BERT se centra en el nivel de token. La configuración del entrenamiento para aprender representaciones a nivel de oración está estrechamente relacionada con la forma en que SPECTER y PAT-SPECTER se entrenan con un objetivo a nivel de documento para generar vectores de documentos.

SciBERT, introducido por Beltagy et.al. en 2019, es una variante de BERT diseñada específicamente para texto científico. Según los autores, los modelos neuronales profundos se entrenados con datos anotados en ese momento. Dado que el dominio científico carecía de datos etiquetados y la anotación es costosa, los autores abordaron esta cuestión entrenando el modelo con los mismos objetivos con los que se entrenó BERT, pero con textos científicos, con un gran corpus de 1,14 millones de publicaciones científicas de informática y del amplio campo biomédico. El objetivo de los autores era entrenar un modelo capaz de producir representaciones de alta calidad a partir de documentos científicos, hecho que les permitió trabajar con diferentes corpus de textos científicos. SciBERT es esencial para SPECTER y PAT-SPECTER, ya que ambos utilizan SciBERT como modelo inicial preentrenado.

Entre las aplicaciones en el campo de la PLN de los modelos basados en BERT encontramos un proyecto premiado financiado por la Oficina Europea de Patentes (EPO) y llevado a cabo por Wang et al. (2023). El proyecto, que utilizaba diferentes modelos BERT, tenía como objetivo "desarrollar un método de aprendizaje automático de alto rendimiento para extraer referencias en el texto de las patentes y luego cotejarlas con la base de datos Web of Science de publicaciones científicas, aplicar este método a las patentes de la EPO y la USPTO para crear un conjunto de datos a gran escala que vincule las patentes con las publicaciones, y descubrir qué tipos de ciencia conducen a patentes más valiosas y en qué se diferencian las referencias en el texto de las referencias en la portada". Este proyecto se basó finalmente en el uso de SciBert, debido a que demostró un mejor rendimiento a la hora de extraer palabras, términos y citas específicas de distintos tipos de corpus de texto científico gracias a su entrenamiento con artículos científicos. A pesar del éxito de este enfoque, el descubrimiento de la relación entre el valor de la patente y el carácter básico de la ciencia, la interdisciplinariedad, la novedad y las citas científicas, resultó inviable aplicarlo a la vinculación ensayo clínico-patente debido a la estructura de los ensayos clínicos, que rara vez presentan referencias en el texto a patentes o documentos científicos.

3.7.4. SPECTER

SPECTER es un codificador de texto de propósito general que puede producir representaciones vectoriales de longitud fija a nivel de documento a partir de artículos científicos. Utiliza como base el anteriormente mencionado SciBERT, pero afinado con un objetivo a nivel de documento. La arquitectura del transformador de SPECTER se hereda de BERT, y también las limitaciones de todos los modelos basados en BERT, como los vectores de longitud fija de 768 dimensiones y el límite de procesamiento de 512 tokens para cada documento. Lo anterior también significa que las incrustaciones del documento fueron producidas por el propio modelo de lenguaje del transformador sin ajuste fino, y las incrustaciones se utilizaron como características para el modelo descendente (Jokinen, 2024; Cohan et al., 2020).

El enfoque particular de SPECTER permitió superar a otros modelos en varias pruebas comparativas, como la clasificación de temas, la predicción de citas y la predicción de la actividad de los usuarios. La necesidad de desarrollar un codificador de texto con las características de SPECTER radica en que los modelos basados en BERT tienen una capacidad limitada para representar documentos como un todo, la razón radica en el entrenamiento de BERT, que se realizó con objetivos de modelado de lenguaje enmascarado y predicción de frases siguientes, que se centra en construir la comprensión sólo a partir de las relaciones a nivel de token y frase de los datos del documento. Sin embargo, la comprensión de los documentos es importante en tareas como la clasificación o la búsqueda, y SPECTER resolvió esta limitación con un objetivo de ajuste fino a nivel de documento. Para estos objetivos, SPECTER se utilizó originalmente como el modelo codificador de documentos de "Logic Mill" para clasificar y recuperar documentos científicos similares en una enorme base de datos (Erhardt et al., 2024). Sin embargo, este modelo fue sustituido recientemente por PAT-SPECTER debido a su mejor rendimiento en la codificación de patentes y a su mayor precisión en otras tareas de PLN.

La figura 14 ilustra el proceso de entrenamiento que, en última instancia, optimiza y actualiza los parámetros del modelo desde las tareas a nivel de token hasta las tareas a nivel de documento. A grandes rasgos, el marco de formación consta de cuatro niveles (de arriba abajo): 1) alimentación de tripletes al modelo, 2) codificación de cada documento con un codificador transformador, 3) construcción de representaciones vectoriales de cada documento, y 4) cálculo de la pérdida mediante una función de activación. Cada nivel se describe en los párrafos siguientes:

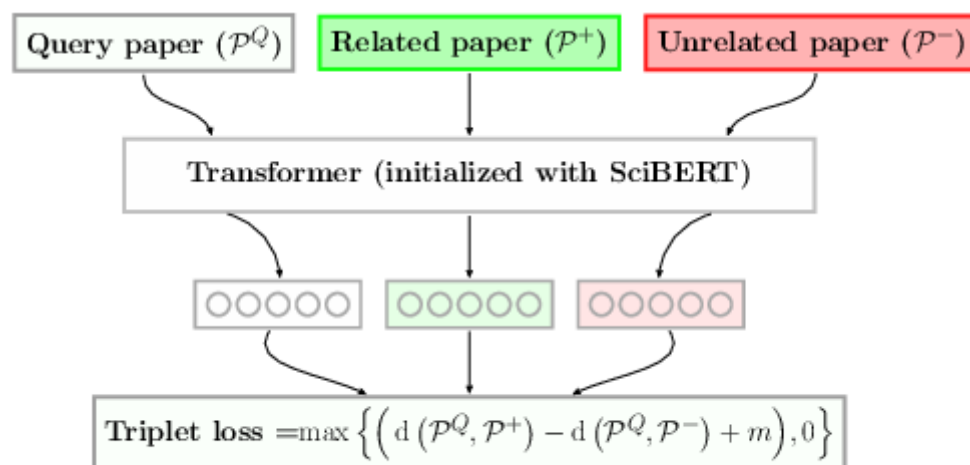


Figura 14: Arquitectura de entrenamiento de SPECTER (Fuente: Cohan et al., 2020)

Entrada del modelo

Uno de los cambios cruciales con respecto a SciBERT relativo y otros modelos basados en BERT es cómo se alimentan las entradas al modelo en el proceso de ajuste. En lugar de introducir entradas y salidas, SPECTER toma tres instancias como entrada: consulta, artículo relacionado y artículo no relacionado. Se toma una muestra de un triplete de tres artículos del gráfico del documento. Para cada instancia de entrada, se concatenan el título y el resumen del artículo.

Cada secuencia de entrada comienza con el CLS (token de clasificación), que captura toda la información de la secuencia/documento de entrada y se utiliza para tareas de clasificación y representación de documentos, y el SEP (token separador), que actúa como delimitador entre distintas secciones de la entrada, como el título y el resumen. Un cambio significativo en comparación con el BERT original es que el *token* CLS se coloca al principio de cada documento en lugar de al principio de una frase, como se presentaba en el BERT original. Alimentando el triplete de consultas, documentos positivos y negativos, la función de pérdida puede aprender la similitud semántica entre documentos. La relación se basa en las citas dentro de los documentos (Cohan et al., 2020).

Parámetros inicializados

Los parámetros del transformador en SPECTER se inicializan a partir del SciBERT preentrenado. Todos los parámetros se actualizan durante el ajuste. La versión de la arquitectura del modelo es BERT-BASE. El tamaño de la capa oculta es 768, un número de codificadores apilados $L=12$, un número de cabezas de atención $h = 12$, y parámetros totales = 110 M. La arquitectura es agnóstica al modelo de lenguaje Transformer con el que se inicializa. La arquitectura es flexible para ser utilizada en

otros dominios y lenguajes sustituyendo los datos y la entrada del modelo, siempre que la estructura de los datos sea la misma.

Incrustación e inferencia

La representación del documento es una representación del token CLS en la última capa oculta del modelo de lenguaje del transformador. La longitud de la incrustación es $d_{model} = 768$. Cada uno de los tres documentos se incrusta a través del modelo como se describe a continuación:

$$v = \text{Transformer}(\text{input})_{CLS} \quad (5)$$

donde Transformer es la función de avance del transformador, y la entrada es la concatenación del token [CLS] y WordPieces del título y el resumen de un artículo, separados por el token [SEP]. La salida del transformador para el token CLS es un vector denso:

$$v_{CLS} \in R_D, \text{ donde } D = 768 \quad (6)$$

que es la representación final del documento en el momento de la inferencia. En la fase de entrenamiento, cada documento se incrusta por separado del token CLS de la última capa oculta, que también se denomina "salida agrupada" (Cohan et al., 2020). Finalmente, los tres vectores se utilizan para calcular una pérdida y optimizar los parámetros mediante retro propagación. Durante el entrenamiento, SPECTER aprovecha la información de citas para ajustar el modelo y aprender representaciones de documentos que capturen la relación entre artículos. Sin embargo, una vez entrenado, el modelo puede generar de forma independiente codificaciones de documentos basándose únicamente en el contenido textual de ellos, sin necesidad de datos de citas ni de la estructura del grafo de documentos. Esta capacidad permite a SPECTER producir incrustaciones para documentos nuevos, incluidos aquellos que aún no han sido citados, lo que lo hace versátil y eficiente para aplicaciones en tiempo real dirigidas a artículos científicos recientes u otros tipos de documentos científicos (Jokinen, 2024).

El modelo puede utilizarse como codificador de uso general en aplicaciones posteriores. El transformador se entrena para comprimir la semántica del documento en un vector de longitud fija. Idealmente, el vector representa el contenido del documento, pero también su relación semántica con los demás documentos, lo que hace que los documentos sean comparables entre sí. De este modo, el modelo es óptimo en tareas a nivel de documento, como la clasificación temática o la recomendación de citas para nuevos artículos no vistos y documentos de patentes (Erhardt et al., 2024).

Pérdida y objetivo de entrenamiento

Como función de pérdida, SPECTER utiliza una pérdida de margen triple. Para cada instancia de entrenamiento, el modelo ve un triplete de un artículo de consulta P^Q , un artículo relacionado positivamente P^+ y un artículo relacionado negativamente P^- . El artículo positivo es un artículo citado por el artículo de consulta, y el artículo negativo es un artículo no citado por el artículo de consulta (aunque puede ser citado por P^+). La pérdida de margen del triplete se define del siguiente modo:

$$L = \max \{d(P^Q, P^+) - d(P^Q, P^-) + m, 0\} \quad (7)$$

donde L es la pérdida calculada, d es la distancia entre dos artículos y m es el margen de pérdida (fijado en 1 por los autores). La distancia d es una norma L2 definida como sigue:

$$d(P^A, P^B) = \|v_A - v_B\|^2 \quad (8)$$

donde v_A y v_B son representaciones vectoriales de longitud fija de los documentos A y B , obtenidas a partir del token de clasificación del codificador en la última capa oculta. La idea de la función de pérdida del triplete es que el modelo intenta aprender representaciones cercanas de P^Q y P^+ y viceversa de P^Q y P^- . Como función de activación, se aplica la unidad lineal rectificadora (ReLU). ReLU se define como:

$$ReLU(x) = \max(0, x) \quad (9)$$

Significa que, si la pérdida calculada es negativa, por ejemplo, si el artículo positivo ya está más cerca del artículo de consulta que el artículo negativo por al menos el margen, la pérdida se establece en 0. Si la pérdida es positiva, ReLU permite que el valor pase sin cambios, señalando que el modelo aún necesita ajustarse. Este uso de ReLU garantiza un aprendizaje eficaz actualizando el modelo sólo cuando la representación actual aún no satisface la condición de margen.

Sin margen m , cuando $d(P^Q, P^+) < d(P^Q, P^-)$ el documento positivo está más cerca del documento de consulta que el negativo. Este es el resultado deseado. En este caso, la pérdida se fija en cero y no es necesario ajustar los parámetros del modelo. La pérdida por triplicado aumenta el margen para separar varios pares, acercando los puntos de datos relacionados y asegurando al mismo tiempo una proyección distante para los no relacionados. Así, mientras $d(P^Q, P^+) - d(P^Q, P^-) > m$, el modelo devuelve una pérdida positiva y se ajustan los parámetros (Jokinen, 2024). El margen pretende definir una distancia mínima en que un paper similar y disímil deben tener en el espacio vectorial en relación con el paper de consulta (Cohan et al., 2020).

3.7.5. PAT-SPECTER

PAT-SPECTER es un modelo de lenguaje para codificar publicaciones científicas y patentes al mismo tiempo. Se trata del modelo SPECTER (versión: 2.0) pero afinado en un conjunto de datos de patentes curado que comprende 1,5 millones de instancias de entrenamiento (5 tripletas para 300.000 patentes focales) con citas de patentes creíbles. Este modelo combina ambos dominios, publicaciones científicas y patentes (Ghosh et al., 2024).

El Instituto Max Planck para la Innovación utiliza PAT-SPECTER como modelo codificador de documentos en "Logic Mill", un sistema de software modular que ayuda a navegar por los conocimientos contenidos en publicaciones científicas y documentos de patentes. Este sistema puede aprovecharse para realizar distintas tareas en el campo de la PLN, como generar representaciones numéricas para los documentos del usuario, calcular similitudes entre documentos o documentos de su base de datos y buscar documentos similares (Erhardt et al., 2024). Esta aplicación real resultó ser la columna vertebral para el desarrollo de esta tesis en la medida en que presentó una forma de utilizar un modelo codificador de lenguaje de última generación para comparar dos tipos de textos diferentes, apoyándose en el entrenamiento extensivo del modelo que considera una enorme cantidad de documentos de diferentes campos. Además, en este trabajo intenté reproducir parcialmente el método de trabajo de Logic Mill, principalmente en la forma de procesar los datos, la codificación y el cálculo de la similitud. La arquitectura de "Logic Mill" se muestra en la Figura 15. Además, en lugar de tener en cuenta las publicaciones científicas, empleé ensayos clínicos e información clave de patentes para generar los vínculos. Por lo tanto, el número de patentes y los indicadores relacionados con determinados ensayos clínicos se utilizan como variables para comprobar los conocimientos y la recombinación de conocimientos incorporados en las respectivas patentes. Todo este proceso en detalle, junto con el modelo y otras variables, se presenta más adelante en la sección Datos y métodos.

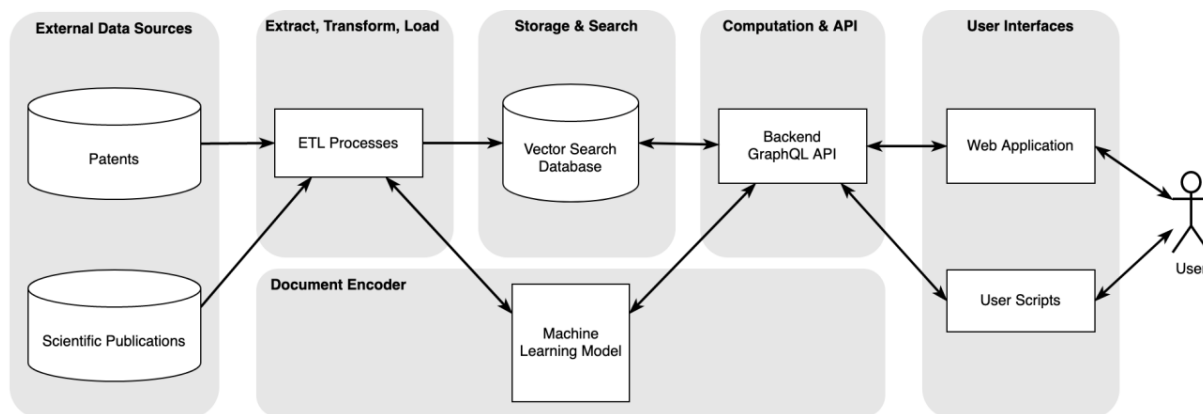


Figura 15: Visión general de la arquitectura de Logic Mill (Fuente: Erhardt et al., 2024)

3.7.6. El papel fundamental de la selección del modelo

La selección del modelo lingüístico adecuado es crucial a la hora de establecer vínculos semánticos entre patentes y ensayos clínicos. La eficacia del modelo repercute directamente en la precisión de la representación de los documentos, los cálculos de similitud y, en última instancia, la calidad de los conocimientos derivados de vincular los avances científicos y tecnológicos. En este contexto, la elección entre SPECTER, PAT-SPECTER y otros modelos basados en Transformers es especialmente relevante.

Como ya se ha dicho, los modelos tradicionales como TF-IDF y Word2vec se basan en la concordancia de palabras clave o en la incrustación estática de palabras, que carecen de comprensión contextual. Estos modelos son inadecuados para captar las relaciones matizadas entre los ensayos clínicos y las patentes, donde el contexto y el significado van más allá de la simple coocurrencia de palabras. Del mismo modo, BERT y SciBERT, aunque potentes en la comprensión de textos generales y científicos, están diseñados principalmente para tareas a nivel de frases o párrafos, más que para representaciones de documentos completos. SciBERT, en concreto, está entrenado en textos biomédicos, lo que lo hace eficaz para tareas relacionadas con datos clínicos, pero está menos optimizado para relacionar patentes, que contienen contenidos muy estructurados y con formato legal.

SPECTER está diseñado para capturar la similitud semántica entre documentos científicos aprovechando los grafos de citas. Este modelo codifica eficazmente los trabajos de investigación y la literatura científica mediante el entrenamiento en relaciones basadas en citas, como se ha explicado en las subsecciones anteriores, lo que lo hace muy eficaz para identificar trabajos académicos conceptualmente relacionados. Sin embargo, dado que SPECTER está optimizado para artículos científicos y no para patentes, su rendimiento en la captura de relaciones entre ensayos clínicos y patentes puede ser subóptimo debido a las diferencias en la estructura, terminología y patrones de citación entre estos tipos de documentos.

En cambio, PAT-SPECTER está explícitamente ajustado para codificar tanto patentes como literatura científica, lo que lo convierte en una opción más adecuada para este estudio. A diferencia de SPECTER, que se centra principalmente en artículos de investigación, PAT-SPECTER incorpora citas de patentes en su proceso de entrenamiento, lo que le permite aprender las características lingüísticas y estructurales particulares de las patentes. Esta distinción es crucial porque las patentes suelen utilizar un lenguaje técnico y jurídico muy que difiere de la escritura académica. Además, PAT-SPECTER tiene en cuenta las relaciones de citación entre las patentes y las publicaciones científicas, lo que le permite establecer conexiones semánticas más sólidas entre los ensayos clínicos y las patentes (Ganguli et al., 2024).

Otra ventaja clave de PAT-SPECTER es su enfoque de entrenamiento basado en tripletes, que mejora su capacidad para distinguir entre documentos fuertemente relacionados, poco relacionados y no relacionados. Este enfoque es especialmente

beneficioso para relacionar patentes con ensayos clínicos, donde las citas directas son raras y la similitud semántica debe inferirse a partir del contenido textual en lugar de referencias explícitas. Dado que los registros de ensayos clínicos incluyen datos estructurados como el diseño del estudio, los resultados del tratamiento y las aprobaciones normativas, la capacidad de PAT-SPECTER para procesar diversos formatos textuales refuerza aún más su idoneidad para esta tarea.

Al seleccionar PAT-SPECTER en lugar de los modelos tradicionales de PLN, los modelos generales de Transformers como BERT y los modelos básicos basados en citas como SPECTER, se garantiza un enfoque más sólido y consciente del contexto para descubrir los vínculos entre los ensayos clínicos y las patentes. Esta elección mejora la precisión de las evaluaciones de similitud y por ende permite una identificación más exacta de los documentos relacionados.

3.8. Pregunta de investigación

Partiendo de las investigaciones previas que han establecido la importancia de fuentes de conocimiento diversas y de alta calidad -como patentes, publicaciones científicas y datos clínicos- para impulsar la innovación, este estudio se centra en un vínculo menos explorado: la relación entre el conocimiento tecnológico y la progresión de los ensayos clínicos. Aunque la recombinación de conocimientos ha sido ampliamente reconocida como un mecanismo clave detrás de las innovaciones revolucionarias, su influencia directa en los resultados del desarrollo clínico sigue siendo poco estudiada. La industria farmacéutica, caracterizada por su dependencia de aportaciones heterogéneas de conocimientos y su complejo entorno normativo, ofrece un contexto convincente para investigar esta relación. El objetivo de esta investigación es evaluar cómo las características específicas del conocimiento basado en patentes -en particular las que reflejan la recombinación, la amplitud tecnológica y la intensidad de las citas- se relacionan con la probabilidad de que los ensayos clínicos avancen a etapas posteriores, lo que sirve como medida de la proximidad a la comercialización. Al vincular las patentes con los ensayos clínicos mediante un modelo de similitud semántica y analizar el conjunto de datos resultante con variables de control a nivel de empresa y de ensayo, el estudio pretende descubrir cómo la composición del conocimiento tecnológico puede moldear la trayectoria del desarrollo clínico.

4 Datos y métodos

En este capítulo se detalla el escenario utilizado para comprender el papel del conocimiento incorporado en las patentes sobre la innovación que representan los ensayos clínicos en fases avanzadas, y las fuentes de datos obtenidas relacionadas con las tecnologías de vacunas de ARNm. Posteriormente, se presenta la metodología seguida para desarrollar el modelo de similitud basado en PAT-SPECTER y una decisión de umbral respectiva basada en un conjunto de datos de referencia, seguida de los pasos metodológicos empleados para abordar adecuadamente las preguntas de la investigación.

4.1. Selección del campo del conjunto de datos

El primer paso consistió en definir el campo de conocimiento para abordar las preguntas de la investigación. En esta tesis, el análisis de los efectos del conocimiento y la recombinación de conocimientos se llevará a cabo utilizando una base de datos de ensayos clínicos y patentes, ambos sobre tecnologías de vacunas de ARNm.

La plataforma de la vacuna de ARNm es un caso reciente y claro de recombinación de conocimientos en la innovación farmacéutica. Su desarrollo se basó en una amplia integración de diferentes ámbitos científicos, sobre todo la biología molecular, la inmunología y la nanotecnología lipídica, para crear productos innovadores unificados. Los investigadores aprovecharon décadas de investigación básica sobre la biología del ARNm y los mecanismos del sistema inmunitario junto con los avances en los sistemas de administración de nanopartículas, fusionando estas corrientes de conocimiento en un novedoso enfoque de la vacuna (Casadevall, 2021). Esta convergencia de conocimientos permitió a los científicos superar obstáculos técnicos clave (como la estabilización del ARNm y su introducción segura en las células a través de nanopartículas lipídicas) y acelerar así el desarrollo de vacunas. El aprovechamiento de esta base de conocimientos recombinados permitió el rápido diseño, ensayo y despliegue mundial de vacunas eficaces de ARNm contra el COVID-19 en menos de un año (EPO, 2023), lo que ilustra cómo este tipo de innovación interdisciplinaria puede acelerar drásticamente la comercialización de productos.

4.2. Estructura de los ensayos clínicos, recuperación y limpieza del conjunto de datos

Los ensayos clínicos siguen un enfoque estructurado y sistemático para evaluar la seguridad, la eficacia y el potencial general de los nuevos tratamientos médicos. Esta estructura garantiza la coherencia entre los ensayos, lo que permite una recopilación de datos fiable y reproducible. Cada ensayo se rige por un protocolo bien definido, en el que se describen los objetivos, la metodología y los criterios de inclusión de los participantes. Los ensayos se diseñan en fases, cada una con una finalidad específica, como evaluar la seguridad, determinar las dosis óptimas o evaluar la eficacia comparativa con los tratamientos existentes. Estas fases son cruciales para proporcionar información en diferentes etapas del desarrollo del tratamiento, desde los primeros ensayos de seguridad hasta los estudios a mayor escala que evalúan la eficacia en una población más amplia (ClinicalTrials.gov, 2021).

Los ensayos clínicos se organizan para garantizar que los datos se recopilan de forma que se minimicen los sesgos y se maximice la validez científica de los resultados. La estructura de los ensayos clínicos se adapta al tratamiento que se está probando, y cada estudio se diseña cuidadosamente para abordar cuestiones de investigación específicas. Este diseño suele incluir aleatorización, grupos de control y métodos estadísticos para analizar los resultados. Además, la recopilación de datos durante los ensayos clínicos se realiza con precisión, haciendo un seguimiento de las respuestas de los pacientes mediante evaluaciones clínicas, pruebas de laboratorio y otras formas de monitorización. Estos datos desempeñan un papel fundamental a la hora de respaldar las decisiones normativas y hacer avanzar el tratamiento a fases posteriores de desarrollo o introducción en el mercado (U.S. Food and Drug Administration [FDA], 2022).

Además de la estructura del ensayo, las funciones de los patrocinadores principales y los colaboradores garantizan el éxito de la ejecución de estos ensayos. El patrocinador principal es el principal responsable de gestionar el estudio, supervisar el análisis de datos y garantizar el cumplimiento de los requisitos normativos. Las organizaciones colaboradoras a menudo proporcionan apoyo adicional, como financiación, experiencia en el diseño o análisis de datos, contribuyendo al esfuerzo de investigación más amplio. A través de esta estructura bien organizada, los ensayos clínicos generan datos valiosos que ayudan a disminuir la brecha entre la investigación científica y la comercialización de nuevos tratamientos, contribuyendo así al avance continuo de la ciencia médica.

Para construir el conjunto de datos de ensayos clínicos para el estudio, los ensayos se recuperaron de clinicaltrials.gov. A partir de su estructura, se obtuvieron los siguientes componentes:

- NCTID: En relación con un ensayo, este número representa un número de registro NCT estadounidense único obtenido por el solicitante a través del sitio web público ClinicalTrials.gov
- Título oficial: En relación con un ensayo, este elemento es el nombre oficial del protocolo proporcionado por el investigador principal o el patrocinador del estudio.
- Breve resumen: En relación con un ensayo, este elemento es una breve descripción del estudio destinada al público no especializado. Incluye una breve exposición de la hipótesis del estudio.
- Descripción detallada: En relación con un ensayo, este punto es una descripción ampliada del protocolo, que incluye más información técnica (en comparación con el Resumen breve)
- Fase: En relación con un ensayo, representa la fase/etapa de desarrollo clínico del medicamento en investigación en el proceso de desarrollo.
- Patrocinador: En relación con un ensayo, este elemento detalla la naturaleza de la participación de la(s) organización(es):
- Patrocinador principal: Organización principal que supervisa la realización del estudio y es responsable del análisis de datos.
- Colaboradores: Otra(s) organización(es) (si las hay) que prestan apoyo, incluida la financiación, el diseño, la ejecución, el análisis de datos y la elaboración de informes (se pueden enumerar hasta 10 organizaciones colaboradoras).
- Total de estudios: Número de ensayos clínicos registrados.
- Años de actividad: Número de años activos desde el primer estudio hasta el año en curso (2025).

Para la recuperación, el conjunto de datos resultante consistió en 14425 ensayos clínicos relacionados con vacunas de ARNm que se recuperaron siguiendo los siguientes criterios:

- Subtecnología: ARNm
- Localización de los ensayos: Sólo se consideraron los ensayos realizados en Europa y/o EE.UU., mientras que los ensayos realizados en otros países se excluyeron para garantizar que la totalidad de los ensayos se someten a procedimientos médicos similares.

- Código ATC: "J" (General Anti-Infectives Systemic) - "J7" (Vaccines) que recoge únicamente las vacunas relacionadas con el ARNm que se aplican a otros tipos de vacunas, como la Malaria o el Coronavirus.
- Estado general: Los ensayos con el estado aprobado para la comercialización, completado, suspendido, terminado o retenido.

Del conjunto de datos resultante de 14425 ensayos clínicos, para cada ensayo repetido, sólo se mantuvieron en el conjunto de datos final los ensayos clínicos que presentaban el patrocinador principal, manteniendo además los colaboradores, lo que dio como resultado un conjunto de datos final de 5231 ensayos para realizar el análisis.

4.3. Estructura de las patentes, recuperación y limpieza del conjunto de datos

Las patentes son herramientas jurídicas esenciales que protegen las nuevas invenciones e innovaciones concediendo derechos exclusivos a sus creadores durante un tiempo limitado. La estructura de un documento de patente está diseñada para proporcionar una descripción exhaustiva pero concisa de la invención, garantizando que esté adecuadamente protegida y permitiendo al mismo tiempo que otros comprendan sus detalles técnicos. Aunque el contenido específico de las patentes varía en función de la tecnología o el campo, por lo general contienen secciones que explican claramente la invención, su novedad y sus aplicaciones potenciales. El objetivo general de esta estructura es hacer que la invención sea públicamente accesible, apoyando la difusión del conocimiento dentro de la comunidad científica y tecnológica (EPO, 2023). Además, la estructura de la patente sirve para establecer los límites de la protección mediante reivindicaciones específicas, en las que se detallan los aspectos únicos de la invención. Al organizar y divulgar estos detalles, las patentes facilitan el progreso tecnológico y la transferencia de conocimientos entre sectores, contribuyendo a impulsar la innovación futura (Brusoni et al., 2005).

Para construir el conjunto de datos de patentes, se obtuvieron los siguientes elementos de su estructura:

- Título: Este campo contiene el nombre o la descripción de la invención, destacando su característica o concepto principal.
- Resumen: Proporciona una rápida visión general de la invención, destacando su función y objetivo principales.
- Descripción: Esta sección abarca el funcionamiento de la invención, sus componentes y detalles técnicos, incluyendo a menudo variaciones o métodos.

- ID de familia: Un identificador único para las familias de patentes. Este ID agrupa patentes relacionadas que pertenecen a la misma familia, indicando que comparten una prioridad común.
- ID de solicitud: Un identificador único para la solicitud de patente. Este número identifica la solicitud específica presentada para la invención.
- Fecha de presentación de la solicitud: La fecha en que se presentó la solicitud de patente. Este campo registra cuándo se presentó oficialmente la solicitud de patente.
- Nombre de la empresa: La entidad o persona que está detrás de la patente. Es el nombre del solicitante o de la organización que posee los derechos de la patente.
- Número completo de publicación: El número de referencia completo de la publicación. Combina la autoridad y el número de la publicación, ofreciendo una referencia completa del documento de patente.
- Tamaño de la familia: El número de presentaciones de patentes relacionadas vinculadas a un origen o tema común.
- Número de IPC: El número de códigos de clasificación distintos asignados a una patente en función de su campo tecnológico.
- Número de documentos citados: El recuento total de documentos (patentes, artículos, etc.) que han sido referenciados por una patente específica.

Las patentes utilizadas para el estudio se recuperaron de PATSTAT Online 2024 empleando una consulta con palabras clave empleadas previamente para estudios de paisajismo de patentes de ARNm por Li et al. (2022), EPO (2023), Martin & Lowery (2020), y Chen et al. (2022). Las patentes consideradas para la recuperación son las que presentan las palabras clave "ribonucleic acid " o "RNA" o "messenger ribonucleic acid " o "mRNA" en el título o en el resumen y presentan al mismo tiempo las palabras "vaccin" o "vaccine" o "vaccination" en el título o en el resumen. Aparte de las palabras clave, las patentes debían clasificarse en determinados códigos IPC (Clasificación Internacional de Patentes) o CPC (Clasificación Cooperativa de Patentes) según Li et al. (2022) y EPO (2023), como puede verse en las Tablas 2 y 3.

Tabla 2: Códigos IPC de vacunas de ARNm y sus campos (Fuentes: Li et al., 2022 y EPO, 2023)

IPC	Campos
A61K003900	Preparados medicinales que contienen antígenos o anticuerpos
A61P003500	Agentes antineoplásicos
A61K004800	Preparados medicinales que contienen material genético que se inserta en células del cuerpo vivo para tratar enfermedades genéticas; Terapia génica
A61P003114	Anti-infecciosos para virus ARN
A61K003912	Preparados medicinales que contienen antígenos virales
C12Q000168	Procesos de medición o ensayo con ácidos nucleicos
C12N000700	Virus, por ejemplo, bacteriófagos; composiciones de estos virus; preparación o purificación de estos virus
C12N001509	Tecnología de ADN recombinante
A61K0039395	Anticuerpos
A61P003112	Antivirales
C12N001586	Vectores virales relativos a la ingeniería genética
C12N000510	Células modificadas por la introducción de material genético extraño, por ejemplo, células transformadas por virus
A61K0039215	Preparados medicinales que contienen coronavirus
A61K003939	Preparados medicinales que contienen antígenos o anticuerpos caracterizados por los aditivos inmunoestimulantes
C12N000701	Virus, por ejemplo, bacteriófagos, modificados por la introducción de material genético extraño

A61K00317105	Preparados medicinales que contengan ácidos ribonucleicos naturales
C12N0015113	Ácidos nucleicos no codificantes que modulan la expresión de genes relativos a la ingeniería genética
A61M0005315	Pistones; Pistones-varillas; Guiar, bloquear o restringir el movimiento de la varilla; Dispositivos en la varilla, para facilitar la dosificación para introducir medios en o sobre el cuerpo
C12Q000170	Procesos de medición o ensayo en los que interviene un virus o bacteriófago
C12N001585	Mutación o ingeniería genética para células animales
A61K39/015	Antígenos hemsporidios
A61K39/02	Antígenos bacterianos

Tabla 3: Códigos CPC de vacunas de ARNm y sus campos (Fuente: Oficina Europea de Patentes, 2023)

CPC	Campos
A61K39/0011	Antígenos cancerígenos
A61K2039/53	Vacunación ADN/ARN
C12N2710/00	Virus ADN
C12N2730/00	Virus ADN de transcripción inversa
C12N2740/00	Transcripción inversa de ARN
C12N2760/00	Virus RNA sentido negativo
C12N2770/00	Virus RNA sentido positivo

Aparte del contenido que las patentes deben incluir en sus títulos, resúmenes y códigos de clasificación, debe haber al menos una patente en la familia de patentes con protección de la EPO o de EE.UU. Esta consideración se tomó porque al mantener patentes con protección EPO o US, el lenguaje de divulgación será en inglés por lo que

se elimina el problema de tener diferentes idiomas para la similitud, y el estar protegida en grandes oficinas de patentes significa que la patente presenta una mayor protección geográfica y es probablemente más relevante en la cartera de la empresa solicitante.

A partir de la consulta descrita anteriormente, se recuperaron 4639 patentes. De ellas, manteniendo sólo una patente por familia, priorizando las patentes protegidas por la EPO (por las razones anteriormente mencionadas) y las que tenían mayor número de citas, se obtuvo un conjunto de datos final de 1136 patentes para el análisis.

4.4. Código para la similitud de documentos

Como se indicó en el capítulo 1, para estudiar el efecto del conocimiento y la recombinación de conocimientos, el primer paso consiste en predecir la relación semántica y contextual entre determinados ensayos clínicos y patentes del mismo campo. Para ello, se eligió el modelo de PLN PAT-SPECTER, entrenado con citas entre distintos documentos. El código se basa en las librerías disponibles para Python "PyTorch", que permite acceder a los diferentes modelos transformadores como PAT-SPECTER, Pandas para gestionar los conjuntos de datos, y la librería de aprendizaje automático y modelado de datos de código abierto "Sklearn" para calcular las similitudes coseno entre los documentos. El funcionamiento general del script puede describirse como sigue:

En primer lugar, se cargan los datos y se revisan en la consola. El siguiente paso es el procesamiento de los datos y la vectorización; para ello, se cargan el modelo de lenguaje y el tokenizador (Figura 16).

```
from transformers import AutoTokenizer, AutoModel
import pandas as pd
# Get the Pat_SPECTER pretrained model path from Allen AI repo
pretrained_model = 'mpi-inno-comp/pat_specter'

# Get the tokenizer from the previous path
pat_specter_tokenizer = AutoTokenizer.from_pretrained(pretrained_model)

model = AutoModel.from_pretrained(pretrained_model)
```

Figura 16: Carga del modelo lingüístico y del tokenizador

Una vez cargado el modelo lingüístico, se utiliza una función para convertir cada corpus de texto individual en una incrustación (Figura 17), de modo que pueda calcularse la similitud coseno empleando los vectores de los documentos. La composición de la información de cada documento (utilizada de forma similar por Erhardt et al., 2024) se emplea de la siguiente manera:

- *Título oficial + Título breve + Resumen breve + Descripción detallada, para ensayos clínicos*
- *Título + Resumen, para patentes*

```
#Function to generate embeddings with PAT-SPECTER
def convert_single_abstract_to_embedding(tokenizer, model, in_text, MAX_LEN=512):

    inputs = tokenizer(
        in_text,
        add_special_tokens=True,
        max_length=MAX_LEN,
        truncation=True,
        padding="max_length",
        return_tensors="pt"
    )

    device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
    model = model.to(device)
    inputs = {key: val.to(device) for key, val in inputs.items()}

    model.eval()

    # Generate embeddings
    with torch.no_grad():
        outputs = model(**inputs)

    embedding = outputs.last_hidden_state[:, 0, :].squeeze().cpu().numpy()

    return normalized_embeddings
```

Figura 17: Función para generar incrustaciones para cada documento con PAT-SPECTER

Previamente a la transformación completa de los textos en incrustaciones, se definen dos funciones: "get_min_viable_data" para controlar el tamaño de la muestra y la información que se tendrá en cuenta, y "convert_overall_text_to_embedding", que trabaja sobre el conjunto de datos convirtiendo cada resumen en un vector mediante la función "convert_single_abstract_to_embedding"(Figura 18).

```
def get_min_viable_data(df, sample_size=6367):

    # Select only the columns we need for the analysis
    useless_cols = []

    df.drop(useless_cols, axis=1, inplace=True)
    df = df.sample(sample_size)

    return df

def convert_overall_text_to_embedding(df):

    # The list of all the embeddings
    embeddings = []

    # Get overall text data
    overall_text_data = data.appln_abstract.values

    # Loop over all the comment and get the embeddings
    for abstract in tqdm(overall_text_data):

        # Get the embedding
        embedding = convert_single_abstract_to_embedding(pat_specter_tokenizer, model, abstract)

        #add it to the list
        embeddings.append(embedding)

    print("Conversion Done!")

    return embeddings
```

Figura 18: Función para controlar la muestra de datos y la conversión global a incrustación

Posteriormente, se añade una columna de incrustaciones para almacenar la representación vectorial de cada cuerpo de texto (Figura 19), información que se utilizará en el siguiente paso para calcular las similitudes coseno entre los documentos. A continuación, se procede a crear el conjunto de datos con las incrustaciones.

```
data = get_min_viable_data(data)
embeddings = convert_overall_text_to_embedding(data)
def create_final_embeddings(df, embeddings):

    df["embeddings"] = embeddings
    df["embeddings"] = df["embeddings"].apply(lambda emb: np.array(emb))
    df["embeddings"] = df["embeddings"].apply(lambda emb: emb.reshape(1, -1))

    return df
data = create_final_embeddings(data, embeddings)
```

Figura 19: Función para añadir la representación vectorial de cada documento al conjunto de datos y creación del conjunto de datos

Finalmente, utilizando el dataset anterior creado con la representación vectorial de los documentos, se calcula la similitud coseno entre los ensayos clínicos y las patentes utilizando la librería "Sklearn". La similitud se calcula considerando por un lado los documentos con "NCT", que son el número de registro del ensayo clínico, y los documentos sin "NCT", que son el ID de solicitud de la patente. La información más relevante para la coincidencia se almacena en un nuevo archivo Excel, teniendo en

cuenta ambos códigos de identificación, la empresa que realiza el ensayo, las patentes coincidentes y la puntuación de similitud coseno entre ambos documentos (Figura 20).

```
#This block calculates all the similarities between the clinical trials and the patents of the dataset and export them to excel

from sklearn.metrics.pairwise import cosine_similarity
import pandas as pd
import numpy as np
from tqdm import tqdm

# Function to compute similarities between "nct" and "non-nct"
def compute_nct_vs_non_nct_similarity(data):
    # Filter data
    nct_data = data[data['appln_id'].str.startswith("nct")].reset_index(drop=True)
    non_nct_data = data[~data['appln_id'].str.startswith("nct")].reset_index(drop=True)

    # Extract embeddings from "non-nct"
    non_nct_embeddings = np.vstack(non_nct_data["embeddings"].values)

    # List to store results
    results = []

    # Iterate over each "nct" abstract
    for idx, row in tqdm(nct_data.iterrows(), total=len(nct_data)):
        # Get the embedding of the current abstract
        nct_embedding = row["embeddings"].reshape(1, -1)

        # Compute similarities with all "non-nct" abstracts
        similarities = cosine_similarity(nct_embedding, non_nct_embeddings).flatten()

        # Store complete results
        for target_idx, similarity in enumerate(similarities):
            results.append({
                "nct_appln_id": row["appln_id"],
                "nct_abstract": row["appln_abstract"],
                "nct_company": row["company_name"],
                "target_appln_id": non_nct_data.iloc[target_idx]["appln_id"],
                "target_abstract": non_nct_data.iloc[target_idx]["appln_abstract"],
                "target_company": non_nct_data.iloc[target_idx]["company_name"],
                "similarity": similarity
            })

    # Create a DataFrame with results and sort by similarity
    results_df = pd.DataFrame(results)
    results_df = results_df.sort_values(by="similarity", ascending=False).reset_index(drop=True)
    return results_df

# Compute full similarities
nct_vs_non_nct_results = compute_nct_vs_non_nct_similarity(data)

# Export results to Excel
nct_vs_non_nct_results.to_excel("nct_vs_non_nct_all_similarities.xlsx", index=False)

print("Export completed. File saved as 'nct_vs_non_nct_all_similarities.xlsx'.")
```

Figura 20: Función para el cálculo de la similitud y la exportación del conjunto de datos

4.5. Establecimiento del umbral de similitud del coseno

Aunque el código desarrollado anteriormente puede proporcionar información sobre el grado de relación semántica y contextual entre el contenido de las patentes y los ensayos clínicos, es necesario establecer un umbral para evaluar las coincidencias. Debido a la falta de un conjunto de datos con coincidencias validadas entre ensayos clínicos y patentes para probar el modelo y establecer un umbral, se seleccionó de clinicaltrials.gov un pequeño conjunto de 94 ensayos clínicos realizados o en curso por las pequeñas empresas farmacéuticas Argenx y Biontech entre el 01/01/2014 y el 01/01/2024, debido a su limitado y manejable número de ensayos y patentes publicadas. A partir del conjunto de datos de ensayos, se obtuvo de EvaluatePharma

un segundo conjunto de datos que revelaba los productos farmacéuticos probados en estos últimos ensayos clínicos y que empleaban patentes concedidas. Posteriormente, estas patentes relacionadas para el análisis se recuperaron de PATSTAT online.

En primer lugar, se extrajeron del conjunto de datos de productos farmacéuticos los códigos ID de publicación de las patentes ciertamente relacionadas, y después se buscaron en PATSTAT online mediante una consulta SQL para recuperar toda la información necesaria sobre las patentes para realizar el cálculo de similitud (Figura 21) y evaluar posteriormente el modelo. Tras la extracción, se limpiaron los datos y se eliminaron los duplicados, manteniendo sólo una patente por cada familia relacionada con un ensayo. El número resultante de patentes únicas y relacionadas con los ensayos clínicos fue de 256.

```
SELECT DISTINCT
  tls201_appln.docdb_family_id,
  tls201_appln.appln_id,
  tls201_appln.appln_filing_date,
  tls202_appln_title.appln_title,
  tls203_appln_abstr.appln_abstract,
  tls206_person.person_name AS applicant_name,
  tls211_pat_publn.publn_nr,
  CONCAT(tls211_pat_publn.publn_auth, tls211_pat_publn.publn_nr) AS full_publication_number
FROM
  tls201_appln
LEFT JOIN
  tls202_appln_title ON tls201_appln.appln_id = tls202_appln_title.appln_id
LEFT JOIN
  tls203_appln_abstr ON tls201_appln.appln_id = tls203_appln_abstr.appln_id
LEFT JOIN
  tls207_pers_appln ON tls201_appln.appln_id = tls207_pers_appln.appln_id
LEFT JOIN
  tls206_person ON tls207_pers_appln.person_id = tls206_person.person_id
LEFT JOIN
  tls211_pat_publn ON tls201_appln.appln_id = tls211_pat_publn.appln_id
WHERE
  tls211_pat_publn.publn_nr IN ('8637027', '8834882', '8444976', '8524231', '2461546', '6281992', '8637027',
    '5543150', '5543150', '4615697', '4814470', '4814470', '5403858', '8026250', '8026250', '7871632', '4814470',
    '6339142', '7074404', '6407213', '9814685', '9814685', '4307100', '4307100', '5543150', '4307100', '4307100',
    '8633178', '8846648', '8846649', '5543150', '5543150', '4615697', '8207290', '8207290', '1804765', '7815929',
```

Figura 21: Fragmento de consulta para la búsqueda de patentes en PATSTAT online

Una vez obtenida la información completa de los ensayos y las patentes, se formateó el conjunto de datos final y se realizó el cálculo de la similitud de los documentos utilizando el código de la sección anterior. La distribución resultante de pares coincidentes (Figura 22) y no coincidentes (Figura 23) de juicios y patentes dejó importantes pistas sobre el umbral a obtener. No obstante, la distribución y el análisis de los datos no fueron suficientes para obtener un punto de corte para los pares, por lo que se tuvo en cuenta otra metodología empleada por Rose et al. (2024) en un estudio en el que utilizaron el mismo codificador lingüístico. El método mencionado anteriormente se denomina "Corte por F1", cuyo objetivo es calcular un valor de corte óptimo de la similitud coseno para separar los pares de patentes reales de otros falsos positivos. La F1 es la media armónica de la precisión y la recuperación; equilibra la tasa de falsos positivos con la de falsos negativos, por lo que el objetivo del método es determinar el valor de similitud del coseno en el que se maximiza la puntuación F1, lo

que indica el mejor equilibrio entre precisión y especificidad para identificar los pares. La fórmula F1 puede describirse como sigue

$$F1 = 2 * \frac{Precisión * Sensibilidad}{Precisión + Sensibilidad} \quad (10)$$

Donde:

$$Precisión = \frac{Verdaderos Positivos}{Verdaderos Positivos + Falsos Positivos} \quad (11)$$

,proporción de verdaderos positivos sobre todos los positivos

$$Sensibilidad = \frac{Verdaderos Positivos}{Verdaderos Positivos + Falsos Negativos}, \text{proporción de positivos detectada} \quad (12)$$

Con esta metodología, mi objetivo era separar pares verdaderos de los falsos positivos, por lo que, probando diferentes umbrales de similitud coseno, encontré el F1 máximo en 0,14 con una similitud coseno de 0,9, que es, por tanto, el valor de umbral óptimo. Los pares de patentes y ensayos con una similitud de coseno igual o superior a este valor deberían clasificarse como posibles coincidencias. Aunque el F1 resultó ser bajo, sirve como un primer filtro importante que se empleará junto con otros métodos posteriormente. También es importante tener en cuenta que la sensibilidad resultó ser de 0,44, por lo que es posible obtener una buena proporción de verdaderos positivos con el modelo.

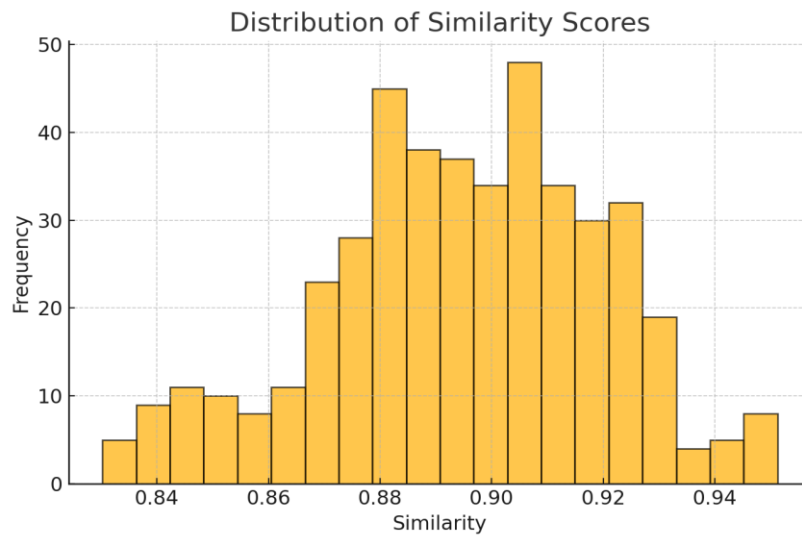


Figura 22: Distribución de las puntuaciones de similitud del coseno para vínculos verdaderos de ensayos clínicos y patentes

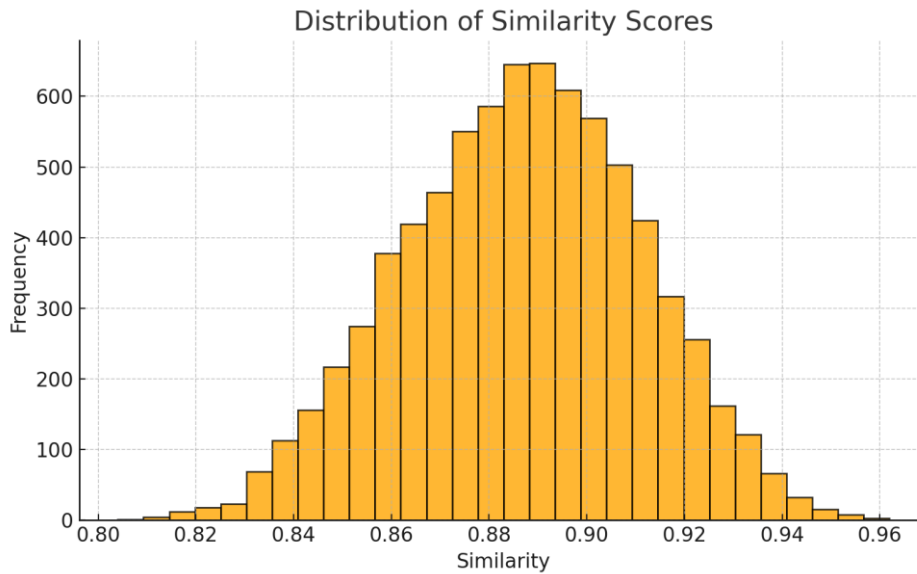


Figura 23: Distribución de las puntuaciones de similitud del coseno para vínculos de ensayos clínicos y patentes no verdaderos

4.6. Vinculación de ensayos clínicos y patentes

Tras obtener el conjunto de datos final de patentes y ensayos clínicos y el script de similitud, el siguiente paso fue ejecutar el modelo de similitud PAT-SPECTER de la sección anterior. Para ello, se consideró el umbral de 0,9 para la similitud coseno obtenido en la sección 2.5. Junto con el umbral, sólo se consideraron los pares en los que la fecha de inicio del ensayo clínico se realizó después de la fecha de solicitud de la patente más antigua de la familia de la familia relacionada, ya que, en este estudio, el enfoque es identificar los ensayos que utilizan conocimientos incorporados en las patentes. Basándonos en los dos primeros filtros descritos anteriormente, se encontraron 1.001.125 coincidencias, con una media de patentes relacionadas por ensayo de 250,84 y con 3991 ensayos clínicos distintos. A pesar de estos números y considerando la baja F1, decidí utilizar como filtro adicional para las coincidencias, que las fases de los ensayos deben ser 1,2,3 o 4, no considerando fases con "-" o vacías, ya que los ensayos sin fases son estudios de dispositivos o intervenciones conductuales. Los resultados muestran un número de patentes relacionadas de 560421 con 1601 ensayos clínicos distintos coincidentes. Así pues, el análisis empírico se realizará con los 1601 ensayos distintos que resulten estar relacionados con su respectivo número de patentes.

2.7 Variables

2.7.1 Variable dependiente

En este estudio se utilizaron datos de ensayos clínicos para analizar los factores de conocimiento y recombinación de conocimientos presentes en patentes que influyen en el desarrollo de fármacos. El éxito de los ensayos puede medirse por la fase del ensayo clínico que se esté llevando a cabo. Con este fin, los datos se codificaron como 1 para los productos posiblemente próximos al mercado que se están probando en ensayos de fase 3 o 4 y 0 para los productos un poco más alejados del mercado que se están probando en fases preclínicas y en fase 1 o 2.

2.7.2 Variables independientes

Las siguientes variables independientes se utilizan para abordar la pregunta de investigación de este estudio, estas variables intentan explicar el efecto del conocimiento incorporado en las patentes y la recombinación del conocimiento en la variable dependiente.

-Número de patentes relacionadas: Representa las patentes relacionadas con un determinado ensayo clínico como resultado del cálculo de similitud coseno con el modelo anteriormente descrito. Estas patentes relacionadas a menudo sirven de base para la tecnología del ensayo y proporcionan información sobre el panorama de conocimientos. Un mayor número de patentes relacionadas puede reflejar una base tecnológica más avanzada y bien establecida para el ensayo.

-Tamaño acumulativo de la familia de patentes: Esta variable representa la suma de los tamaños de las familias de patentes asociadas al ensayo clínico. El tamaño de la familia de patentes se refiere al número de países en los que se presenta una patente. Un mayor tamaño de la familia indica que la invención se ha considerado lo suficientemente valiosa como para ser protegida en múltiples jurisdicciones. Esto puede sugerir la importancia mundial y la comerciabilidad potencial de la tecnología subyacente al ensayo (Wagner y Wakeman, 2016).

-Documentos citados acumulativos: Representa la suma de todos los distintos documentos citados en todas las familias de patentes que están relacionados con el ensayo clínico. Estos documentos citados suelen incluir referencias a otras patentes, publicaciones científicas o presentaciones normativas que las solicitudes de patentes han citado en su solicitud de patente. Esta información es crucial para comprender el panorama de la propiedad intelectual y el estado de la técnica considerado durante los procesos de creación y examen de patentes (Sharma y Tripathi, 2017). Este indicador

podría ser un indicador indirecto de cómo los productos farmacéuticos de éxito se nutren indirectamente de diferentes fuentes de conocimiento integradas en las patentes.

-Códigos IPC acumulativos: Esta variable representa la suma del número de Clasificaciones Internacionales de Patentes (IPC) asociadas a las patentes relevantes para el ensayo clínico. Los IPC se utilizan para categorizar el contenido tecnológico de las patentes. Un mayor número de IPC indica un alcance tecnológico más amplio, una mayor complejidad de las tecnologías desarrolladas en el ensayo y una recombinación de conocimientos presente en las patentes relacionadas (Xie et al., 2025).

2.7.3 Variables de control

También se incluyen varias variables de control para tener en cuenta factores distintos de la dependencia de un determinado acervo de conocimientos, representado por las patentes. Estas variables pretenden cubrir la influencia de otros ensayos y factores relacionados con la empresa.

-Número de colaboradores: Esta variable representa el número total de organizaciones o entidades que colaboran en un ensayo clínico concreto. Cuanto mayor sea el número de colaboradores, más multidisciplinar será el ensayo. Esto puede aumentar las probabilidades de éxito, ya que varias organizaciones aportan su experiencia, recursos y capacidades de investigación únicos. Una mayor colaboración también puede indicar una red más amplia de apoyo y financiación.

-Experiencia de la firma: Representa el número de ensayos clínicos realizados anteriormente por la misma organización. Pretende captar la experiencia y pericia que tiene una organización en términos de desarrollo de fármacos y ensayos (Garzón-Vico et al., 2020).

-Antigüedad de la firma: Esta variable representa el tiempo que la institución que realiza el ensayo lleva activa desde el inicio de su primer ensayo clínico. Está relacionada con el avance tecnológico y también puede indicar la dotación de conocimientos de una organización (Wang et al., 2023).

-Competencia: Esta variable representa el nivel de competencia en el mercado o área tecnológica del ensayo clínico, en este caso, el campo de las vacunas de ARNm. Se mide como el número ensayos llevadas a cabo por instituciones competidoras en el mismo campo tecnológico y en el momento en que comenzó el ensayo focal. Un valor más alto de competencia sugiere que hay muchos ensayos en el campo, lo que puede aumentar la dificultad de obtener la aprobación regulatoria o lograr una posición de liderazgo en el mercado. En áreas de investigación muy concurridas, los ensayos

clínicos pueden tener más dificultades para avanzar por las fases debido a la competencia por la aprobación reguladora y a la financiación limitada.

-Año de inicio (variable categórica): Esta variable recoge el año natural en el que comenzó cada ensayo clínico. Para tener en cuenta los posibles efectos temporales en la progresión y el éxito de los ensayos, esta variable se incorpora como un conjunto de variables categóricas en el análisis de regresión. Cada variable representa un año concreto, lo que permite al modelo controlar factores temporales no observados, como cambios en el entorno normativo, avances tecnológicos o cambios en la financiación, que podrían influir en la probabilidad de que un ensayo pase a fases posteriores. Al incluir estos indicadores específicos de cada año, el análisis pretende aislar el efecto de la dinámica temporal en la variable dependiente, garantizando que los impactos estimados de otras variables explicativas no se vean confundidos por influencias relacionadas con el tiempo.

-Tipo de patrocinador (variable categórica): Esta variable clasifica la organización principal responsable de la realización del ensayo clínico en tres grupos distintos: Empresas comerciales, Universidades/Centros de investigación y Agencias gubernamentales/Centros de investigación. Esta clasificación se introduce como una variable categórica utilizando una codificación ficticia, con una categoría que sirve como grupo de referencia. La inclusión del tipo de patrocinador permite que el análisis evalúe cómo la naturaleza de la institución patrocinadora influye en la progresión de los ensayos clínicos. Los distintos tipos de patrocinador pueden variar en cuanto a recursos, objetivos estratégicos, experiencia regulatoria y tolerancia al riesgo, todo lo cual puede influir en el diseño, la ejecución y el avance de los ensayos. Al examinar los efectos asociados a cada categoría de patrocinador, el estudio pretende comprender el papel que desempeñan las características institucionales en la trayectoria de éxito de los ensayos clínicos.

2.8 Modelo empírico

Teniendo en cuenta la naturaleza binaria de la variable dependiente, se utilizó un modelo de regresión logística para estimar la probabilidad de que un ensayo clínico avance a una fase posterior en función de múltiples variables explicativas relacionadas con el conocimiento tecnológico, el esfuerzo de colaboración y las características organizativas. La significación estadística de las variables independientes que afectan al éxito de los ensayos clínicos se midió mediante valores p de $<0,01$, $<0,05$ y $<0,1$. Mientras que la magnitud de los efectos se interpretó empleando razón de

probabilidades (odds ratio). Los datos recopilados se analizaron mediante el software RStudio, y el modelo logístico se ajustó de la siguiente manera:

$$\text{logit}(\text{Pr}(\text{Phase} = 1)) = \beta_0 + \beta_1 X_1 \dots \dots \dots \beta_i X_i \quad (13)$$

Donde

- $\text{logit}(p) = \log\left(1 - \frac{p}{1-p}\right)$, p es la probabilidad de que el ensayo clínico se encuentre en la fase avanzada 3 o 4 (Fase = 1).
- $I=\{1\dots 8\}$

5 Resultados y discusión

En este capítulo, ilustraré los resultados del análisis realizado con respecto a las coincidencias ensayo clínico-patente encontradas en el apartado anterior. El objetivo del capítulo es reflexionar sobre el comportamiento de las variables y el efecto que las variables independientes y controles previamente establecidos tienen sobre la variable dependiente, y cómo en conjunto pueden explicar el éxito en alcanzar fases avanzadas de desarrollo y por tanto la cercanía a la comercialización.

3.1 Análisis exploratorio

Antes de interpretar los modelos de regresión, esta sección presenta estadísticas descriptivas, correlaciones y distribuciones de variables. Esto permite obtener una visión preliminar de la calidad de los datos, las posibles transformaciones y las relaciones entre variables.

Empezando por la variable dependiente, la distribución de los ensayos clínicos entre las fases de desarrollo revela que 1.045 ensayos, equivalentes a aproximadamente el 65% del conjunto de datos, se encuentran en las fases 1 o 2, mientras que el 35% restante ha progresado a fases más avanzadas, concretamente las fases 3 o 4. Esta distribución desigual refleja los retos bien documentados y los obstáculos normativos asociados con el avance de los ensayos clínicos basados en ARNm hacia fases más avanzadas de desarrollo. Volviendo a las variables independientes, la Tabla 4 presenta las estadísticas descriptivas, basadas en un total de 1.601 ensayos clínicos únicos, como se explicó anteriormente en el capítulo anterior. Cada uno de estos ensayos está vinculado a un conjunto de patentes relacionadas. Por término medio, cada ensayo está asociado a aproximadamente 350 patentes únicas, y el 67% de todas las patentes relacionadas están vinculadas a ensayos de fase inicial, mientras que el 33% restante corresponde a fases más avanzadas. Como era de esperar, todas las observaciones incluyen al menos una patente clasificada bajo múltiples códigos IPC, lo que pone de relieve la presencia sistémica de la recombinación tecnológica y señala la naturaleza multidimensional del conocimiento que se moviliza en la innovación del ARNm. En cuanto a la colaboración, el número medio de colaboradores por ensayo se sitúa en torno a dos. Sin embargo, en el 60% de los casos, sólo participa una entidad, normalmente el patrocinador principal, lo que refleja que una parte sustancial de los ensayos se realizan con una estructura de agente único. Este detalle ofrece una visión de las estrategias organizativas empleadas en el proceso de desarrollo y sugiere que,

en muchos casos, las empresas optan por un modelo de innovación más centralizado y gestionado internamente.

En cuanto al tipo de patrocinador principal (Tabla 5), puede observarse que aproximadamente el 46% de los ensayos clínicos están dirigidos por empresas comerciales, seguidas de cerca por el 44% patrocinado por universidades, hospitales e instituciones de investigación sin ánimo de lucro. Una proporción menor, en torno al 10%, es llevada a cabo por organismos gubernamentales, lo que refleja una participación relativamente modesta pero no por ello menos significativa del sector público en el patrocinio de ensayos clínicos. Por lo que respecta a la variable categórica que representa el año de inicio de los ensayos, es especialmente destacable que el 45% de los ensayos comenzaron en 2020. Esta concentración se corresponde con el estallido de la pandemia de COVID-19, una crisis sanitaria mundial que catalizó una oleada de actividad investigadora sin precedentes, especialmente en el campo de las tecnologías basadas en el ARNm y el desarrollo de vacunas (Organización Mundial de la Propiedad Intelectual [OMPI], 2023). Esta distribución temporal pone de relieve el impacto directo de las perturbaciones externas en el ritmo y el enfoque de la innovación biomédica.

Tabla 4: Estadísticas descriptivas de las variables independientes y los controles

Variable independiente	n	Media	Desv.Est.	Mín	Max
# Patentes relacionadas	1601	350.04	278.70	1.00	1059.00
Tamaño acumulado de las familias	1601	2835.35	2226.51	6	8038.00
Códigos IPC acumulados	1601	2646.77	2081.84	5	7602.00
Documentos citados acumulados	1601	6653.38	5572.55	0	20428.00
# Colaboradores	1601	1.87	1.89	1.00	29.00
Experiencia de la firma	1601	2079.05	3525.74	1.00	20237.00
Edad de la firma	1601	28.79	16.25	2.00	125.00
Competencia I+D	1601	545.79	295.60	0.00	931.00

Tabla 5: Frecuencias de las diferentes categorías de promotor principal del ensayo

Categoría	Recuento	Porcentaje
Agencia gubernamental/centro de investigación	151	9.43
Empresa comercial	734	45.85
Universidad/Centro de investigación	716	44.72

Como puede observarse en la Tabla 6, la correlación entre las variables relacionadas con las patentes resultó ser muy alta ($\text{cor} > 0,9$), lo que puede explicarse por su origen común en las patentes relacionadas de cada ensayo. En el caso de los controles, no están correlacionados significativamente con ninguna de las variables.

Tabla 6: Tabla de correlaciones de las variables independientes y los controles

	# Collaborators	Firm experience	Firm age	R&D Competition	# Related patents	Cumulative family size	Cumulative IPC codes	Cumulative cited documents
# Collaborators	1							
Firm experience	0.06	1						
Firm age	0.15	0.65	1					
R&D Competition	0	-0.15	-0.2	1				
# Related patents	-0.06	-0.11	-0.16	0.02	1			
Cumulative family size	-0.06	-0.11	-0.16	0.05	1	1		
Cumulative IPC codes	-0.05	-0.1	-0.14	0.01	1	1	1	
Cumulative cited documents	-0.06	-0.14	-0.21	0.15	0.96	0.96	0.94	1

Las variables independientes, como puede verse en la Figura 24, están muy sesgadas a la derecha y tienen colas largas. Por lo tanto, para mejorar la interpretabilidad, en la siguiente sección se consideraron las transformaciones logarítmicas de estas variables.

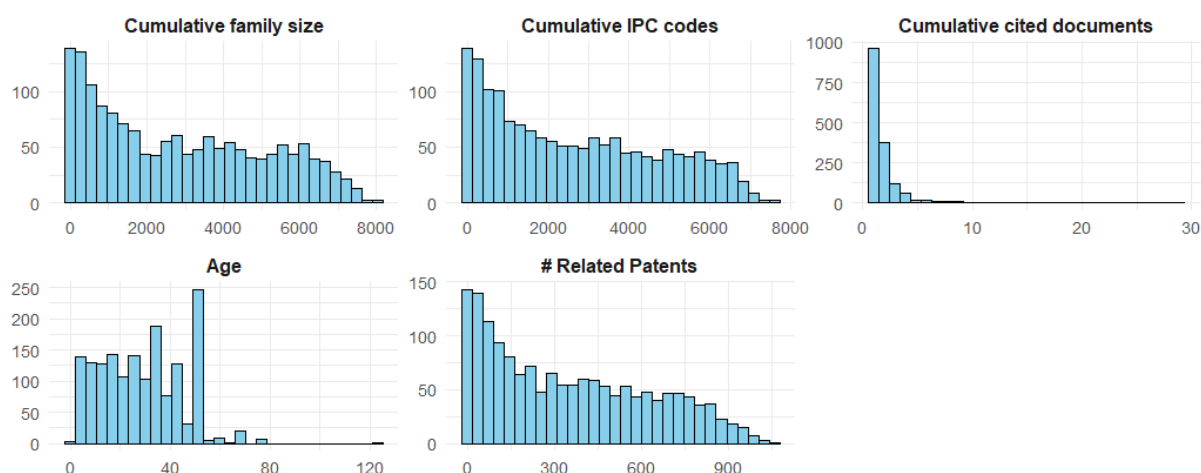


Figura 24: Histogramas de variables sesgadas

3.2 Resultados de la regresión logística

Tras el análisis exploratorio y un examen minucioso tanto de los patrones de correlación como de las distribuciones de las variables, se estimaron cuatro modelos de regresión logística, cada uno de los cuales incorporaba una variable distinta relacionada con las patentes. Esta estrategia de modelización se adoptó en respuesta a la importante colinealidad observada entre los indicadores relacionados con las patentes, que de otro modo podría comprometer la fiabilidad de las estimaciones. En consecuencia, cada variable relacionada con las patentes se introdujo por separado junto con las variables de control para aislar su efecto directo sin aumentar la multicolinealidad.

Los resultados de la regresión y los correspondientes ajustes del modelo se presentan en el cuadro 7. Como se ha señalado, tanto las variables independientes como la variable de control, la edad de la empresa, se transformaron logarítmicamente. Esta transformación fue motivada por la presencia de distribuciones muy asimétricas y de cola larga, como se ha comentado anteriormente. Además, como se ilustra en el Apéndice A.1, aunque las variables independientes originales (sin transformar) arrojaron resultados estadísticamente significativos, el tamaño de sus efectos fue mínimo. Por lo tanto, se aplicó la transformación logarítmica para mejorar la interpretabilidad y captar mejor la magnitud de sus asociaciones. Un razonamiento similar se aplicó a la edad de la empresa, cuya transformación logarítmica dio lugar a una mejora consistente en el ajuste del modelo, como se documenta en el Apéndice A.2.

Para facilitar la interpretación del efecto de las variables sobre la variable dependiente teniendo en cuenta las transformaciones logarítmicas, se empleó la razón de probabilidades para la duplicación de cada variable de la siguiente manera:

$$\text{Razón de probabilidades por duplicación de } x = \exp(\text{coef} * \ln(2)) - 1 \quad (14)$$

Mientras que para las variables no transformadas se utilizó de forma similar la razón de probabilidades simple:

$$\text{Razón de probabilidades de } x = \exp(\text{coef}) - 1 \quad (15)$$

La variable transformada logarítmicamente que representa el número de patentes relacionadas muestra una asociación negativa estadísticamente significativa con la probabilidad de que un ensayo clínico avance a fases de desarrollo avanzadas. En concreto, una duplicación del número de patentes relacionadas corresponde a una disminución del 8% en las probabilidades de alcanzar fases posteriores. Esta relación inversa concuerda con las prácticas establecidas en la industria farmacéutica, donde las solicitudes de patentes suelen presentarse antes del inicio de los ensayos clínicos. Esta estrategia se emplea para salvaguardar la propiedad intelectual, ya que las divulgaciones públicas durante los ensayos clínicos, como los listados en ClinicalTrials.gov, los comunicados de prensa o las presentaciones en conferencias que pueden considerarse estado de la técnica, comprometen potencialmente la novedad requerida para la patentabilidad. Dado que la mayoría de los ensayos del conjunto de datos se encuentran en fases tempranas de desarrollo, la correlación negativa observada puede reflejar la práctica común de presentar patentes tempranamente para mitigar los riesgos asociados a las divulgaciones públicas durante el desarrollo clínico. Sin embargo, este hallazgo no implica necesariamente que un mayor número de patentes relacionadas provoque directamente que los ensayos permanezcan en fases tempranas. En cambio, puede indicar que se produce una amplia actividad de patentamiento al principio del proceso de desarrollo, en particular para las innovaciones con resultados clínicos inciertos.

Partiendo de la observación anterior, el tamaño acumulado de la familia de patentes también muestra una asociación negativa estadísticamente significativa con el avance de los ensayos clínicos, y una duplicación del tamaño de la familia corresponde a una disminución del 7,8% en las probabilidades de alcanzar fases de desarrollo posteriores. Esta relación puede entenderse en el contexto de las prácticas estratégicas de patentamiento, en particular la formación de los denominados "matorrales de patentes". En el sector biofarmacéutico, las empresas suelen construir redes de patentes complejas y superpuestas para proteger compuestos biológicos que abarcan formulaciones, procesos de fabricación, mecanismos de administración y tecnologías auxiliares. Aunque estas protecciones multicapa pueden aumentar la exclusividad y disuadir a la competencia, también pueden generar importantes obstáculos jurídicos y administrativos. Esta maraña puede dar lugar a un aumento de los litigios, a negociaciones prolongadas para la concesión de licencias y la colaboración, y a una mayor incertidumbre en cuanto a la libertad para operar. En conjunto, estos factores desalientan las colaboraciones externas y complican la ejecución de los ensayos, lo que puede retrasar la aprobación reglamentaria y la entrada en el mercado (Goode y Chao, 2022). El caso de las tecnologías basadas en el ARNm ofrece un ejemplo claro, ya que

este campo se caracteriza por un panorama de patentes denso y fragmentado que, según se ha reconocido ampliamente, impide la innovación y crea un entorno difícil para el desarrollo clínico (Adekola, 2023).

Al abordar el efecto de recombinación de conocimientos a lo largo de los ensayos, la variable que representa los códigos IPC acumulativos muestra una asociación negativa estadísticamente significativa con el éxito del ensayo clínico, lo que indica que una duplicación en el número de clases de IPC asociadas con un ensayo dado corresponde a una disminución del 8,5% en las probabilidades de avanzar a fases de desarrollo posteriores. Este hallazgo sugiere que un ámbito tecnológico más amplio, reflejado en la diversidad de códigos de la Clasificación Internacional de Patentes, puede introducir complejidad o diluir el enfoque estratégico durante el desarrollo clínico. Aunque la investigación en las primeras fases puede beneficiarse de los esfuerzos exploratorios que abarcan múltiples ámbitos tecnológicos, la progresión hacia innovaciones revolucionarias en fases posteriores suele exigir una trayectoria tecnológica más específica y coherente. Así pues, la relación negativa observada puede reflejar los retos que plantea la gestión de aportaciones tecnológicas polifacéticas en entornos clínicos muy regulados y con un uso intensivo de recursos.

La variable que representa los documentos citados acumulados muestra una asociación negativa estadísticamente significativa con la progresión de los ensayos clínicos, con una duplicación del número de patentes citadas asociada a una disminución del 6,5% en las probabilidades de alcanzar fases avanzadas de desarrollo. Este resultado es coherente con los patrones observados para otros indicadores relacionados con la amplitud, como los códigos IPC acumulativos. Desde la perspectiva de la teoría de la recombinación de conocimientos, un mayor número de citas puede indicar la existencia de una base tecnológica amplia y diversa. Si bien esto puede permitir una amplia exploración y una adaptación incremental, también puede dar lugar a mayores cargas de coordinación, ineficiencias de ejecución y estrategias de desarrollo fragmentadas. En entornos clínicos complejos, tal dispersión puede limitar la claridad y especificidad de la dirección tecnológica, complicando el diseño de los ensayos y ralentizando el avance normativo o científico. En consecuencia, los programas clínicos que se basan en conjuntos de conocimientos previos más estrechamente enfocados y cohesionados pueden estar mejor posicionados para lograr resultados de desarrollo impactantes que los que se difunden a través de numerosas fuentes tecnológicas.

Por el contrario, las variables de control muestran en su mayoría una relación positiva con la variable dependiente. El número de colaboradores presenta una asociación positiva con la variable dependiente, mientras que cada colaborador adicional en el ensayo aumenta las probabilidades de alcanzar fases altas de desarrollo en un 14,5%. Este dato subraya el papel del intercambio colaborativo de conocimientos en la innovación, en tanto que los colaboradores aportan experiencia diversa, amplitud de conocimientos, acceso a redes más amplias y capacidades complementarias, todo lo

cual mejora la capacidad de la empresa para superar las barreras técnicas, clínicas o normativas. La competencia en I+D también desempeña un papel. Asociado positivamente con la dependiente, el valor de esta variable significa que, por cada ensayo activo adicional realizado por competidores, las probabilidades de alcanzar fases avanzadas aumentan en un 0,1%. Esto sugiere que los entornos competitivos pueden impulsar un mayor rendimiento porque las empresas en ecosistemas de I+D más dinámicos pueden beneficiarse de la difusión de conocimientos, la presión de la evaluación comparativa o la aceleración de la resolución de problemas debido a la proximidad de otros agentes activos. El control transformado logarítmicamente Edad de la firma presentó una asociación positiva con la variable dependiente, aumentando las probabilidades en un 14,1% y duplicando su valor. Lógicamente, las empresas más antiguas tienen sistemas internos más maduros con mayor capacidad de integración de conocimientos, un capital de reputación más sólido y características de conocimiento tácito acumulado que pueden favorecer el éxito del ensayo clínico.

Inesperadamente, la variable de control Experiencia de la firma, medida como el número acumulado de ensayos clínicos previos realizados por una empresa, muestra una asociación estadísticamente significativa pero negativa con la probabilidad de que un ensayo avance a fases de desarrollo posteriores. En concreto, cada ensayo previo adicional se asocia con una disminución muy pequeña del 0,01% en las probabilidades de avance. Aunque la dirección del efecto puede parecer contraintuitiva, su importancia estadística sugiere un patrón coherente en los datos. Una interpretación plausible es que las empresas con una cartera más amplia de ensayos anteriores pueden participar más activamente en estudios exploratorios en fase inicial, que son intrínsecamente menos propensos a progresar. Dado que el conjunto de datos incluye una elevada proporción de ensayos en fases iniciales, esto puede explicar en parte la asociación negativa. Por lo tanto, el resultado no debe interpretarse como una prueba de que la experiencia obstaculiza el progreso, sino más bien de que una mayor experiencia puede correlacionarse con una mayor actividad de investigación en fases iniciales, lo que reduce estadísticamente las probabilidades medias de progreso en toda la cartera de ensayos de una empresa.

Tabla 7: Resultados del modelo de regresión logística

	<i>Dependent variable:</i>			
	Phase			
	Model 1 (1)	Model 2 (2)	Model 3 (3)	Model 4 (4)
# Collaborators	0.136*** (0.030)	0.135*** (0.030)	0.136*** (0.030)	0.135*** (0.030)
Firm experience	-0.0001*** (0.00002)	-0.0001*** (0.00002)	-0.0001*** (0.00002)	-0.0001*** (0.00002)
log.Firm age	0.196** (0.091)	0.195** (0.092)	0.197** (0.091)	0.190** (0.092)
R&D Competition	0.001*** (0.0002)	0.001*** (0.0002)	0.001*** (0.0002)	0.001*** (0.0002)
log(# Related patents)	-0.120*** (0.035)			
log(Cumulative family size)		-0.119*** (0.034)		
log(Cumulative IPC codes)			-0.128*** (0.036)	
log(Cumulative cited documents)				-0.097*** (0.029)
Constant	-1.358*** (0.370)	-1.110*** (0.411)	-1.048** (0.416)	-1.224*** (0.394)
Observations	1,601	1,601	1,601	1,601
Log Likelihood	-986.156	-985.892	-985.442	-986.171
Akaike Inf. Crit.	1,984.313	1,983.784	1,982.885	1,984.342
<i>Note:</i>			*p<0.1; **p<0.05; ***p<0.01	

3.3 Discusión e implicaciones

Los resultados presentados anteriormente concuerdan en gran medida con investigaciones anteriores sobre innovación y desarrollo farmacéutico, y ofrecen explicaciones teóricas para las asociaciones negativas observadas entre las variables relacionadas con la propiedad intelectual y el avance de los ensayos clínicos. Los coeficientes negativos y estadísticamente significativos tanto para el número de patentes relacionadas como para el tamaño acumulado de la familia de patentes sugieren que una actividad de patentes más extensa, aunque comúnmente interpretada como un marcador de valor tecnológico, también puede correlacionarse con retrasos estratégicos en la progresión del desarrollo. Como destacan Wagner y Wakeman (2016), en el contexto de los productos farmacéuticos de alto valor o aquellos con una incertidumbre considerable respecto a sus especificaciones finales, las empresas a menudo posponen las principales decisiones de desarrollo hasta que se aclaran sus posiciones en materia de propiedad intelectual. Este comportamiento refleja un intento deliberado de resolver la incertidumbre jurídica y técnica antes de comprometerse con fases de prueba que exigen muchos recursos. En estos casos, el gran número de patentes en las primeras fases puede ser un indicador del retraso en

el desarrollo, ya que las empresas consolidan sus estrategias de protección antes de proceder al desarrollo clínico a gran escala.

El caso del tamaño acumulativo de las familias de patentes también refuerza esta interpretación. Las grandes familias suelen surgir en sectores caracterizados por una gran complejidad tecnológica y un escrutinio reglamentario, como el de los productos biológicos. Si bien estas familias expansivas ayudan a garantizar la ventaja competitiva y a disuadir de la imitación, también tienden a formar los llamados matorrales de patentes, que son redes densas y superpuestas de derechos que pueden complicar significativamente la concesión de licencias, la colaboración y la navegación reguladora (Goode y Chao, 2022). Esta complejidad puede crear cuellos de botella en la ejecución clínica, sobre todo cuando intervienen múltiples titulares de derechos, jurisdicciones o componentes tecnológicos. Como se ha visto en el campo de las tecnologías basadas en el ARNm (el campo en el que se centra este estudio), la proliferación de marañas de patentes ha sido ampliamente criticada por ralentizar la innovación e introducir una incertidumbre sustancial en las vías de desarrollo y comercialización (Adekola, 2023).

Al ampliar el análisis a los indicadores de recombinación de conocimientos, surgen capas adicionales de explicación. La variable que recoge el número acumulado de códigos IPC, utilizada como indicador de la amplitud tecnológica, también muestra una asociación negativa con la progresión. Esto concuerda con los hallazgos de Xie et al. (2025), que subrayan que más del 90% de las invenciones farmacéuticas surgen de la recombinación de múltiples dominios tecnológicos. Aunque esta amplitud puede promover combinaciones novedosas y una investigación exploratoria temprana, no siempre favorece una ejecución eficaz. La diversidad de bases tecnológicas puede plantear problemas de coordinación, aumentar la carga cognitiva de los equipos de desarrollo y reducir la armonización con los marcos normativos. El paso de la exploración a la ejecución, sobre todo en las últimas fases de desarrollo, exige a menudo una trayectoria tecnológica más centrada y coherente. Esta tensión entre la amplitud y la ejecución puede ayudar a explicar por qué tantas invenciones basadas en la recombinación finalmente no logran progresar más allá de las primeras fases clínicas. En apoyo de esto, BIO et al. (2021) informan de que aproximadamente el 85% de los productos farmacéuticos que se basan en gran medida en conocimientos recombinados no llegan a la fase III, lo que refuerza la idea de que la diversidad tecnológica por sí sola es insuficiente para el éxito del avance clínico.

Una dinámica comparable es evidente en la variable que recoge los documentos citados acumulados. Esta medida, que refleja la amplitud de los conocimientos tecnológicos y científicos previos a los que se hace referencia en las patentes relacionadas con un ensayo, también se correlaciona negativamente con el avance clínico. Desde la perspectiva de la teoría de la recombinación de conocimientos, esto sugiere que la dependencia de una amplia gama de trabajos anteriores puede facilitar avances incrementales en la investigación en fase inicial, pero a costa de la coherencia

estratégica en el desarrollo posterior. La literatura reciente ha destacado un estancamiento en la novedad combinatoria de las nuevas patentes farmacéuticas, con empresas que reciclan cada vez más combinaciones conocidas en lugar de desarrollar cambios tecnológicos transformadores (Xie et al., 2025). En este contexto, una mayor atención a la profundidad, más que al mero alcance, de los conocimientos previos puede ser más propicia para el éxito clínico. Como señalan Zhang et al. (2021), Kaplan y Vakili (2015), y Breschi et al. (2003), la innovación de alto impacto se suele basar en un compromiso centrado en una base tecnológica bien desarrollada, en lugar de en la difusión a través de muchos ámbitos desconectados. En consecuencia, los programas clínicos que se basan en un conjunto de capacidades concentradas y estratégicamente alineadas pueden estar mejor equipados para hacer frente a las exigencias rigurosas y de gran intensidad de recursos de la ejecución de ensayos en fases avanzadas.

Los resultados relativos a las variables de control corroboran la interpretación más amplia del modelo y se ajustan a los marcos teóricos establecidos en la gestión de la innovación y el desarrollo clínico. El efecto positivo y estadísticamente significativo del número de colaboradores corrobora la premisa bien documentada de que los esfuerzos de colaboración mejoran los resultados de la innovación al amalgamar capacidades complementarias, experiencia interdisciplinar y acceso a redes científicas e institucionales expansivas. Estas configuraciones colaborativas son especialmente ventajosas en la intrincada labor de los ensayos clínicos, donde la coordinación multidisciplinar, la infraestructura compartida y el intercambio de conocimientos en tiempo real son fundamentales para avanzar en las fases del ensayo. Esta observación es coherente con las pruebas empíricas presentadas por Dong et al. (2018), que subrayan que la diversidad de la colaboración se correlaciona significativamente con la producción de innovación científica. Además, el National Center for Advancing Translational Sciences (NCATS) ha demostrado que las colaboraciones público-privadas pueden acelerar el desarrollo terapéutico aprovechando las fortalezas de diversos socios (Wegner et al., 2022).

Una tendencia paralela es evidente en lo que respecta a la competencia en I+D, que también muestra una influencia positiva y significativa en la progresión de los ensayos clínicos. Este hallazgo respalda la idea de que los entornos de innovación competitivos catalizan el rendimiento al obligar a las empresas e instituciones a acelerar los plazos de desarrollo, mitigar las ineficiencias y garantizar una entrada temprana en el mercado. En los sectores caracterizados por una dinámica en la que el ganador se lo lleva todo, como el farmacéutico, la necesidad imperiosa de lograr la ventaja del pionero puede impulsar una ejecución más concentrada y rápida de los programas clínicos (Yang et al., 2025). Esto es especialmente pertinente en ámbitos terapéuticos en los que los periodos de exclusividad son limitados y el tiempo de comercialización está intrínsecamente ligado tanto a la viabilidad comercial como a los resultados en materia de salud pública.

La asociación positiva entre la edad de la firma y la progresión de los ensayos clínicos subraya el papel fundamental de la madurez organizativa en la eficiencia de la investigación. Las empresas consolidadas suelen desarrollar una amplia experiencia, una infraestructura sólida y una profunda familiaridad con los requisitos normativos, lo que les permite gestionar eficazmente las complejidades inherentes a las fases posteriores del desarrollo clínico. Estas capacidades facilitan una mejor coordinación, cumplimiento e inscripción de pacientes, todo lo cual es esencial para avanzar con éxito en los ensayos. Como destacan Glass et al. (2016), las empresas con una investigación clínica más antigua tienden a lograr una mayor eficiencia en los ensayos y mayores tasas de éxito, lo que refleja su experiencia acumulada y su conocimiento institucional. Por lo tanto, la longevidad sirve como un valioso indicador de la solidez y sofisticación de los sistemas internos que apoyan la innovación clínica y la ejecución de los ensayos.

Por el contrario, el coeficiente negativo y pequeño pero estadísticamente significativo, asociado a la experiencia de la empresa, cuantificada como el número de ensayos previos realizados, no estaba previsto. Aunque el tamaño del efecto es mínimo, su dirección puede ser indicativa de la composición del conjunto de datos o de una propensión inherente de las empresas con experiencia a participar en un mayor volumen de ensayos exploratorios en fase inicial. Estos ensayos, aunque contribuyen a la experiencia acumulada de una empresa, pueden tener intrínsecamente una menor probabilidad de progresión. Esta interpretación difiere de hallazgos anteriores, como los de Wang (2023) o Glass et al (2016), que sugieren que la experiencia de la empresa tiene una correlación neutra o positiva con el avance de los ensayos. En consecuencia, es probable que el resultado actual pueda atribuirse a matices contextuales en lugar de representar una relación causal sólida o generalizable.

En conjunto, estos resultados tienen implicaciones significativas para abordar la pregunta central de la investigación: ¿cómo influye el conocimiento, en particular el incorporado en las patentes, en el avance de los ensayos clínicos? Las persistentes asociaciones negativas observadas para los indicadores relacionados con las patentes, incluidos los recuentos de patentes relacionadas, el tamaño acumulado de la familia, el alcance tecnológico (medido por los códigos IPC) y los conocimientos citados, ponen de relieve la tensión entre la innovación exploratoria y la ejecución clínica. Si bien las estructuras de conocimiento expansivas y diversas pueden fomentar la ideación, señalar la sofisticación tecnológica y facilitar la recombinación, al mismo tiempo pueden imponer costes de coordinación, complejidades jurídicas y ambigüedades estratégicas. Estas disyuntivas resultan especialmente llamativas en contextos clínicos, donde el éxito de los avances requiere no sólo potencial científico, sino también claridad operativa, alineación normativa y enfoque organizativo.

En cambio, los efectos positivos de la colaboración, la madurez institucional y la competencia en I+D ponen de relieve el papel fundamental de las capacidades estructurales y relacionales para movilizar e integrar eficazmente el conocimiento.

Estos resultados sugieren que los factores determinantes del éxito clínico van más allá de la mera cantidad o diversidad de conocimientos, siendo más primordial la capacidad de las empresas e instituciones para gobernar y aplicar esos conocimientos de forma selectiva, estratégica y orientada a la ejecución. Esta perspectiva concuerda con la creciente literatura que subraya que la innovación de alto impacto depende no sólo de los activos tecnológicos previos, sino también de las capacidades organizativas posteriores.

6 Conclusiones y limitaciones

Este capítulo final resume las principales conclusiones de la investigación, cuyo objetivo era examinar la relación entre el conocimiento tecnológico incorporado en las patentes y el éxito de los ensayos clínicos. Para investigar esto, se desarrolló un modelo de similitud semántica para emparejar textos de ensayos clínicos con documentos de patentes utilizando PAT-SPECTER, un modelo codificador de lenguaje de última generación entrenado en datos de citas de varios dominios científicos. Este modelo facilitó la construcción de un nuevo conjunto de datos que vincula patentes y ensayos clínicos en el campo de las tecnologías de vacunas de ARNm, un área que se ha vuelto particularmente prominente e intensiva en innovación desde 2020.

Los datos, extraídos de PATSTAT online (para patentes) y ClinicalTrials.gov (para ensayos), permitieron un análisis de regresión logística multifactorial que incorporaba tanto variables relacionadas con las patentes como variables de control relevantes a nivel de firma y de ensayo. El objetivo principal era determinar si el panorama tecnológico y de propiedad intelectual asociado a un ensayo clínico podía ayudar a explicar su probabilidad de progresión. Los resultados revelan una asociación negativa consistente entre varios indicadores de conocimiento basados en patentes -a saber, el número de patentes relacionadas, el tamaño acumulativo de la familia de patentes, el alcance tecnológico (medido a través de códigos IPC) y el volumen de conocimiento previo citado- y la progresión del ensayo clínico. Aunque a menudo se considera que una amplia actividad de patentes es un indicador de sofisticación o valor tecnológico, esta investigación sugiere que también puede indicar retrasos estratégicos, una mayor complejidad jurídica o barreras de coordinación. En particular, la formación de marañas de patentes, redes densas de derechos superpuestos, se identificó como un impedimento potencial para el desarrollo, especialmente en áreas de alto riesgo como las tecnologías de ARNm. Además, el análisis sugiere que el amplio alcance tecnológico y la fuerte dependencia de los conocimientos re combinados, aunque beneficiosos en las primeras fases de la innovación, pueden obstaculizar la ejecución focalizada necesaria en las fases posteriores de desarrollo. Esto es coherente con la noción de que la amplitud excesiva puede socavar la coherencia estratégica y la alineación operativa, ambas cruciales para la navegación reguladora y la eficiencia de los ensayos.

En cambio, las variables a nivel institucional y de ensayo, como el número de colaboradores, la edad institucional y la intensidad de la I+D competitiva, muestran asociaciones positivas y estadísticamente significativas con el avance de los ensayos. Estos resultados ponen de relieve la importancia de la capacidad organizativa, la experiencia y las redes de colaboración para sortear las complejidades del desarrollo clínico. Las instituciones maduras, en particular, parecen estar mejor equipadas para gestionar los procesos normativos, coordinar los esfuerzos interdisciplinarios y satisfacer las exigencias operativas de las fases avanzadas de los ensayos.

Un resultado inesperado pero digno de mención es la experiencia de las empresas, que mostró una relación negativa pequeña pero estadísticamente significativa con la progresión de los ensayos. Esto puede reflejar el hecho de que las empresas con más experiencia suelen realizar un mayor número de ensayos exploratorios en fases tempranas, que intrínsecamente tienen menos probabilidades de progresar. En lugar de indicar una deficiencia en la capacidad, es probable que este resultado se deba a la naturaleza de los propios ensayos y al comportamiento estratégico de las empresas incluidas en el conjunto de datos.

A pesar de la importancia de estos resultados, el estudio tiene sus limitaciones. Aunque PAT-SPECTER es una herramienta de vanguardia basada en la arquitectura BERT y entrenada en citas científicas, no está específicamente entrenada en textos de ensayos clínicos. Como tal, aunque funciona bien a la hora de establecer relaciones semánticas y contextuales aproximadas, la puntuación F1 relativamente baja del modelo se traduce en una notable tasa de falsos positivos. Para solucionar este problema, se aplicaron filtros adicionales durante el procesamiento de los datos. Incluso con estos ajustes, las coincidencias de similitud identificadas deben interpretarse como aproximadas más que exactas, ya que reflejan una alineación temática general más que un solapamiento de contenido detallado. Otra limitación reside en la imposibilidad de analizar las reivindicaciones específicas (en inglés *claims*) protegidas en las patentes coincidentes. Las reivindicaciones representan los límites legalmente exigibles de la protección de patentes y pueden aportar información valiosa sobre la precisión con la que una patente apoya o se relaciona con un esfuerzo de desarrollo clínico. Sin embargo, debido a las restricciones técnicas de la interfaz de PATSTAT Online 2024, no fue posible realizar una extracción y un análisis a gran escala de las reivindicaciones de patentes. Esto representa una vía prometedora para futuras investigaciones.

En conclusión, si bien el acceso a un conocimiento tecnológico rico y diverso sigue siendo esencial para la innovación en el sector farmacéutico, este estudio subraya que el éxito del avance clínico depende igualmente de la eficacia con que dicho conocimiento se gestione, integre y ponga en funcionamiento dentro de marcos institucionales competentes y estratégicamente orientado.

Bibliografía

Adekola, T. A. (2023). *When mRNA Technology Meets Patent Law: Innovation, Barriers, and the Public Health*. TC Beirne School of Law, University of Queensland. SSRN

Advisory Committee on Measuring Innovation in the 21st Century Economy. (2008). *Innovation Measurement: Tracking the State of Innovation in the American Economy*. U.S. Department of Commerce.

Agarwal S., “Word to Vectors — Natural Language Processing”, 2017. <https://towardsdatascience.com/word-to-vectors-natural-language-processing-b253dd0b0817>

Ahmadpoor, M., & Jones, B. F. (2017). The dual frontier: Patented inventions and prior scientific advance. *Science*, 357(6351), 583-587

Algethami, J. S., El-Wahed, A. A. A., Elashal, M. H., Ahmed, H. R., Elshafiey, E. H., Omar, E. M., ... & El-Seedi, H. R. (2022). Bee pollen: Clinical trials and patent applications. *Nutrients*, 14(14), 2858.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *ACL*, 2020.

Arora, A., Belenzon, S., & Pataconi, A. (2018). *The decline of science in corporate R&D*. *Strategic Management Journal*, 39(1), 3–32

Arora, A., Belenzon, S., & Sheer, L. (2021). *Knowledge spillovers and corporate investment in scientific research*. *American Economic Review*, 111(3), 871–898

Arora, A., Cohen, W. M., & Walsh, J. P. (2016). The acquisition and commercialization of invention in American manufacturing: Incidence and impact. *Research Policy*, 45(6), 1113–1128

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Baker, W. E., Sinkula, J. M., Grinstein, A., & Rosenzweig, S. (2014). *The effect of radical innovation in/congruence on new product performance*. *Industrial Marketing Management*, 43(8), 1314–1323

Biotechnology Innovation Organization (BIO), Biomedtracker, & Amplion. (2021). *Clinical development success rates and contributing factors 2011–2020*

Breschi, S., Lissoni, F., & Malerba, F. (2003). Knowledge-relatedness in firm technological diversification. *Research Policy*, 32(1), 69–87

Brusoni, S., Criscuolo, P., & Geuna, A. (2005). The knowledge bases of the world's largest pharmaceutical groups: What do patent citations to non-patent literature reveal? *Economics of Innovation and New Technology*, 14(5), 395–415

Casadevall, A. (2021). *The mRNA vaccine revolution is the dividend from decades of basic science research*. *The Journal of Clinical Investigation*, 131(19), e153721

ClinicalTrials.gov. (2021). Clinical Trials: Structure and Design. Retrieved from <https://clinicaltrials.gov>

Dai, X., Karimi, S., Hachey, B., & Paris, C. (2019). Biomedical entity extraction and relation mapping

Daniel S. Hain, Roman Jurowetzki, Tobias Buchmann, Patrick Wolf, A text-embedding-based approach to measuring patent-to-patent technological similarity, *Technological Forecasting and Social Change*, Volume 177,2022,121559,ISSN 0040-1625, <https://doi.org/10.1016/j.techfore.2022.121559>. (<https://www.sciencedirect.com/science/article/pii/S0040162522000919>)

Dasgupta, P., David, P., 1994. Towards a new economics of science. *Research Policy* 23, 487–521

Daud A., Khan W. & Che D., “Urdu language processing: a survey.”, *Artificial Intelligence Review*, 1-33, DOI 10.1007/s10462-016-9482-x, 2016.

DrugPatentWatch. (2023). *Revolutionizing the fight against biologic drug patent thickets*.

E. Mankolli, V. Guliashki, "Machine Learning and Natural Language Processing: Review of Models and Optimization Problems", October 30, 2020. https://link.springer.com/chapter/10.1007/978-3-030-62098-1_7

European Patent Office (EPO). (2023). *Patent Knowledge and Technology Transfer*

Finnegan. (2021). *How clinical trials affect patentability in the US and Europe*.

Dong, Y., Ma, H., Tang, J., & Wang, K. (2018). Collaboration Diversity and Scientific Impact

Wegner, C. D., Mount, B. A., & Colvis, C. M. (2022). A public-private collaboration model for clinical innovation. *Clinical and Translational Science*, 15(7), 1581–1591.

Yang, G., Smith, J., Pérez, L., & Sartori, V. (2025). *Charting the path to patients: Optimizing drug pipelines*. McKinsey & Company

Li, M., Ren, J., Si, X., Sun, Z., Wang, P., Zhang, X., Liu, K., & Wei, B. (2022). *The global mRNA vaccine patent landscape*. *Human Vaccines & Immunotherapeutics*, 18(6), e2095837.

European Patent Office. (2023). *mRNA technologies: Insight report*. European Patent Office.

Goode, R., & Chao, B. (2022). Biological patent thickets and delayed access to biosimilars: An American problem. *Journal of Law and the Biosciences*

Glass, H. E., Glass, L. M., Tran, P., & Alghamdi, H. (2016). Pharmaceutical organizational size and phase 3 clinical trial completion times. *Therapeutic Innovation & Regulatory Science*, 50(6), 801–807.

Gadiya, R., Chundawat, N., Joshi, H., & Patel, H. (2023). *Pharmaceutical patent landscaping: A novel approach to understand patents from the drug discovery perspective*. *World Patent Information*, 74, 102192

Ganguli, I., Lin, J., Meursault, V., & Reynolds, N. F. (2024). *Patent text and long-run innovation dynamics: The critical role of model selection* (NBER Working Paper No. 32934). National Bureau of Economic Research. <https://www.nber.org/papers/w32934>

Garzón-Vico, A., Rosier, J., Gibbons, P., & McNamara, P. (2020). The Impact of Failure and Success Experience on Drug Development. *Journal of Product Innovation Management*, 37(1), 74–96

Ghosh, M., Erhardt, S., Rose, M.E., Buunk, E., & Harhoff, D. (2024). PaECTER: Patent-level Representation Learning using Citation-informed Transformers. *ArXiv, abs/2402.19411*.

Haas, M. R., & Ham, W. (2015). Microfoundations of knowledge recombination: Peripheral knowledge and breakthrough innovation in teams. *Advances in Strategic Management*, 32, 47–87.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. pages 3613–3618. Association for Computational Linguistics, 2019.

J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: <http://aclweb.org/anthology/N19-1423>. doi:10.18653/v1/N19-1423.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018

Jokinen, O. (2024). *Document-level embeddings and graph-augmented learning for Finnish articles* (Master's thesis). University of Helsinki

K. Sparck Jones, A Statistical Interpretation of Term Specificity and its Application in Retrieval, *Journal of Documentation* 28 (1972) 11–21. URL: <https://doi.org/10.1108/eb026526>. doi:10.1108/eb026526, publisher: MCB UP Ltd.

Kaplan, S., & Vakili, K. (2015). The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal*, 36(10), 1435-1457

Khan W., Daud A., Nasir J. A., Amjad T, “A survey on the state-of-the-art machine learning models in the context of NLP”, *Kuwait Journal of Science* 43 (4),pp.95-113, 2016.

Khanna, R. (2015). *Learning from small failures: Role of relatedness, familiarity, and structure of knowledge base* (Doctoral dissertation, University of North Carolina at Chapel Hill).

Kirchhoff, B. A. (1991). *Entrepreneurship's contribution to economics. Entrepreneurship Theory and Practice*, 16(2), 93-112

Leten, B., Kelchtermans, S., & Belderbos, R. (2022). How does basic research improve innovation performance in the world's major pharmaceutical firms? *Industry and Innovation*, 29(3), 396-424

Magazzini, L., Brusoni, S., & Franco, M. (2012). Learning from failures or failing to learn? Lessons from pharmaceutical R&D. *European Management Review*, 9(2), 117-137

Mahan, V. L. (2014). Clinical trial phases. *International Journal of Clinical Medicine*, 5(12), 1374-1383. <https://doi.org/10.4236/ijcm.2014.521175>

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Bibliography 51 Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Meyer M. Are patenting scientists the better scholars? An exploratory comparison of inventor-authors with their non-inventing peers in nano-science and technology[J]. *Research Policy*, 2006, 35(10): 1646-1662

Motohashi, K., Koshiba, H. & Ikeuchi, K. Measuring science and innovation linkage using text mining of research papers and patent information. *Scientometrics* **129**, 2159–2179 (2024). <https://doi.org/10.1007/s11192-024-04949-w>

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. The MIT Press. ISBN: 978-0-262-01802-9.

National Heart, Lung, and Blood Institute (NHLBI). (2021). *How clinical trials work*. Retrieved from <https://www.nhlbi.nih.gov/research/clinical-trials/how-studies-work>

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019

Noh, J., & Kavuluru, R. (2021). *Improved biomedical word embeddings in the transformer era*. *Journal of Biomedical Informatics*, 120, 103867. <https://doi.org/10.1016/j.jbi.2021.103867>

OECD & World Bank. (2009). *Innovation and growth: Chasing a moving frontier*. OECD

OECD. (2015). *The innovation imperative: Contributing to productivity, growth and well-being*.

Poege, F., Harhoff, D., Gaessler, F., & Baruffaldi, S. (2019). *Science quality and the value of inventions*. *Science Advances*, 5(12), eaay7323

ROBERT HECHT-NIELSEN. Iii.3 - theory of the backpropagation neural network**based on ânonindentâ by robert hecht-nielsen, which appeared in proceedings of the international joint conference on neural networks 1, 593â611,june1989.Â© 1989 ieee. In Harry Wechsler, editor, *Neural Networks for Perception*, pages 65–93. Academic Press, 1992

Romer, P. M. (1989). *Endogenous technological change* (No. w3210). National Bureau of Economic Research.

Roper, S., & Hewitt-Dundas, N. (2015). Knowledge stocks, knowledge flows and innovation: Evidence from matched patents and innovation panel data. *Research Policy*, 44(7), 1327-1340

Rose, M. E., Buunk, E., Erhardt, S., Li, C., Ghosh, M., & Harhoff, D. (2024). *Tracing the Flow of Knowledge from Science to Technology Using Deep Learning: Pat-SPECTER Application for Threshold*. Max Planck Institute for Innovation and Competition.

Sampat, B. N. (2016). Patent citations—An analysis of quality differences and citing practices. *Management Science*, 62(2), 513–531

Schumpeter, J.A.: The theory of economic development : an inquiry into profits, capital, credit, interest, and the business cycle. Harv. Econ. Stud. (1934)

Sebastian Erhardt et al. (2024). Logic Mill - A Knowledge Navigation System. InPatentSemTech@SIGIR (pp. 25-35). <https://doi.org/10.48550/arXiv.2301.00200>

Sharma, P., & Tripathi, R. C. (2017). Patent citation: A technique for measuring the knowledge flow of information and innovation. *World Patent Information*, 51, 31-42.

Shuo Xu, Zhen Liu, Xin An, Hong Wang, Hongshen Pang, Linkages among science, technology, and industry on the basis of main path analysis, *Journal of Informetrics*, Volume 19, Issue 1, 2025, 101617, ISSN 1751-1577, <https://doi.org/10.1016/j.joi.2024.101617>

Singh, N., Vayer, P., Tanwar, S., Poyet, J.-L., Tsaioun, K., & Villoutreix, B. O. (2023). Drug discovery and development: Introduction to the general public and patient groups. *Frontiers in Drug Discovery*, 3, 1201419

Sorenson, O., & Fleming, L. (2004). Science and the diffusion of knowledge. *Research Policy*, 33(10), 1615–1634

T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, 2013. URL: <http://arxiv.org/abs/1301.3781>.doi:10.48550/arXiv.1301.3781,arXiv:1301.3781 [cs]

U.S. Food and Drug Administration (FDA). (2022). Overview of Clinical Trials. Retrieved from <https://www.fda.gov>

Wagner, S., & Wakeman, S. (2016). *What do patent-based measures tell us about product commercialization? Evidence from the pharmaceutical industry*. *Research Policy*, 45(6), 1091-1102

Wang, J., Verberne, S., Abbasiantaeb, Z., & Zhang, J. (2023). *Linking patents to scientific publications through in-text reference mining*. Leiden University. EPO ARP Grant 2020.

Wang, T. (2023). Toward an understanding of innovation failure: The timing of failure experience. *Technovation*, 125, 102787

Weitzman, M. L. (1998). Recombinant growth. *The Quarterly Journal of Economics*, 113(2), 331-360

World Intellectual Property Organization. (2023). COVID-19 vaccines and therapeutics: Insights into related patenting activity throughout the pandemic. WIPO.

Xiao, T., Makhija, M., & Karim, S. (2022). *A knowledge recombination perspective of innovation: Review and new research directions*. *Journal of Management* (Forthcoming). Retrieved from <https://ssrn.com/abstract=3912331>

Xie, Z., Wu, K., Du, J. T., Xie, Y., & Chen, Y. (2025). *How knowledge recombination fuels technological innovation? Insights from IPC co-occurrence networks*. In G. Oliver et al. (Eds.), *ICADL 2024, Lecture Notes in Computer Science* (Vol. 15494, pp. 19–38). Springer Nature Singapore. https://doi.org/10.1007/978-981-96-0868-3_2

Zhang, Y., Wang, D., & Xu, L. (2021). *Knowledge search, knowledge integration and enterprise breakthrough innovation under the characteristics of innovation ecosystem network: The empirical evidence from enterprises in Beijing-Tianjin-Hebei region*. *PLOS ONE*, 16(12), e0261558

Zhong, C., Huang, R., Duan, Y., Sunguo, T., & Dello Strologo, A. (2024). Exploring the impacts of knowledge recombination on firms' breakthrough innovation: The moderating effect of environmental dynamism. *Journal of Knowledge Management*, 28(3), 698-723

A Apéndice

A.1. Regresión logística con variables base

Logistic Regression: Determinants of Phase				
	Dependent variable:			
	Phase (Logit)			
	Model 1	Model 2	Model 3	Model 4
	(1)	(2)	(3)	(4)
# Collaborators	0.140*** (0.030)	0.140*** (0.030)	0.140*** (0.030)	0.141*** (0.030)
Institution experience	-0.0001*** (0.00003)	-0.0001*** (0.00003)	-0.0001*** (0.00003)	-0.0001*** (0.00003)
Institution age	0.003 (0.005)	0.003 (0.005)	0.003 (0.005)	0.003 (0.005)
R&D Competition	0.001*** (0.0002)	0.001*** (0.0002)	0.001*** (0.0002)	0.001*** (0.0002)
# Related patents	-0.0004* (0.0002)			
Cumulative family size		-0.00005* (0.00002)		
Cumulative number of IPCs			-0.0001** (0.00003)	
Cumulative number of cited documents				-0.00001 (0.00001)
Constant	-1.342*** (0.200)	-1.341*** (0.199)	-1.312*** (0.200)	-1.424*** (0.196)
Observations	1,601	1,601	1,601	1,601
Log Likelihood	-992.958	-992.849	-992.266	-994.161
Akaike Inf. Crit.	1,997.915	1,997.698	1,996.533	2,000.323
Note:			* p<0.1; ** p<0.05; *** p<0.01	

Tabla A.1: Regresión logística con variables sin transformar

A.2. Regresión logística con variables independientes log-transformadas

Logistic Regression: Determinants of Phase				
	<i>Dependent variable:</i>			
	Phase (Logit)			
	Model 1	Model 2	Model 3	Model 4
	(1)	(2)	(3)	(4)
# Collaborators	0.140*** (0.030)	0.139*** (0.030)	0.140*** (0.031)	0.139*** (0.030)
Institution experience	-0.0001*** (0.00003)	-0.0001*** (0.00003)	-0.0001*** (0.00003)	-0.0001*** (0.00003)
Institution age	0.002 (0.005)	0.002 (0.005)	0.002 (0.005)	0.002 (0.005)
R&D Competition	0.001*** (0.0002)	0.001*** (0.0002)	0.001*** (0.0002)	0.001*** (0.0002)
# Related patents	-0.126*** (0.035)			
log.Cumulative family size		-0.125*** (0.034)		
log.Cumulative IPC codes			-0.134*** (0.036)	
log.Cumulative cited documents				-0.103*** (0.029)
Constant	-0.806*** (0.265)	-0.547* (0.317)	-0.484 (0.324)	-0.667** (0.293)
Start Year	No	No	No	No
Sponsor Type	No	No	No	No
Observations	1,601	1,601	1,601	1,601
Log Likelihood	-988.372	-988.081	-987.671	-988.271
Akaike Inf. Crit.	1,988.744	1,988.162	1,987.343	1,988.541

Note:

* p<0.1; ** p<0.05; *** p<0.01

Tabla A.2: Regresión logística con variables independientes transformadas logarítmicamente

A.3. Regresión logística con variables de control categóricas activas

Logistic Regression: Determinants of Phase				
	<i>Dependent variable:</i>			
	Phase (Logit)			
	Model 1	Model 2	Model 3	Model 4
	(1)	(2)	(3)	(4)
# Collaborators	0.164*** (0.033)	0.164*** (0.033)	0.164*** (0.033)	0.164*** (0.033)
Institution experience	-0.0001*** (0.00003)	-0.0001*** (0.00003)	-0.0001*** (0.00003)	-0.0001*** (0.00003)
log.Institution age	0.330*** (0.100)	0.329*** (0.100)	0.330*** (0.100)	0.327*** (0.100)
R&D Competition	0.001 (0.001)	0.001 (0.001)	0.001 (0.001)	0.001 (0.001)
# Related patents	-0.126*** (0.037)			
log.Cumulative family size		-0.125*** (0.036)		
log.Cumulative IPC codes			-0.131*** (0.037)	
log.Cumulative cited documents				-0.112*** (0.030)
Constant	14.525 (1,455.398)	14.747 (1,455.398)	14.596 (1,455.398)	14.727 (1,455.398)
Start Year	Yes	Yes	Yes	Yes
Sponsor Type	Yes	Yes	Yes	Yes
Observations	1,601	1,601	1,601	1,601
Log Likelihood	-959.342	-959.171	-959.023	-958.368
Akaike Inf. Crit.	2,000.684	2,000.342	2,000.045	1,998.735
<i>Note:</i>			* p<0.1; ** p<0.05; *** p<0.01	

Tabla A.3: Regresión logística con variable de control categórica activa

Lista de figuras

Figura 1: Modelo de flujos de conocimiento (Fuente: Sorenson & Fleming, 2004).....	16
Figura 2: Uso de la ciencia en la innovación: edad media de la ciencia citada en patentes y media de citas por patente, 1980-2006. (Fuente: Arora et al., 2018).....	17
Figura 3: Publicación, patentes e investigación: Empresas estadounidenses, 1980-2006. (Fuente: Arora et al., 2018)	17
Figura 4: Tendencias en el uso de la ciencia por las patentes corporativas, 1985-2015. (Fuente: Arora et al., 2021).....	19
Figura 5: Curva de distribución año-patente en el campo farmacéutico (Fuente: Xie et al., 2025)	22
Figura 6: Distribución de los innovadores diversificados según el número de campos tecnológicos en los que solicitaron patentes (valores porcentuales, 1982-1993) (Fuente: Breschi et al., 2003)	23
Figura 7: Distribución del total de patentes solicitadas por innovadores diversificados según el número de campos tecnológicos en los que posteriormente solicitaron patentes (valores porcentuales, 1982-1993) (Fuente: Breschi et al., 2003).	23
Figura 8: Sinopsis del análisis del panorama de las patentes. (Fuente: Gadiya et al. (2023)	27
Figura 9: Tendencia temporal de la investigación básica interna y del aprovechamiento de la investigación básica interna y externa (Fuente: Leten et al. 2021)	29
Figura 10: Cuatro formas en que los tres tipos de literatura académica se relacionan entre sí (Fuente: Xu et al. 2025).....	31
Figura 11: Descubrimiento y desarrollo de fármacos (Fuente: Singh et al., 2023)	33
Figura 12: Nueva arquitectura del modelo. La arquitectura CBOW predice la palabra actual basándose en el contexto, y la Skip-gram predice las palabras circundantes a partir de la palabra actual (Fuente: Mikolov et al., 2013).	39
Figura 13: La arquitectura transformer (Fuente: Vaswani et al., 2017).....	43
Figura 14: Arquitectura de entrenamiento de SPECTER (Fuente: Cohan et al., 2020)	47
Figura 15: Visión general de la arquitectura de Logic Mill (Fuente: Erhardt et al., 2024)	50

Figura 16: Carga del modelo lingüístico y del tokenizador	60
Figura 17: Función para generar incrustaciones para cada documento con PAT-SPECTER.....	61
Figura 18: Función para controlar la muestra de datos y la conversión global a incrustación	62
Figura 19: Función para añadir la representación vectorial de cada documento al conjunto de datos y creación del conjunto de datos.....	62
Figura 20: Función para el cálculo de la similitud y la exportación del conjunto de datos	63
Figura 21: Fragmento de consulta para la búsqueda de patentes en PATSTAT online	64
Figura 22: Distribución de las puntuaciones de similitud del coseno para vínculos verdaderos de ensayos clínicos y patentes	65
Figura 23: Distribución de las puntuaciones de similitud del coseno para vínculos de ensayos clínicos y patentes no verdaderos	66
Figura 24: Histogramas de variables sesgadas.....	74

Lista de tablas

Tabla 1: Número de productos por etapas del proceso de desarrollo de productos (Fuente: Wagner & Wakeman, 2016)	20
Tabla 2: Códigos IPC de vacunas de ARNm y sus campos (Fuentes: Li et al., 2022 y EPO, 2023).....	58
Tabla 3: Códigos CPC de vacunas de ARNm y sus campos (Fuente: Oficina Europea de Patentes, 2023)	59
Tabla 4: Estadísticas descriptivas de las variables independientes y los controles	72
Tabla 5: Frecuencias de las diferentes categorías de promotor principal del ensayo .	73
Tabla 6: Tabla de correlaciones de las variables independientes y los controles.....	73
Tabla 7: Resultados del modelo de regresión logística	78

Agradecimientos

Esta tesis representa el final de un ciclo de aprendizaje, adaptación y construcción. Un periodo lleno de experiencias con personas hermosas que me aportaron crecimiento personal, confianza y un futuro lleno de posibilidades.

En primer lugar, quiero agradecer a la profesora Daniela Silvestri por su orientación, constancia, por brindarme siempre reflexiones sobre los puntos de vista de la investigación, y ayudarme a organizar y encuadrar de la mejor manera posible el trabajo de investigación, un tipo de trabajo académico con el que no estaba relacionado. La mejor asesora que podría pedir.

Igualmente importante fue el apoyo del profesor Daniel De Souza, que me apoyó en particular, pero también los hitos más relevantes de este trabajo. Gracias por la paciencia y la amabilidad.

Quisiera expresar mi más sincero agradecimiento al profesor Pablo Viveros por su constante apoyo y gestión durante mi estancia en el extranjero. Su compromiso y disposición fueron fundamentales para facilitar todos los trámites relacionados con el programa de doble título, así como con la organización y desarrollo de mi defensa de tesis posterior en la Universidad Técnica Federico Santa María.

Asimismo, agradezco de manera especial a todos los profesores de la USM que formaron parte de este proceso y contribuyeron con su orientación y conocimientos en las distintas etapas de mi formación académica.

Quiero agradecer fuertemente y con todo mi corazón a toda mi familia, incluyendo abuelos, tíos y primos.

A mi padre, por compartir siempre su cariño, opiniones, consejos y esas palabras de orgullo que siempre me han hecho creer fuertemente en mí mismo, mantener la tranquilidad y siempre poner el pecho a las balas para superar todo con valentía.

A mi madre, que igualmente siempre ha estado para mí y me ha apoyado a lograr todos los planes que me he propuesto, su inmenso cariño y amor que me han hecho sentir siempre en casa, y por su gran preocupación en todo aspecto.

Y juntos ambos padres me han demostrado que no importa donde este, siempre estarán conmigo y siempre lo estarán para todo.

A mi hermano, por confiar en mí, por todas esas charlas ocasionales, por hacerme sentir orgulloso de quien soy y por todo lo que compartimos, y seguiremos

compartiendo en nuestro futuro que siempre estará relacionado. Siempre estaré para ti.

A mi tío Walter, por mostrarme el camino de la ingeniería y aconsejarme en los nuevos pasos que conviene tomar para formar mi carrera y mi vida. Siempre ha sido un ejemplo de libertad, responsabilidad y futuro para mí.

A mi tío Pablo, por sus innumerables consejos y apoyo en varias etapas de mi vida, siempre presente tanto en el ámbito académico como recreativo.

Ringrazio di cuore Martina, con la quale ho condiviso un'infinità di momenti in quest'ultimo anno, imparando tanto l'uno dall'altro. Hemos superado muchas aventuras y atravesado períodos de transición. Hemos viajado y vivido todo lo que nos hace felices: la comida, el cine, el globo y las salidas con los amigos, siempre juntos. Le doy valor a todo lo que tiene, no puedo pedir más. Ma soprattutto, abbiamo costruito un amore sincero e vero, e sono certi che insieme supereremo sempre ogni difficoltà: *"si se quiere se puede"*. Grazie per tutto l'amore, i consigli, l'organizzazione, per avermi fatto sentire parte della tua famiglia e, soprattutto, per farmi sentire a casa ogni volta che sono con te.

Me gustaría dar las gracias a otro pilar importante durante este periodo: mis íntimos amigos internacionales Ehab, Hilal y Enzo, con los que completamos juntos el máster.

Ehab, por estar siempre a mi lado, compartiendo preocupaciones, pensamientos y consejos de moda. Por ser un amigo con el que siempre puedo hablar.

Hilal, por ser siempre una persona de confianza, colaborar académicamente, ofrecer consejos de vida y compartir los últimos cotilleos.

Y a Enzo, por las innumerables aventuras que hemos vivido, por ser el mejor compañero de viaje, por todas las conversaciones profundas y por el trabajo que hemos realizado juntos.

Por último, a mis amigos de toda la vida de Chile y a los amigos que he conocido por el camino. Gracias por tanta amabilidad.

Este capítulo de mi vida ha llegado a su fin, y estoy listo para comenzar uno nuevo que me dirigirá hacia nuevos desafíos y nuevas metas.

Gracias a todos.

Patricio

