Análisis de correlación de representaciones basadas en aprendizaje automático de características nucleares en imágenes histopatológicas con su expresión génica espacial

Jorge Daniel Robles Ardila y Deciré Dayana Jaimes Rodríguez

Trabajo de grado presentado como requisito parcial para optar al título de Ingeniero de Sistemas

Director
David Romo Bucheli
Doctor en Ingeniería – Ingeniería Eléctrica

Codirectora
Nydia Paola Rondón Villarreal
Doctora en Ingeniería – Ingeniería Electrónica

Universidad Industrial de Santander
Facultad de Ingenierías Fisicomecánicas
Escuela de Ingeniería de Sistemas e Informática
Bucaramanga
2026

Dedicatoria

A mi madre, Sonia Ardila, por haberme criado con amor, paciencia y los mejores valores, enseñándome a ser responsable, disciplinado y respetuoso. Gracias por tu apoyo incondicional, por tu entrega constante y por cuidar de nosotros siempre con una preocupación genuina y desinteresada. Todo lo que soy hoy es en gran parte gracias a ti.

A mi padre, Jorge Robles, por ser un ejemplo de fortaleza y superación. Por enseñarme a esforzarme por lo que deseo y a no rendirme ante las dificultades. Gracias por velar siempre por nosotros y por trabajar incansablemente para que nunca nos faltara nada, brindándonos una vida digna.

A mi hermana, Brigitte Robles, por acompañarme desde siempre con su cariño y apoyo. Por haber cuidado de mí en tantos momentos y estar presente cuando más lo necesité.

A mi sobrino, Christopher Torres, quien, aunque apenas comienza su camino, llena de alegría a nuestra familia. Le deseo una vida llena de amor, oportunidades y sueños cumplidos, y espero que crezca rodeado de los valores que nos han formado.

A mi mascota, Milú, por su compañía incondicional a lo largo de su vida. Por cada momento de alegría, por su nobleza y por haber sido una fuente de tranquilidad, motivación y fortaleza en este proceso. Su presencia hizo este camino más llevadero.

Jorge Daniel

A mi madre, Liliana, mi motivación día a día para seguir.

A mi padre, Arturo, cuyo amor no entiende de distancias.

A mis abuelos, Dora y Luis, mis segundos padres;
y a Jose y Josefina, quienes me cuidan desde las estrellas.

A mis duquesas, Angie y Hayley, quienes entienden mi mundo y me invitan al suyo.

A mi mascota, Zeus, por cada noche de desvelo que estuvo a mi lado.

A mi pareja, Daniel, la chispa que ilumina mi camino y mi refugio en los días oscuros.

Deciré

Agradecimientos

A Dios, por las bendiciones recibidas, por guiar mi camino y por darme la fortaleza necesaria para alcanzar esta meta.

A mi familia, por su apoyo constante y por estar siempre presentes. Gracias por enseñarme el valor de la unión, la solidaridad y el compromiso. Saber que cuento con ustedes ha sido fundamental en cada etapa de este proceso.

A la Universidad Industrial de Santander, por abrirme sus puertas y brindarme una formación académica y personal que marcará mi vida. Por cada experiencia, aprendizaje y oportunidad que contribuyó a mi crecimiento.

A mi director, David Romo, por su acompañamiento, disposición y orientación a lo largo de este trabajo. Gracias por su actitud siempre positiva, por su apoyo constante y por brindarme el espacio para desarrollar mis ideas de la mejor manera.

A mis amigos, tanto los de toda la vida como aquellos que conocí durante la carrera, por su compañía y apoyo en este camino. Gracias por cada momento compartido, por las risas, por el ánimo en los momentos difíciles y por motivarme a seguir adelante. Las amistades que formé en esta etapa son, sin duda, de las más valiosas que pude haber construido.

Jorge Daniel

Agradezco primero a mi familia, por acompañarme incondicionalmente en este difícil camino de cinco años, sin su apoyo nada de esto sería real. También a mis amigos, los que siempre supieron sacarme una sonrisa incluso en mis peores momentos. Gracias también a todas esas personas que no puedo mencionar aquí, pero que formaron parte de mi transcurso por la universidad y me acompañaron de alguna forma.

A la Universidad Industrial de Santander por haberme formado y acogido durante estos años que marcaron mi vida. De igual manera, quiero agradecer a mis directores: a David, por ser esa guía con su conocimiento durante todo el desarrollo de este trabajo; y a Paola, por ser un apoyo tan importante y por esa alegría que siempre me ayudó a seguir adelante. También a mi compañero, Robles, por continuar conmigo a pesar de todas las dificultades que enfrentamos en el camino.

Finalmente, gracias a mí. Gracias por creer siempre en lo que soy capaz de hacer y por tener la fuerza para continuar sin importar qué tan difíciles se pusieron las cosas.

Deciré

Índice general

Índice de figuras

Índice de tablas

Lista de apéndices

Resumen

Título: Análisis de correlación de representaciones basadas en aprendizaje automático de características nucleares en imágenes histopatológicas con su expresión génica espacial.

Autores: Jorge Daniel Robles Ardila y Deciré Dayana Jaimes Rodríguez

Palabras clave: Aprendizaje profundo, imágenes histopatológicas, transcriptómica espacial, expresión génica, cáncer.

Descripción: El análisis de imágenes histopatológicas ofrece una visión detallada de la estructura celular; por su parte, la transcriptómica espacial permite mapear la actividad génica conservando la arquitectura física del tejido. Relacionar ambas modalidades es fundamental para comprender la heterogeneidad tumoral en el cáncer. Sin embargo, a pesar de los avances del aprendizaje profundo en patología digital, continúa el desafío de identificar qué representaciones visuales de los núcleos celulares reflejan con mayor fidelidad estos patrones de expresión molecular.

Para abordar este problema, el presente trabajo analiza cómo se correlacionan diversas representaciones de núcleos celulares con la expresión génica espacial, mediante muestras de tejido renal sano y canceroso del conjunto de datos HEST-1K. El estudio evalúa descriptores morfológicos tradicionales, grafos de interacción celular y representaciones latentes (*embeddings*) de modelos fundacionales como Virchow y UNI. A nivel metodológico, el proceso abarcó la mitigación de efectos de lote, la extracción de características multiescala y la construcción de un modelo de integración multimodal que fusiona la morfología visual con la topología celular.

Los resultados demuestran que los *embeddings* extraídos de modelos fundacionales capturan de manera efectiva la asociación con perfiles génicos específicos. Aunque las representaciones basadas en grafos mostraron una capacidad predictiva limitada al evaluarse de forma aislada, su verdadero valor radica en la información estructural complementaria que aportan. Al integrar ambas modalidades, se supera de forma consistente a las representaciones individuales, alcanzando incrementos promedio en el coeficiente de determinación (R^2) de 0.010 en UNI, 0.009 en UNI2-h, 0.012 en Virchow y 0.009 en Virchow2. De manera complementaria, en este trabajo se plantea un método para la identificación de genes con dos características de interés: una mejor asociación con características morfológicas y mayor independencia con la información que pueda proveer una etiqueta (tejido tumoral o sano).

---

* Trabajo de Grado de Pregrado

** Facultad de Ingenierías Fisicomecánicas. Escuela de Ingeniería de Sistemas e Informática.

Director: David Romo Bucheli, PhD.

Abstract

Title: Correlation analysis of machine learning-based representations of nuclear features in histopathological images with their spatial gene expression.

Authors: Jorge Daniel Robles Ardila and Deciré Dayana Jaimes Rodríguez

Keywords: Deep learning, histopathological imaging, spatial transcriptomics, gene expression, cancer.

Description: The analysis of histopathological images provides a detailed view of cellular structure; in turn, spatial transcriptomics enables the mapping of gene activity while preserving the physical architecture of the tissue. Relating these two modalities is essential for understanding tumor heterogeneity in cancer. However, despite advances in deep learning applied to digital pathology, identifying which visual representations of cell nuclei most faithfully reflect these molecular expression patterns remains an open challenge.

To address this problem, this work analyzes the correlation between different representations of cell nuclei and spatial gene expression, using samples of healthy and cancerous renal tissue from the HEST-1K dataset. The study evaluates traditional morphological descriptors, cell interaction graphs, and latent representations (embeddings) derived from foundation models such as Virchow and UNI. Methodologically, the approach includes batch effect correction, multiscale feature extraction, and the development of a multimodal integration model that combines visual morphology with cellular topology.

The results show that embeddings extracted from foundation models effectively capture associations with specific gene expression profiles. While graph-based representations exhibit limited predictive performance when evaluated in isolation, they provide complementary structural information. By integrating both modalities, the proposed approach consistently outperforms individual representations, achieving average improvements in the coefficient of determination (R^2) of 0.010 for UNI, 0.009 for UNI2-h, 0.012 for Virchow, and 0.009 for Virchow2. Additionally, this work proposes a method for identifying genes of interest based on two criteria: stronger association with morphological features and greater independence from the information provided by the label (tumor or healthy tissue).

---

* Bachelor's Thesis

** Faculty of Physical-Mechanical Engineering. School of Systems Engineering and Informatics. Thesis Advisor: David Romo Bucheli, PhD.

Introducción

El cáncer sigue siendo un desafío de salud global apremiante. Las estimaciones de la *American Cancer Society* para 2022 reportaron cerca de 20 millones de diagnósticos nuevos y casi 10 millones de decesos (Bray et al. 2024). Las proyecciones hacia 2050 advierten una escalada a 35 millones de casos anuales. Ante esta trayectoria, la optimización clínica no es suficiente; resulta imperativo profundizar en nuestra comprensión biológica sobre el desarrollo de la enfermedad para diseñar estrategias de intervención más efectivas.

La evaluación de tejidos histopatológicos mediante microscopía ha operado históricamente como el estándar de oro en patología (Unger y Kather 2024). Hoy en día, la digitalización de biopsias expande estas capacidades hacia el terreno del análisis computacional. El aprendizaje profundo (*deep learning*) ha demostrado una particular eficacia en este frente, automatizando tareas complejas de segmentación y clasificación con una precisión en ocasiones comparable al criterio experto (Andrew H. Song et al. 2023). A pesar de estos hitos técnicos, la adopción clínica de dichas herramientas enfrenta una barrera crítica: la falta de interpretabilidad. Los modelos operan frecuentemente sin arrojar certidumbre biológica sobre sus decisiones. Por consiguiente, la investigación reciente busca anclar estas abstracciones computacionales a la realidad molecular del tejido.

la transcriptómica, definida como el estudio de la expresión génica a nivel global en una muestra biológica, ha permitido caracterizar la actividad molecular de los tejidos. Sin embargo, los enfoques tradicionales de esta área suelen perder la información espacial en el análisis de muestras de forma agregada. En este contexto, la transcriptómica espacial destaca como una tecnología verdaderamente disruptiva, cuya principal ventaja radica en medir la expresión génica preservando la arquitectura física del tejido (Ståhl et al. 2016). De esta forma, es posible cruzar lo que se observa en la imagen histológica con lo que se expresa a nivel génico, obteniendo un mapa espacial de la enfermedad. Esta integración bidireccional es clave para desentrañar la heterogeneidad de los tumores y comprender, a un nivel mucho más profundo, los procesos biológicos subyacentes.

Bajo este contexto, en este trabajo se pretende examinar de qué manera las representaciones obtenidas de imágenes histopatológicas, utilizando distintas técnicas, se correlaciona con los perfiles de expresión génica que se obtienen mediante la transcriptómica espacial en muestras de tejido canceroso. El objetivo es contribuir al avance en la integración entre morfología y transcriptómica, aportando evidencia y metodologías que impulsen el desarrollo de sistemas de apoyo al diagnóstico y la investigación en oncología computacional.

El presente documento cuenta con su información organizada en seis capítulos fundamentales. Inicialmente, el capítulo 1 presenta el objetivo general del trabajo junto con los objetivos específicos, los cuales orientan el desarrollo metodológico y experimental de la

investigación. A continuación, el capítulo 2 expone el marco referencial, en el que se abordan los principales conceptos teóricos necesarios para la comprensión del problema, incluyendo aspectos como cáncer renal, patología digital, modelos fundacionales, características nucleares, así como fundamentos de transcriptómica y transcriptómica espacial, y su integración en el análisis histológico. Posteriormente, el capítulo 3, describe el procedimiento seguido durante la realización del trabajo, detallando el conjunto de datos utilizado, los procesos de selección y curación de genes, así como el desarrollo de las distintas representaciones morfológicas, presentando el esquema de integración entre representaciones estructurales y visuales. Seguidamente, el capítulo 4 presenta el análisis de correlación entre las representaciones morfológicas y la expresión génica, así como la evaluación mediante modelos de regresión, comparando distintos enfoques de representación junto con su capacidad predictiva, incluyendo una comparación entre tejido sano y tejido canceroso. El capítulo 5 recoge los principales hallazgos derivados de la experimentación, destacando las implicaciones de la integración multimodal en la modelación del microambiente tisular, y presenta las posibles líneas de trabajo futuro, orientadas a la mejora del enfoque propuesto y su extensión a nuevos escenarios y tecnologías.

Planteamiento y Justificación del Problema

A pesar de los notables avances en patología computacional y transcriptómica espacial, la relación exacta entre las representaciones morfológicas y los perfiles de expresión génica sigue sin caracterizarse de manera sistemática (Shulman et al. 2024; Andrew H. Song et al. 2023). A la fecha, la literatura carece de un marco experimental comparativo que determine qué enfoque, ya sean características nucleares tradicionales, grafos de interacción celular o *embeddings* de modelos fundacionales, captura con mayor fidelidad esta información molecular.

A esta brecha se suma un desafío analítico crítico inherente a la transcriptómica espacial: la maldición de la dimensionalidad. Frecuentemente, el espacio de características supera los 30.000 genes distribuidos en un número muy reducido de muestras (J. Wang 2024). Este severo desequilibrio propicia la aparición de correlaciones espurias que no responden a vínculos biológicos reales, lo que dificulta la identificación de asociaciones estadísticamente sólidas (Gu et al. 2021).

Ante la necesidad de aislar señales biológicas válidas del ruido estadístico, surge la siguiente pregunta de investigación: *¿De qué manera podemos identificar conjuntos de genes que tengan correlaciones significativas con las características morfológicas y nucleares obtenidas de imágenes histopatológicas y la expresión génica espacial, considerando el reto de la alta dimensionalidad y baja cantidad de casos disponibles con datos transcriptómicos?*

Este trabajo se justifica en la necesidad de proponer y evaluar métodos que integren de manera efectiva la información morfológica con la transcriptómica. El objetivo es generar evidencia empírica sobre cómo se relacionan ambas fuentes de datos y, al mismo tiempo, sentar las bases para herramientas computacionales que faciliten la integración multimodal en el análisis de tejidos, contribuyendo así a la investigación en oncología computacional y potencialmente, a futuras aplicaciones clínicas.

1. Objetivos

1.1. Objetivo General

Desarrollar representaciones morfológicas y latentes de núcleos celulares basadas en modelos de aprendizaje automático, y analizar estadísticamente su correlación con la expresión génica observada vía transcriptómica espacial para caracterizar la heterogeneidad del cáncer en imágenes histopatológicas.

1.2. Objetivos Específicos

- Seleccionar información de una base de datos con imágenes histopatológicas e información asociada de transcriptómica espacial.
- Extraer características cuantitativas y morfológicas de los núcleos en las imágenes histopatológicas mediante técnicas de procesamiento digital de imágenes.
- Representar las imágenes histopatológicas a través de modelos de aprendizaje automático para la caracterización de núcleos.
- Modelar la correlación entre las características nucleares y la expresión génica observada vía transcriptómica espacial.
- Comparar los modelos de correlación en términos de su capacidad predictiva de la expresión génica.

2. Marco Referencial

En este capítulo se presentan y contextualizan los conceptos fundamentales para el desarrollo del trabajo. En primer lugar, la sección 2.1 introduce el problema central, el cáncer renal. A continuación, la sección 2.2 aborda la patología digital, su funcionamiento y su relevancia. Posteriormente, en la sección 2.3, se describen los modelos fundacionales, junto con sus bases y su papel en este campo. La sección 2.4 presenta las características nucleares desde dos enfoques: en la subsección 2.4.1 se abordan las tradicionales, mientras que en la subsección 2.4.2 se analizan desde una perspectiva basada en grafos. Luego, la sección 2.5 introduce la transcriptómica, iniciando con sus conceptos básicos en la subsección 2.5.1 y continuando con la transcriptómica espacial en la subsección 2.5.2. Finalmente, la sección 2.6 explora la integración entre imágenes histopatológicas y datos transcriptómicos, destacando cómo esta combinación permite analizar la relación entre la estructura del tejido y los perfiles de expresión génica.

2.1. Cáncer Renal

El cáncer denomina los padecimientos relacionados con el crecimiento celular descontrolado en algunos tejidos del cuerpo humano, caracterizados por la acumulación progresiva de alteraciones o mutaciones en mecanismos reguladores de la proliferación y muerte celular (Trichopoulos et al. 1996). Como se establece en la investigación señalada anteriormente y al estudio de Evans y Woodward (2023), el funcionamiento normal del ciclo celular permite el crecimiento y la división de células humanas de forma controlada, facilitando la formación de nuevas células; este proceso suele fallar debido a mutaciones genéticas ocasionadas por diferentes causas, como la replicación del ADN durante la división celular, la exposición a agentes externos (como radiación o sustancias químicas) o factores epigenéticos como el envejecimiento (Evans y Woodward 2023).

Estas mutaciones permiten el crecimiento anormal de células y la posterior aparición de tumores, también conocidos como neoplasias, las cuales pueden ser cancerosas (malignos) o no cancerosas (benignos). Los tumores cancerosos o malignos se caracterizan por la invasión de tejido cercano y por su capacidad de viajar a otras partes del cuerpo para formar nuevos tumores mediante el proceso conocido como metástasis. Por otro lado, los tumores benignos no se dispersan ni invaden otros tejidos, ni vuelven a crecer después de ser removidos (Villa Díaz 2025).

Según el Instituto de Investigación del Cáncer (CRI) de Estados Unidos, el cáncer comprende mas de 200 tipos de enfermedades que se caracterizan por el crecimiento descontrolado de células. Uno de los tipos de cáncer de mayor incidencia y mortalidad es el cáncer de riñón, el cual se clasifica en ocho tipos diferentes: carcinoma de células renales, carcinoma urotelial, sarcoma renal, tumor de Wilms, tumores renales benignos (no cancerosos), adenoma renal, oncocitoma y angiomiolipoma.

De todos los tipos de cáncer renal mencionados previamente, el más común alrededor del mundo corresponde al carcinoma de células renales (CCR o RCC por sus siglas en inglés), donde de cada diez casos de cáncer de riñón, nueve son de este tipo, el cual se origina principalmente en el epitelio de los túbulos renales encargados del equilibrio corporal de agua, absorción y filtrado de liquidos (Hsieh et al. 2017).

Al igual que la mayoría de enfermedades cancerosas para distintos órganos, el cáncer de riñón tiene un sistema de graduación para medir la etapa o nivel de avance de la enfermedad mediante la observación de aspectos morfológicos en dicho órgano. Este sistema corresponde al TNM, donde sus siglas hacen referencia al tumor (tamaño/extensión del tumor primario), nodos o ganglios linfáticos (afectación de ganglios cercanos) y metástasis (propagación a otras partes del cuerpo). Este sistema de estadificación fue desarrollado por el Comité Conjunto Estadounidense sobre Cáncer (AJCC, por sus siglas en inglés) y define

la clasificación en cuatro etapas (I–IV), donde la etapa I corresponde al estadio más temprano y la IV al más avanzado (American College of Surgeons. y American Joint Committee on Cancer. 2024, pp. 739–748). A continuación, se definen brevemente las cuatro etapas principales.

1. **Etapa I:** El tumor mide 7 cm o menos y se encuentra únicamente dentro del riñón. Además, no se ha extendido a ganglios linfáticos ni a órganos distantes.

2. **Etapa II:** El tumor mide más de 7 cm, pero sigue estando únicamente dentro del riñón. Al igual que en la etapa I, no hay afectación de ganglios ni metástasis.

3. **Etapa III:** En esta etapa, el cáncer comienza a ser más invasivo. Puede ocurrir de dos formas:

   Extensión local: El tumor se extiende a las venas principales (como la vena cava), pero no ha atravesado la fascia de Gerota (la capa exterior que envuelve el riñón).

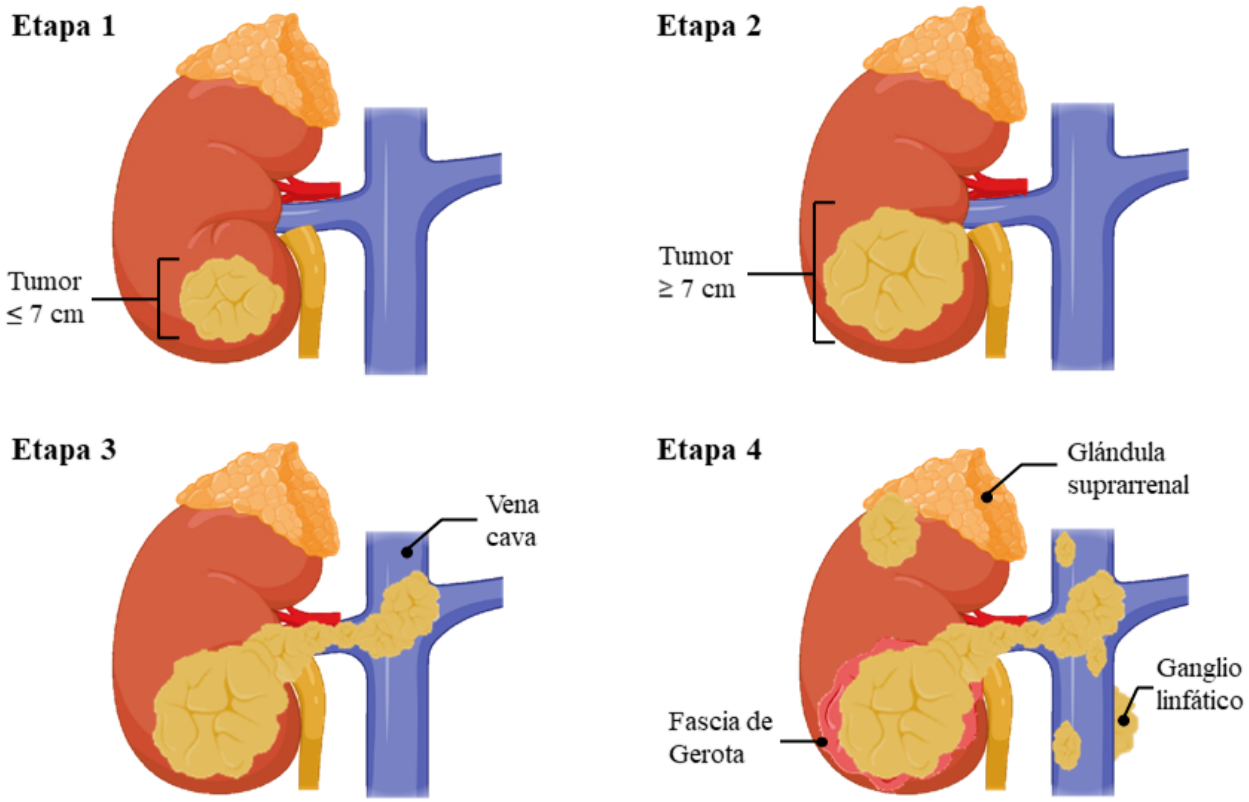
   Ganglios linfáticos: El tumor puede ser de cualquier tamaño y se propaga a los ganglios linfáticos cercanos.

4. **Etapa IV:** El tumor crece más allá de la fascia de Gerota y puede haber invadido la glándula suprarrenal del mismo riñón. El cáncer ha hecho metástasis, propagándose a órganos lejanos o a ganglios linfáticos distantes.

Las diferentes propiedades de cada una de las cuatro etapas o estadios del cáncer de riñón mencionadas previamente son ilustradas en la figura 1.

Figura 1

Esquema de la progresión del cáncer de riñón según las cuatro etapas principales.



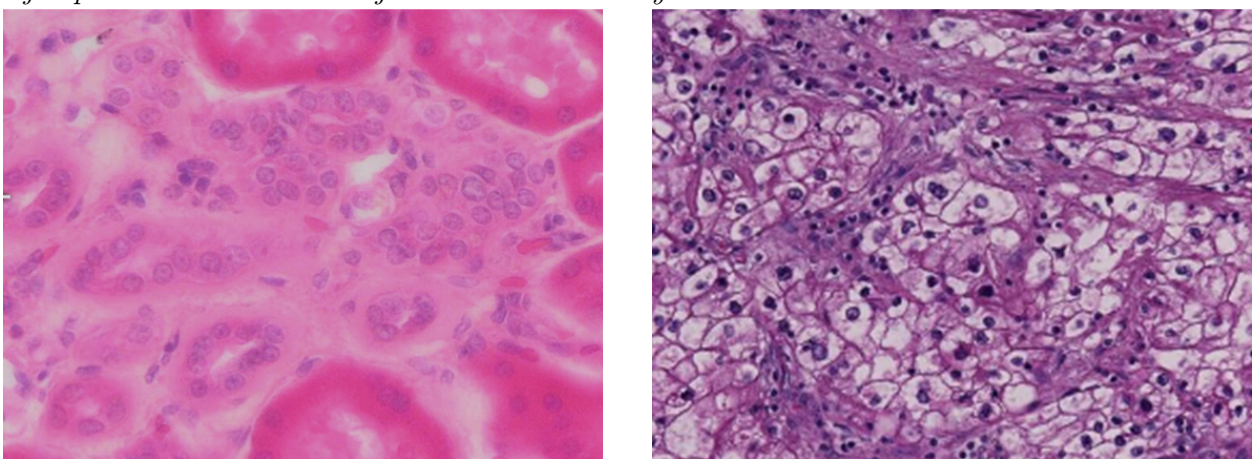
Nota. Adaptado de Smit (s.f.).

El CCR puede presentarse en diferentes subtipos, dependiendo del tipo de células afectadas, siendo el de mayor incidencia el carcinoma renal de células claras (CRCC o ccRCC por sus siglas en inglés), representando siete de cada diez pacientes de carcinoma de células renales (Koul 2019). El carcinoma renal de células claras se caracteriza por una apariencia clara o transparente de las mismas. En la figura 2 se tiene una comparación entre un tejido sano de tejido renal y tejido renal con CRCC, logrando apreciar una mayor transparencia en las células del tejido canceroso.

En muestras de tejido canceroso se puede realizar una graduación o clasificación de la agresividad o etapa del cáncer observando la morfología y estructura de las células dentro del tejido. Para el carcinoma de células renales, se empleaba el grado de Fuhrmann, propuesto en 1982 por Gary A. Fuhrman, el cual tiene en cuenta tres aspectos: el tamaño del núcleo, la forma nuclear y la prominencia de los nucleólos (Fuhrman et al. 1982). Este método clasificaba los núcleos celulares dentro de cuatro grados, siendo el grado cuatro considerado el de mayor avance o agresividad del cáncer, con núcleos demasiado grandes y formas irregulares.

Figura 2

Ejemplos de muestras de tejidos de riñón sano y canceroso.



(a) *Tejido de riñón sano* **(b)** *Tejido de riñón con CRCC*

Nota. Muestras tomadas del conjunto de datos HEST-1K de Jaume et al. (2024).

Sin embargo, esta forma de clasificación no define pesos o nivel de importancia para cada una de las características que eran tenidas en cuenta, ni tenía criterio de la magnificación o escala de observación de la muestra al momento de analizar el tamaño nuclear. Es por esto que en 2016 la Organización Mundial de la Salud (WHO) junto a la Sociedad Internacional de Patología Urológica (ISUP) definió un nuevo sistema de clasificación para tumores del sistema urinario y de los órganos genitales masculinos (Delahunt et al. 2014).

Este nuevo sistema es similar al grado de Fuhrmann en el uso de cuatro grados, teniendo en cuenta únicamente la prominencia del nucleólo y asignando un nivel de magnificación a cada grado para la clasificación nuclear (Moch et al. 2016).

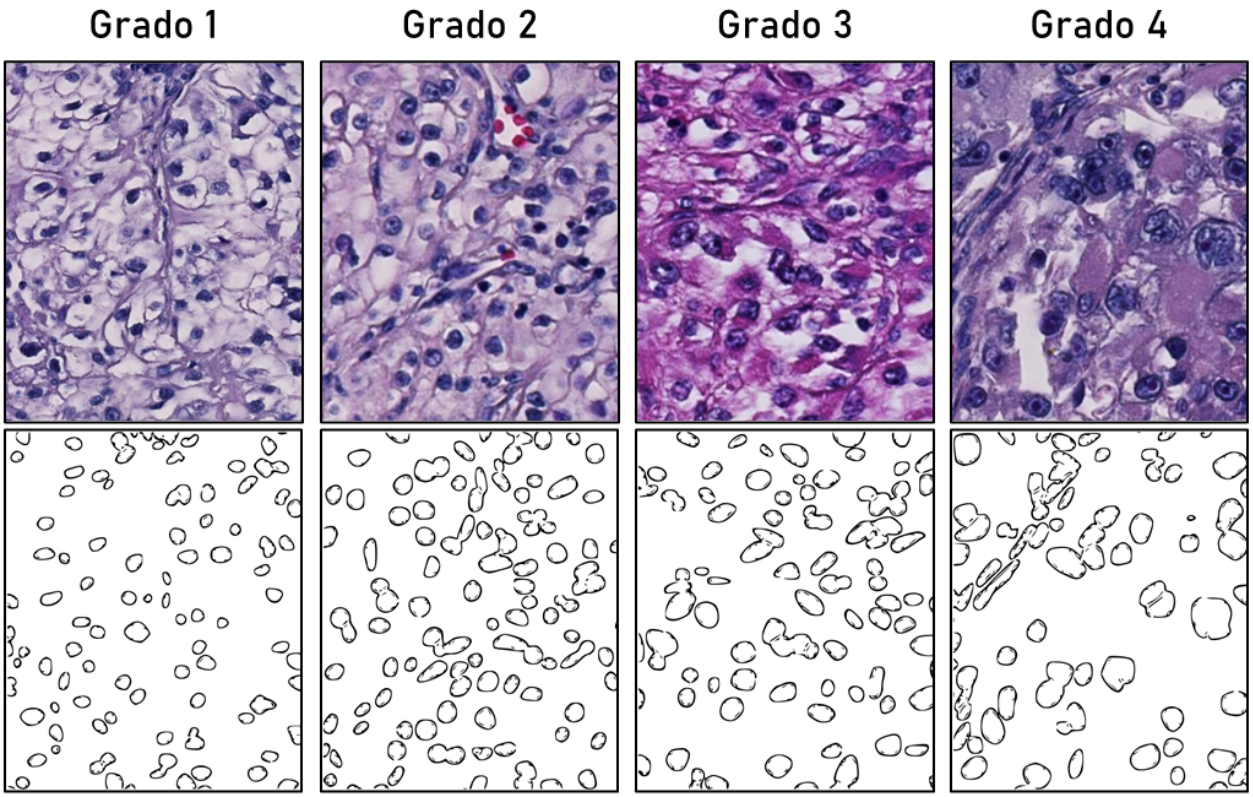
- Grado 1: Nucléolos de células tumorales invisibles o pequeños y basófilos con un aumento de $\times 400$.
- Grado 2: Nucléolos de células tumorales visibles con un aumento de $\times 400$, pero discretos con un aumento de $\times 100$.
- Grado 3: Nucléolos de células tumorales eosinófilos y claramente visibles con un aumento de $\times 100$.
- Grado 4: Tumores que muestran pleomorfismo nuclear extremo y/o que contienen células gigantes tumorales y/o la presencia de cualquier proporción de tumor que muestra desdiferenciación sarcomatoide y/o rabdoide.

Una representación aproximada para los grados de carcinoma renal de células claras mencionados previamente, se ilustra mediante la figura 3, donde se organizan los cuatro

grados en cada columna mediante secciones de tejido tomadas manualmente del dataset HEST-1K, presentando en la fila superior la imagen original del tejido y en la parte inferior la silueta correspondiente de los núcleos segmentados en el tejido para apreciar mejor su forma y tamaño.

Figura 3

Ejemplos de graduación de carcinoma renal de células claras (ccRCC).



Nota. La fila superior muestra el tejido original y la fila inferior las siluetas de núcleos segmentados, permitiendo observar las diferencias morfológicas según el grado tumoral. Adaptado de HEST-1K (Jaume et al. 2024).

2.2. Patología Digital

La patología es la disciplina médica dedicada a investigar las enfermedades desde su base estructural. Esta investiga células, órganos y tejidos intentando descifrar tanto el origen como los mecanismos detrás de cada afección (Cross y Underwood 2013). Así se logra que las observaciones a nivel celular se correlacionen verdaderamente con el cuadro clínico que presenta el paciente. Particularmente la histopatología lleva este análisis al espectro microscópico empleando cortes de tejido teñidos, donde domina la aplicación de hematoxilina y eosina (H&E) (Andrew H. Song et al. 2023). Al usar hematoxilina, los núcleos celulares

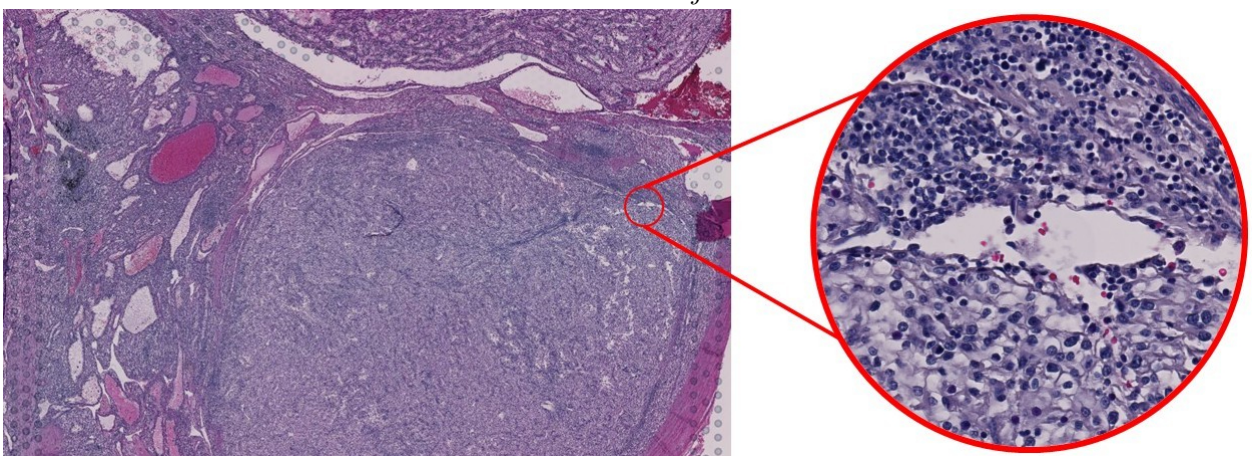
adquieren una tonalidad morada. La eosina reacciona de forma diferente y tiñe de rosa la matriz extracelular junto con el citoplasma. Semejante contraste visual resulta indispensable en oncología clínica, pues revela la arquitectura tisular completa y expone cualquier cambio patológico evidente (Fischer et al. 2008).

Hoy en día la rápida evolución tecnológica ha forzado la entrada de herramientas de digitalización al entorno médico. De este cruce nace la patología digital. Ya los patólogos no dependen exclusivamente de observar láminas físicas de cristal a través de las lentes de un microscopio. En la actualidad el estándar avanza hacia la captura de imágenes digitales de altísima resolución conocidas como Whole Slide Images (WSI). Estas son generadas directamente tras escanear las preparaciones histológicas tradicionales (Litjens et al. 2017). Semejante salto digital simplificó radicalmente la forma de almacenar y evaluar las muestras. Pero el verdadero punto de inflexión llegó al habilitar la aplicación de métodos computacionales sobre estos tejidos. Destaca recientemente la adopción de inteligencia artificial para ejecutar análisis histopatológicos complejos. Estos modelos logran procesar miles de WSI de manera simultánea, alcanzando una escala analítica que la medicina puramente manual sencillamente no podía sostener en el pasado (Baxi et al. 2022).

En la Figura 4 se presenta un ejemplo de una WSI correspondiente a una muestra de tejido renal con cáncer, extraída del conjunto de datos HEST-1K (Jaume et al. 2024).

Figura 4

Visualización a distintas escalas de una WSI de tejido renal.



Nota. La imagen muestra la progresión de escalas en una lámina completa, destacando la morfología mediante tinción H&E en muestras de HEST-1K (Jaume et al. 2024). Resalta núcleos (morado) y citoplasma (rosa) a distintas escalas.

Dentro de este contexto, la patología computacional busca ampliar el alcance de la patología digital a través del desarrollo de métodos algorítmicos que permiten extraer información cuantitativa a partir de las WSI (Andrew H. Song et al. 2023). Una de sus tareas más

fundamentales es la localización de regiones de interés. En términos simples, esto significa identificar automáticamente zonas relevantes dentro de la muestra como áreas tumorales, inflamatorias o necróticas, lo que permite enfocar el análisis en las partes con mayor valor diagnóstico y aliviar en algo el problema del tamaño tan grande que tienen estas imágenes. Sobre estas regiones se construye otra tarea igualmente importante: la segmentación de estructuras celulares y tisulares, que delimita componentes concretos como núcleos o glándulas para extraer medidas morfológicas como el área, la forma o la densidad celular.

Partiendo de estas representaciones ya construidas, se llevan a cabo tareas de clasificación histopatológica. Aquí modelos entrenados reconocen patrones morfológicos asociados a tipos de tejido o estados patológicos, apoyando el diagnóstico y reduciendo parte de la variabilidad que existe entre observadores. Yendo un paso más lejos, el análisis pronóstico relaciona características visuales del tejido con desenlaces clínicos como la supervivencia o la respuesta a tratamientos. Y más recientemente, ha cobrado fuerza la identificación de biomarcadores morfológicos y espaciales, una línea que busca encontrar patrones visuales en el tejido que se correspondan con alteraciones moleculares específicas, abriendo la puerta a integrar el análisis histopatológico con información genómica o transcriptómica de una manera más directa (Baxi et al. 2022; Kather et al. 2019; Litjens et al. 2017).

2.3. Modelos Fundacionales en Patología Digital

El aprendizaje profundo ha transformado el análisis de imágenes médicas mediante arquitecturas capaces de extraer características relevantes directamente de los datos brutos. En particular, las redes neuronales convolucionales (CNN) han permitido automatizar tareas de clasificación y segmentación al identificar representaciones jerárquicas que van desde texturas locales hasta estructuras tisulares complejas (Litjens et al. 2017; Andrew H. Song et al. 2023). No obstante, el rendimiento de estas redes suele estar limitado por la necesidad de grandes volúmenes de datos etiquetados y una baja capacidad de generalización ante la variabilidad morfológica de diferentes órganos.

Ante la complejidad que supone la alta resolución de las imágenes histopatológicas y la demanda de análisis a gran escala, el enfoque ha evolucionado hacia la creación de modelos de propósito general. Este cambio de paradigma dio origen a los modelos fundacionales, los cuales se preentrenan mediante aprendizaje autosupervisado con volúmenes masivos de datos histológicos (R. J. Chen et al. 2024). A diferencia de las CNN tradicionales, estos modelos capturan representaciones universales que pueden transferirse a múltiples tareas diagnósticas sin necesidad de reentrenamiento exhaustivo, optimizando la reproducibilidad y reduciendo significativamente los costos de anotación manual (Awais et al. 2025).

Entre los primeros modelos fundacionales propuestos para patología computacional se

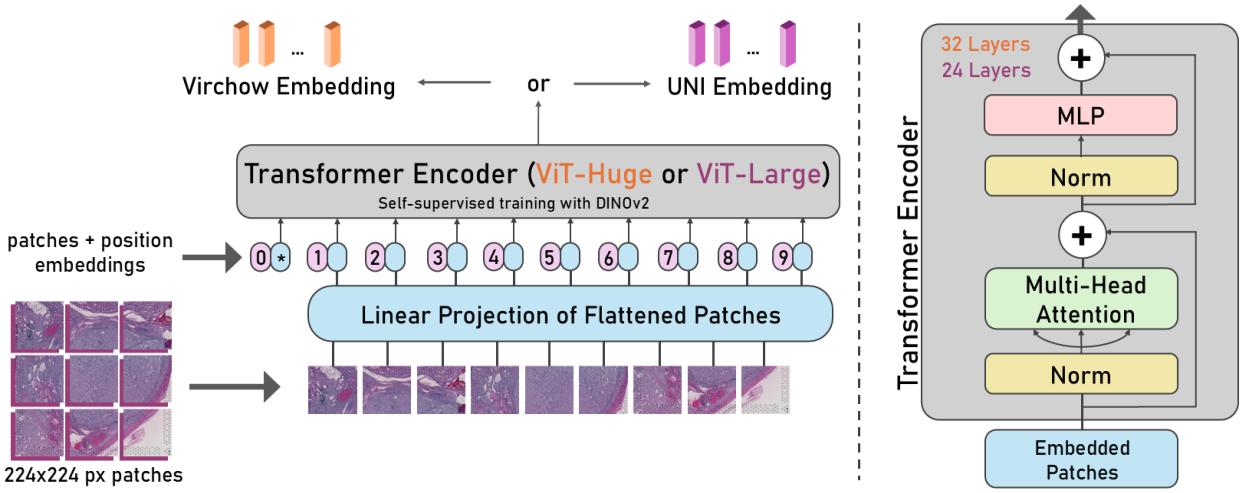
encuentra UNI, modelo de aprendizaje autosupervisado con base en DINOv2 (Oquab et al. 2024) diseñado específicamente para histopatología. Se concibe como un encoder visual de propósito general, capaz de aprender representaciones transferibles a múltiples aplicaciones posteriores sin ser optimizado para una enfermedad específica. Se basa en una arquitectura Vision Transformer (ViT-L/16) y es preentrenado con más de 100 millones de imágenes de más de 100.000 WSI de diagnóstico teñidas con H&E en 20 tipos de tejidos principales para la primera versión, y con más de 200 millones de imágenes H&E e IHC muestreadas de más de 350.000 WSI diversas para la segunda versión. En el HEST-Benchmark para predicción de expresión génica a partir de histología, UNI alcanza correlaciones promedio de Pearson de 0,3862 en su primera versión y de 0,4139 en la segunda, el valor más alto entre los modelos comparados, superando a modelos CNN tradicionales y mostrando un desempeño competitivo frente a arquitecturas de mayor escala (Jaume et al. 2024). UNI no está orientado a una tarea puntual. Su enfoque consiste en aprender patrones morfológicos generales y relaciones espaciales del tejido a gran escala, de modo que las representaciones obtenidas puedan reutilizarse con facilidad en distintos problemas de patología computacional (R. J. Chen et al. 2024).

En contraste con UNI, se presenta Virchow, un modelo diseñado explícitamente para aprender representaciones orientadas a la patología en lugar de características visuales generales. Al igual que UNI, se trata de un modelo autosupervisado. En su versión inicial, Virchow emplea una arquitectura Vision Transformer (ViT-H/14) que se basa en DINOv2 (Oquab et al. 2024). Esta versión es preentrenada con datos de aproximadamente 100.000 pacientes, correspondientes a cerca de 1,5 millones de WSI. Estas son divididas en mosaicos muestreados a una resolución de 0,5 micras por píxel (aumento de 20x) teñidas con H&E adquiridas del Memorial Sloan Kettering Cancer Center (MSKCC) (Vorontsov et al. 2024). En el HEST-Benchmark esta primera versión alcanza una correlación promedio de Pearson de 0.3977 (Jaume et al. 2024). La segunda versión introduce modificaciones sobre DINOv2: el regularizador KoLeo fue reemplazado por un estimador de densidad de kernel y el aumento de recorte y cambio de tamaño fue reemplazado por una traducción de contexto extendida. También mantiene la arquitectura ViT-H/14 preentrenada con un conjunto de datos interno de aproximadamente 3,1 millones de WSI del MSKCC, que abarca una amplia gama de afecciones patológicas. Las imágenes se muestrean a múltiples resoluciones espaciales de 2,0, 1,0, 0,5 y 0,25 micras por píxel, correspondientes a aumentos de 5x, 10x, 20x y 40x (Zimmermann et al. 2024). Este diseño de muestreo multirresolución permite al modelo capturar la organización celular y la estructura tisular a escala fina, lo que refuerza su especialización en patrones morfológicos relevantes para el cáncer y la patología diagnóstica. A diferencia de UNI, que se orienta a la transferibilidad general entre tareas, Virchow prioriza representacio-

nes patológicas profundas en múltiples escalas. En el HEST-Benchmark, la segunda versión alcanza una correlación promedio de Pearson de 0,3984 (Jaume et al. 2024).

Figura 5

Arquitecturas de los modelos fundacionales Virchow y UNI.



Nota. Ambos modelos se basan en la arquitectura Vision Transformer (ViT). Virchow emplea una variante ViT-Huge con 32 capas, mientras que UNI utiliza una ViT-Large con 24 capas. Adaptado de R. J. Chen et al. (2024), Dosovitskiy et al. (2021) y Vorontsov et al. (2024).

Más allá de los modelos puramente visuales, OmiCLIP propone un enfoque multimodal basado en una arquitectura CLIP (Kim et al. 2021). El modelo combina un encoder visual para parches histológicos con un encoder molecular entrenado sobre datos transcriptómicos y se entrena con el conjunto ST-Bank, que incluye aproximadamente 2,2 millones de parches provenientes de 1.007 muestras distribuidas en 32 órganos. Cada parche está emparejado con perfiles de transcriptómica espacial obtenidos mediante la tecnología 10x Visium. Inspirado en enfoques basados en modelos de lenguaje a gran escala, la transcriptómica de cada parche se representa como una "frase"formada por los genes de mayor expresión separados por espacios (‘ ’). Aunque OmiCLIP no fue evaluado dentro del HEST-Benchmark, sus resultados se reportan en tareas de estructuración multimodal y alineamiento espacial. Tras el entrenamiento contrastivo se observó un incremento significativo en la calidad del clustering según la métrica Calinski–Harabasz ($p < 0{,}001$ en muestras anotadas y $p < 0{,}05$ en el resto de órganos). Además, el encoder visual superó a modelos como UNI y GigaPath en tareas de organización tisular, lo que sugiere que esta alineación explícita entre imagen y transcriptómica permite capturar mejor la heterogeneidad del tejido (W. Chen et al. 2025).

Siguiendo la línea hacia la integración de múltiples fuentes de información, PS3 (Predicting Survival from Three Modalities) combina imágenes histológicas, datos transcriptómicos y reportes de patología clínica para la predicción de supervivencia. A diferencia de

enfoques que integran únicamente imagen y transcriptómica, el modelo incorpora explícitamente la información textual de los reportes diagnósticos como una tercera modalidad. Para abordar la heterogeneidad entre estas fuentes, PS3 adopta un esquema basado en prototipos. Las WSI se representan mediante prototipos histológicos obtenidos a partir de agrupamientos morfológicos recurrentes, los datos transcriptómicos se transforman en 50 prototipos asociados a rutas biológicas (Cancer Hallmarks) y los reportes se codifican en prototipos diagnósticos mediante mecanismos de autoatención. Estas representaciones se integran mediante un Transformer multimodal que modela interacciones intra e intermodales entre imagen, transcriptómica y texto. En la etapa de extracción de características visuales y textuales el modelo emplea arquitecturas visión-lenguaje preentrenadas como PLIP (Huang et al. 2023) y, en estudios de ablación, QUILT-Net (Ikezogwo et al. 2025), ambas basadas en CLIP (Kim et al. 2021). Evaluado sobre seis cohortes del consorcio The Cancer Genome Atlas (TCGA) en tareas de predicción de supervivencia específica por enfermedad, PS3 alcanza un C-Index de 0,699 según lo reportado en su trabajo original (Raza et al. 2025).

2.4. Representaciones Centradas en Núcleos para Imágenes Histopatológicas

La caracterización nuclear constituye una de las bases de mayor importancia para el área de la histopatología computacional, ya que a partir del núcleo celular se puede obtener múltiples señales o características morfológicas relacionadas con el estado o la estructura biológica del tejido. La presencia de alteraciones en el tamaño, la forma y la organización espacial de las células suele estar relacionada con procesos relativos al cáncer, como la proliferación descontrolada de células tumorales o inestabilidad genética. Cuantificar estas características permite obtener descriptores numéricos analizables a través de patrones visuales que se observen en imagenes histopatológicas.

Este tipo de características pueden ser estudiadas desde dos enfoques complementarios: el análisis y descripción de cada núcleo celular de forma individual mediante descriptores tradicionales de su morfología y textura; y el análisis de las relaciones estructurales y espaciales entre núcleos utilizando representaciones basadas en grafos. Mediante los enfoques mencionados previamente, se puede obtener información de la organización tisular desde distintas dimensiones, siendo de gran utilidad para el análisis del microambiente tumoral. A continuación se explica de manera más detallada en qué consisten estos enfoques.

2.4.1. Características Nucleares Tradicionales

Las características nucleares tradicionales, también conocidas como *handcrafted features*, ofrecen una descripción cuantitativa acerca de la morfología, textura e intensidad de

los nucleos celulares de manera individual en imágenes histológicas. De esta forma, es posible relacionar estas observaciones con procesos biológicos internos del tejido en estudio, siendo el núcleo celular el orgánulo más importante de la célula, actuando como centro regulador de la genética y el metabolismo celular (Gao et al. 2025).

Estas características pueden organizarse en distintas categorías según el tipo de información que obtienen y analizan. Estas incluyen descriptores basados en morfología, textura y enfoques híbridos que combinan rasgos manuales con representaciones aprendidas.

1. **Características morfológicas:** Estas propiedades incluyen medidas como el perímetro, el área, la circularidad, la excentricidad y las relaciones de aspecto, reflejando variaciones en términos de tamaño y forma nuclear (Ji et al. 2019; Kumar et al. 2022). Estas características son ampliamente utilizadas para el estudio de la malignidad y el avance tumoral, por ejemplo, mediante el grado de Fuhrmann para el cáncer de riñón mencionado en la sección 2.1.

2. **Características de textura e intensidad:** Las características de textura e intensidad, tales como las métricas de Haralick y el Patrón Binario Local (LBP), cuantifican la interacción espacial entre regiones de tejido, analizando la interacción entre píxeles y de los patrones locales de color del tejido en estudio. Del mismo modo se puede obtener estadísticas a partir de histogramas de color, los cuales ofrecen información de la heterogeneidad cromática y organización de la cromatina en la muestra. Este tipo de características resultan útiles para evaluar la calidad de las segmentaciones nucleares realizadas por determinado modelo, detectando por ejemplo, subsegmentación o núcleos no detectados debido al bajo contraste o intensidad de la muestra (Porebski et al. 2008; Wen et al. 2017).

3. **Características híbridas (handcrafted + deep learning):** Integran características aprendidas por redes convolucionales a partir de datos brutos (como la intensidad de los píxeles) y descriptores manuales o morfológicos, esto con el objetivo de agregar información con un mayor contenido informativo al modelo, logrando una representación mas clara y acertada de la apariencia de los núcleos celulares. Las características híbridas resultan útiles en escenarios donde los datos sean escasos o haya una baja calidad de las muestras en estudio (Kashif et al. 2016).

El análisis y la cuantificación multivariada de este tipo de características nucleares permiten tanto la detección como la diferenciación entre distintos niveles de avance de las enfermedades, incluido el cáncer. Un ejemplo de ello es el estudio realizado por Konstandinou et al., donde utilizaron un conjunto de 63 características nucleares tradicionales para distinguir

entre niveles bajos y altos de neoplasia intraepitelial cervical (CIN), la cual es una condición precancerosa para el cáncer cervical invasivo. Mediante el análisis de las muestras de 44 pacientes diagnosticados con esta condición, se destacaron 19 descriptores morfológicos relacionados con el tamaño, la forma y la textura de los núcleos celulares, siendo útiles en la diferenciación entre niveles bajos y altos de esta enfermedad (Konstandinou et al. 2018).

Un aspecto a tener en cuenta para el estudio de características nucleares es el método de extracción de las mismas, proceso el cual suele llevarse a cabo mediante métodos computacionales automatizados o modelos de segmentación nuclear ya definidos. Dependiendo del método de extracción, la calidad de las características observadas puede variar significativamente, lo que puede generar resultados acertados o, por el contrario, sesgados debido a una mala extracción (Sumon et al. 2025).

2.4.2. Características Nucleares Basadas en Grafos

Un grafo se define como una estructura compuesta por un conjunto finito de nodos (o vértices) y un conjunto de aristas que representan relaciones o conexiones entre dichos nodos (Diestel 2025; Gutiérrez García 2024). En el contexto de la histopatología computacional, los nodos representan núcleos celulares observados en el tejido teñido con H&E, mientras que las aristas representan relaciones espaciales o morfológicas entre ellos, tales como interacción, proximidad espacial, similitud, entre otras (Gadiya et al. 2019).

En el análisis histológico mediante grafos, existen varias formas de construir este tipo de representaciones, dependiendo del tipo de relaciones o atributos que se deseen modelar entre los núcleos celulares. A continuación, se describen los principales tipos de grafos utilizados en el estudio de la arquitectura celular.

1. **Grafos de proximidad y relaciones espaciales:** Los grafos de proximidad se construyen mediante la conexión de núcleos cercanos en función de distancias mediante un radio predefinido, ya sea a través de distancias euclidianas o de vecindad. Cada arista dentro del grafo representa la proximidad espacial entre nodos, y puede tener atributos como distancia, diferencia de forma o intensidad del núcleo, entre otros. El objetivo de este tipo de grafos es capturar y representar la arquitectura del tejido en estudio, siendo un ejemplo de ello el estudio realizado por Romo-Bucheli et al. para la identificación de regiones asociadas al carcinoma de células basales, o también conocido como cáncer de piel, mediante características topológicas extraídas de un grafo de distancia basado en núcleos (Romo-Bucheli et al. 2017).

2. **Grafos multi-atributo:** Este tipo de grafos se caracterizan por la implementación de múltiples atributos para cada nodo, los cuales describen distintas características

nucleares tradicionales, tales como tamaño, textura o forma, y cada arista representa distintas relaciones espaciales o de interacción entre distintas células o núcleos (Gadiya et al. 2019). De esta forma, es posible obtener simultáneamente información relacionada con la morfología de los núcleos y la organización espacial de los mismos en el tejido, facilitando la implementación de algoritmos de aprendizaje supervisado como redes convolucionales de grafos (GCNs) en tareas de clasificación e identificación de tumores o tejidos cancerosos. (Ahmedt-Aristizabal et al. 2022).

3. **Cell–ECM graphs:** Los *Cell–ECM graphs* son grafos nucleares que incorporan elementos de la matriz extracelular (ECM), la cual es una estructura o red tridimensional de proteínas y polisacáridos que se encargan de sostener y dar estructura a las células presentes en los tejidos vivos, además de otras funciones como la adhesión, la comunicación y la migración celular. Esta matriz es representada como nodos adicionales, y las aristas representan interacciones de tipo célula-célula, ECM-ECM y célula-ECM, permitiendo modelar de forma más detallada la arquitectura del tejido. Esta representación facilita el desarrollo de tareas como la detección de nichos celulares y el descubrimiento no supervisado de subregiones (Ghafoor et al. 2025).

Los grafos celulares o nucleares pueden ser estudiados desde una perspectiva estructural para comprender la organización global del tejido, permitiendo el análisis de propiedades topológicas, la identificación de patrones de agrupamiento y la relación entre morfología nuclear con características biológicas. Lou et al. utilizan este método para la calsificación de núcleos, tacando una mejoría en la discriminación de tipos de células (Lou et al. 2024). Asimismo, existen otros enfoques orientados al estudio de subestructuras locales y distintas métricas de centralidad para caracterizar la heterogeneidad tumoral (Ghafoor et al. 2025).

La construcción de estas representaciones basadas en grafos nucleares permite desarrollar una gran variedad de modelos computacionales, como las redes neuronales de grafos, enfocadas en el aprendizaje de la organización espacial celular para el análisis de patrones complejos.

Redes Neuronales de Grafos (Graph Neural Networks)

Las *Graph Neural Networks (GNNs)* utilizan operaciones de convolución y agregación sobre los nodos y aristas de un grafo, aprendiendo a partir de información local y global del tejido (Ahmedt-Aristizabal et al. 2022; Brussee et al. 2025). Estas redes utilizan esquemas de propagación de mensajes (*message passing*) para la actualización iterativa de la representación de cada nodo a partir de la información de sus nodos vecinos. Las GNNs integran

características nucleares morfológicas con información topológica proveniente de interacciones entre células, lo que permite la clasificación de núcleos, la detección de comunidades de células y la predicción de estados clínicos mediante la organización celular. Estos modelos son ampliamente utilizados dentro de la histopatología computacional, siendo ejemplos de ello el desarrollo del modelo CGC-Net para la graduación de cáncer colorrectal (Zhou et al. 2019) y SlideGraph+ para predecir el estado del gen HER2 relacionado con el cáncer de mama (W. Lu et al. 2022).

Refinamiento y Alineamiento de grafos

Previo a la utilización de modelos de aprendizaje basados en grafos, es posible implementar técnicas de refinamiento y alineamiento para optimizar la estructura inicial del grafo. Esto con el objetivo de dar una mayor solidez al modelo en caso de presentarse variaciones en los datos histológicos utilizados, facilitando su generalización entre distintos contextos tisulares. Por ejemplo, Hassan et al. proponen un método de refinamiento para el ajuste de las aristas y los atributos de los nodos mediante un proceso de aprendizaje representacional sobre el grafo (Hassan et al. 2024). Este proceso mejora la identificación de comunidades nucleares con distintas propiedades patológicas y la diferenciación de patrones celulares similares.

2.5. Transcriptómica

Entender cómo funciona una célula implica, en buena medida, saber qué genes tiene activos en un momento dado. Esto es precisamente lo que aborda la transcriptómica: el estudio del conjunto de moléculas de ARN (ácido ribonucleico) que se produce en una célula o tejido bajo condiciones específicas. El hecho de que todas nuestras células compartan el mismo genoma hace aún más relevante esta pregunta, porque lo que diferencia a una neurona de una célula hepática no es su ADN, sino qué parte de ese ADN se transcribe, cuándo y en qué cantidad. Esas diferencias en la expresión génica son las que determinan la identidad y el comportamiento celular (Dong e Y. Chen 2013). El proceso subyacente sigue lo que se conoce como el dogma central de la biología molecular: la información genética fluye del ADN (ácido desoxirribonucleico) al ARN (ácido ribonucleico) mediante el proceso de transcripción y de ahí a la proteína a través de la traducción (Zhong et al. 2025). Sin embargo, el salto conceptual que ofrece la transcriptómica no es explicar ese flujo, que lleva décadas siendo conocido, sino poder medirlo de forma global y sistemática en miles de genes simultáneamente, lo que permite identificar el estado funcional de una célula en un momento determinado.

2.5.1. La Expresión Génica y su Transcripción

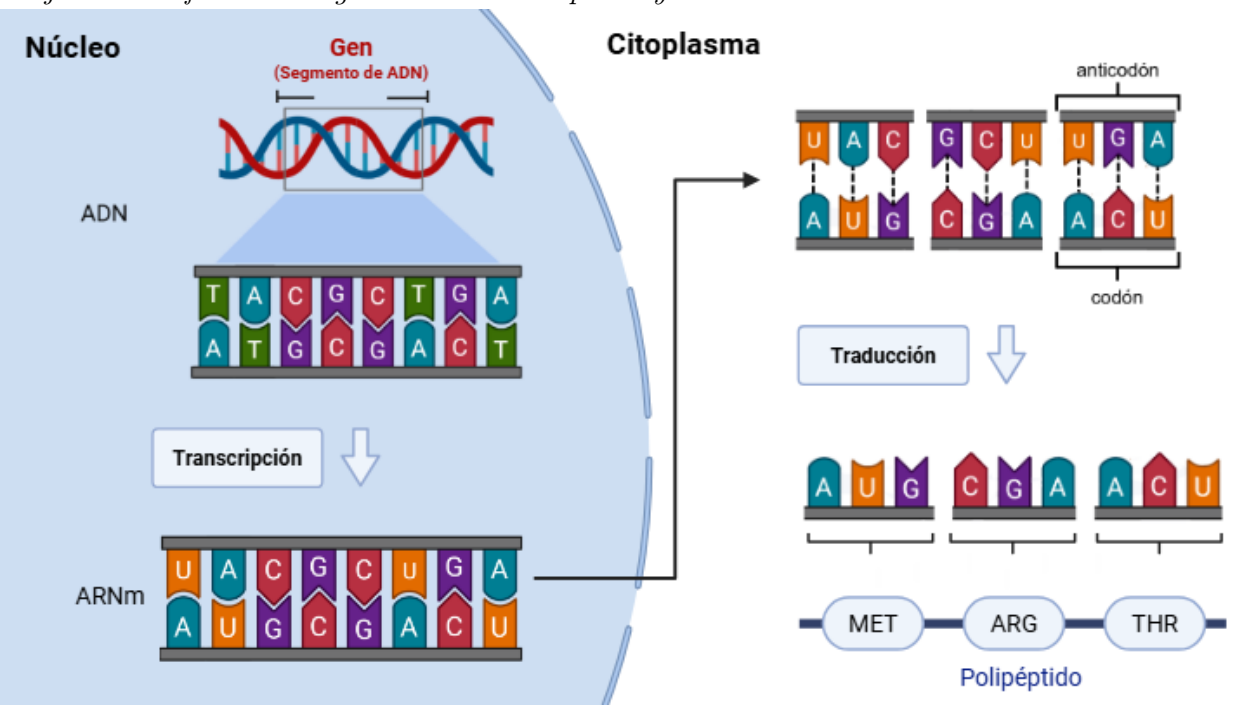
El material de trabajo de la transcriptómica son los transcritos, es decir, los fragmentos de ARN generados cuando un gen se activa, y el conjunto total de estos transcritos presentes en una célula o tejido se denomina transcriptoma. A diferencia del genoma, el cual suele ser el mismo en la mayoría de células, el transcriptoma varía según el tipo de célula, el estado fisiológico, el entorno celular y la presencia de alteraciones patológicas o estímulos externos (Stark et al. 2019). El estudio del transcriptoma equivale a tomar una fotografía del estado funcional de la célula en un instante determinado, capturando información útil sobre los procesos biológicos en curso.

Proceso de Transcripción y Traducción:

La expresión génica se constituye de varios procesos mediante los cuales la información contenida dentro de un gen se convierte en un producto funcional o proteína (Kukurba y Montgomery 2015). El paso inicial corresponde al proceso de transcripción, durante el cual una región del ADN actúa como molde para producir ARN mensajero (ARNm), el cual se encarga de transportar la información genética desde el núcleo de la célula hacia los ribosomas (Alhalmi et al. 2020). Posteriormente, se realiza el proceso de traducción, el cual consiste en la lectura del ARNm en el ribosoma mediante secuencias de tres nucleótidos o bases llamados codones. Cada triplete o codón especifica un aminoácido, y la cadena resultante de aminoácidos se pliega para dar lugar a una proteína funcional (Zhong et al. 2025). La Figura 6 esquematiza el flujo de información desde el ADN hasta la formación de proteínas funcionales.

Figura 6

Flujo de la información genética: transcripción y traducción.



Nota. Esquema del proceso de síntesis de proteínas, iniciando con la transcripción del ADN a ARNm en el núcleo y finalizando con la traducción a una cadena polipeptídica en el citoplasma. Imagen creada con BioRender.com.

Alteraciones en la secuencia de ADN o en la regulación de la expresión génica pueden modificar el comportamiento celular. Las alteraciones más comunes corresponden a las mutaciones puntuales de nucleótidos que ocurren durante el proceso de replicación del ADN en el ciclo celular, dando lugar a tres posibles escenarios (Banoon et al. 2022):

1. **Sustitución:** una base se agrega incorrectamente y reemplaza el par en la posición correspondiente en la cadena.
2. **Inserción:** uno o más nucleótidos son insertados en la cadena durante la replicación.
3. **Eliminación o silenciamiento:** uno o más nucleótidos son omitidos ya sea durante la replicación o algún otro proceso.

Estas mutaciones pueden desregular procesos como la proliferación celular o la apoptosis, favoreciendo el desarrollo de enfermedades como el cáncer (Hanahan 2022). Asimismo, mecanismos epigenéticos, como la metilación del ADN o modificaciones en histonas influyen

en la expresión génica sin alterar la secuencia original (Villanueva et al. 2020). En este contexto, el análisis transcriptómico permite identificar rutas moleculares alteradas, subtipos de enfermedad y posibles biomarcadores diagnósticos o terapéuticos (Shulman et al. 2024).

El avance de la transcriptómica ha favorecido en gran medida el desarrollo de tecnologías de secuenciación masiva, como RNA-seq, y la creación de herramientas computacionales avanzadas, permitiendo cuantificar millones de transcritos simultáneamente (Z. Chen et al. 2022). Esto ha facilitado la construcción de atlas de expresión génica en diversos tejidos y órganos humanos, como el GTEx (Genotype-Tissue Expression), favoreciendo la identificación de patrones moleculares que pueden estar relacionados con distintas condiciones de salud o enfermedades (Lonsdale et al. 2013).

2.5.2. Transcriptómica Espacial

Mientras que la transcriptómica convencional cuantifica niveles de expresión génica sin considerar su ubicación dentro del tejido, la transcriptómica espacial incorpora información sobre la localización de los transcritos. Esta tecnología fue presentada por Ståhl et al. y permite relacionar directamente la actividad molecular con la arquitectura histológica, facilitando el estudio de la heterogeneidad tisular y de las interacciones celulares en su contexto anatómico (Ståhl et al. 2016).

El procedimiento general consiste en fijar el tejido sobre una lámina que contiene oligonucleótidos con códigos de barras espaciales. Las moléculas de ARN capturadas se etiquetan según su posición dentro del corte histológico, se secuencian y posteriormente se reconstruye un mapa bidimensional que representa la distribución espacial de la expresión génica (T.-Y. Chen et al. 2023). En este contexto, la unidad fundamental de medición en muchas de estas tecnologías es el denominado *spot*, el cual corresponde a un punto de captura sobre la superficie del portaobjetos que contiene múltiples oligonucleótidos con un código de barras espacial único. Cada *spot* captura ARN mensajero proveniente de una región específica del tejido, permitiendo asociar los niveles de expresión génica con una localización concreta. En plataformas basadas en secuenciación como Visium, esta expresión se cuantifica típicamente en términos de *conteos crudos de transcritos* (*raw counts*), es decir, el número de moléculas de ARN detectadas por gen en cada *spot*. En el caso particular del conjunto de datos HEST-1k, estos conteos se almacenan sin normalización adicional en estructuras tipo `AnnData`, permitiendo aplicar posteriormente distintas estrategias de procesamiento según el análisis requerido (por ejemplo, normalización por conteo total o transformaciones logarítmicas como *log1p*). Adicionalmente, cada *spot* se encuentra alineado espacialmente con la imagen histológica mediante un proceso automatizado de detección de referencias y extracción de parches, lo que garantiza la correspondencia precisa entre la información mor-

fológica y transcriptómica. No obstante, debido a su tamaño (por ejemplo, alrededor de 55 μm en plataformas como Visium), un *spot* puede contener múltiples células, lo que limita la resolución espacial a nivel subcelular.

Debido a limitaciones de este estilo, con el paso del tiempo se han desarrollado diferentes tecnologías de transcriptómica espacial, las cuales pueden clasificarse según el mecanismo mediante el cual adquieren la información espacial y el tipo de datos generados. En términos generales, existen cuatro categorías principales: (i) métodos basados en secuenciación, (ii) métodos basados en sondas, (iii) métodos basados en imágenes y (iv) enfoques de secuenciación de célula única guiados por imágenes (T.-Y. Chen et al. 2023). Entre estas, las dos primeras aproximaciones más ampliamente utilizadas corresponden a los métodos basados en secuenciación y los métodos basados en imágenes.

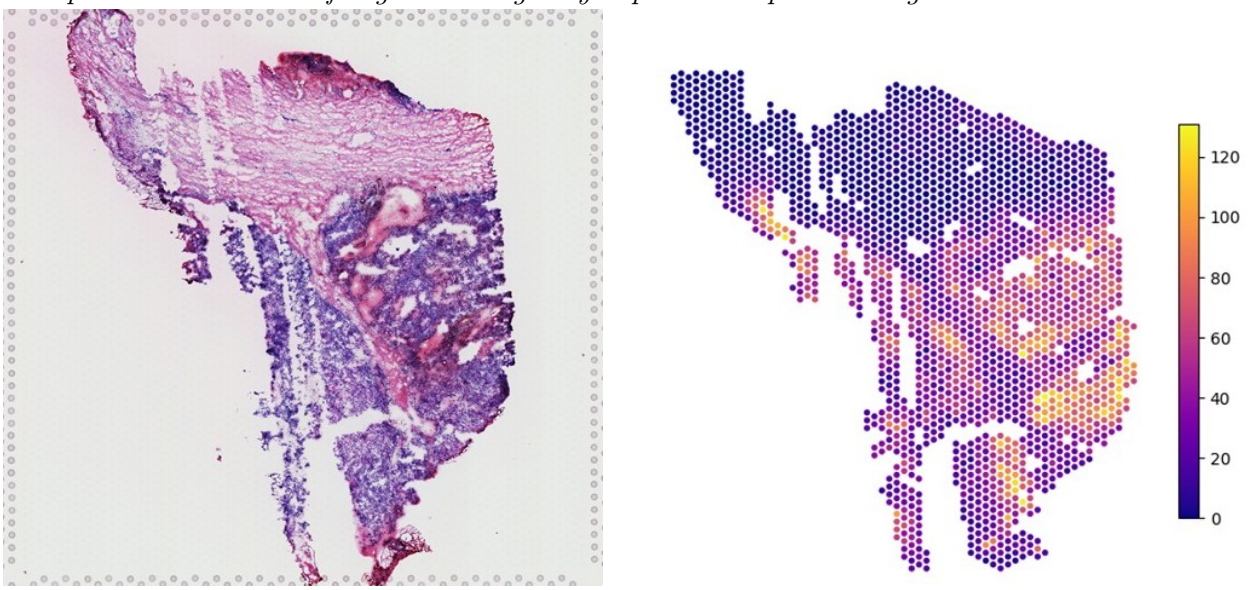
Los métodos basados en secuenciación (NGS), como Spatial Transcriptomics, Visium y sus variantes más recientes, capturan ARN poliadenilado de manera no sesgada mediante matrices de captura con códigos de barras espaciales. Posteriormente, los transcritos son secuenciados y asignados a sus coordenadas originales, permitiendo reconstruir el transcriptoma completo en cada *spot* mediante el conteo de transcritos ya explicado anteriormente. Este enfoque permite el análisis del transcriptoma completo, pero su resolución espacial está limitada por el tamaño de los puntos de captura.

Por otro lado, los métodos basados en imágenes, como Xenium, se fundamentan en la hibridación in-situ mediante sondas fluorescentes y ciclos iterativos de adquisición de imágenes. Estas técnicas permiten localizar directamente los transcritos dentro del tejido con resolución subcelular, proporcionando la ubicación exacta de cada molécula de ARN. Sin embargo, requieren la definición previa de un conjunto de genes de interés, lo que limita su capacidad para explorar el transcriptoma completo.

La Figura 7 muestra un ejemplo de una imagen histológica teñida con H&E junto con el mapa espacial de expresión del gen UBC relacionado con la presencia de carcinoma renal de células claras (Jiang et al. 2024), evidenciando cómo la expresión génica puede variar dentro de distintas regiones del tejido con niveles elevados de expresión en zonas tumorales.

Figura 7

Comparación entre morfología histológica y expresión espacial del gen UBC.



(a) *Imagen histológica (H&E)* **(b)** *Expresión del gen*

Nota. Comparación entre una WSI y el mapa de expresión espacial del gen UBC en el mismo tejido, tomadas del conjunto de datos HEST-1K (Jaume et al. 2024).

Esta capacidad de integrar morfología y molecularidad ha hecho de la transcriptómica espacial una herramienta valiosa para estudiar enfermedades complejas, particularmente cáncer, donde la organización y las variaciones del microambiente tisular tienen un papel determinante en el desarrollo de procesos patológicos, incluido el avance tumoral (Levy-Jurgenson et al. 2020). Con todo, su adopción masiva sigue siendo limitada debido a los altos costos y al número reducido de muestras procesables por experimento. Frente a esto, han surgido tanto variantes tecnológicas más asequibles como aproximaciones computacionales que buscan predecir perfiles de expresión espacial directamente a partir de imágenes histológicas, utilizando los datos de transcriptómica espacial como señal de supervisión (Zahedi et al. 2024). De manera similar, Juwayria et al. propusieron la técnica Microarray Integrated Spatial Transcriptomics (MIST), la cual integra microarreglos de tejidos (TMAs) con la plataforma Visium, utilizando corte láser y herramientas de impresión 3D para mejorar el rendimiento por muestra y reducir costos en comparación con otros métodos comerciales (Juwayria et al. 2024).

2.6. Representaciones en Histopatología y Transcriptómica Espacial

La relación entre la morfología tisular y la expresión génica ha sido abordada desde distintos enfoques representacionales. En esencia, el problema consiste en definir cómo se modela la información visual extraída de imágenes histopatológicas y cómo esta se vincula con los perfiles transcriptómicos espaciales. En lugar de depender de una única metodología, la literatura reciente ha explorado diversas estrategias, que abarcan desde descriptores tradicionales hasta modelos multimodales entrenados de extremo a extremo.

Un primer enfoque se basa en características morfológicas tradicionales, descritas en la sección 2.4.1, que incluyen propiedades como forma, tamaño, textura e intensidad nuclear. Estas representaciones permiten cuantificar patrones celulares y analizar su relación con estados moleculares específicos. Aunque no fueron diseñadas originalmente para escenarios multimodales, han servido como base para estudiar asociaciones entre rasgos histológicos y expresión génica, especialmente en contextos con datos limitados. En esta línea, trabajos como el de Gao et al. han evidenciado que variaciones en características nucleares se asocian con distintos patrones de expresión génica en múltiples tejidos (Gao et al. 2025).

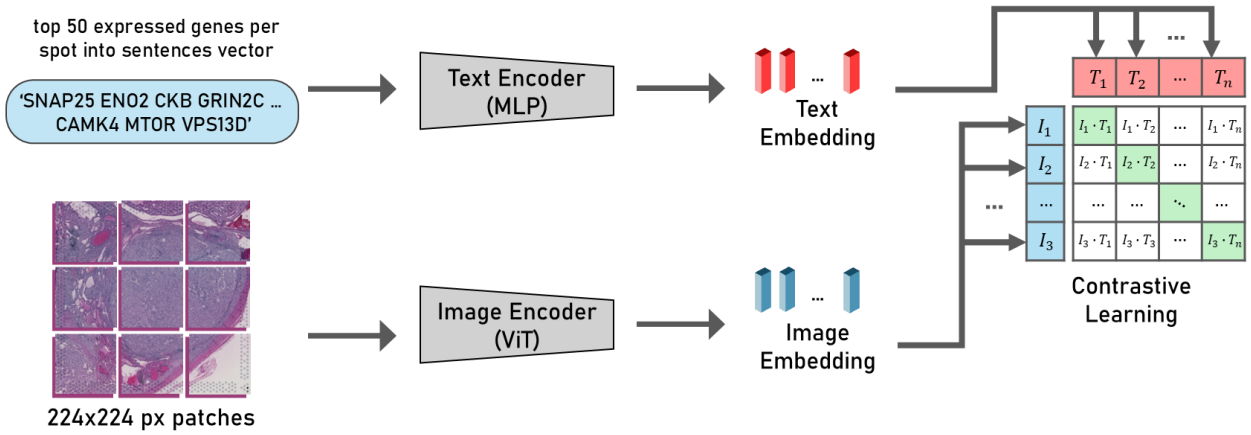
Posteriormente, con el avance del aprendizaje profundo, surgieron representaciones aprendidas automáticamente a partir de grandes volúmenes de imágenes histológicas. Modelos convolucionales y arquitecturas basadas en mecanismos de atención han demostrado su capacidad para capturar patrones morfológicos complejos, los cuales presentan correlaciones significativas con la expresión génica (Jia et al. 2023; Levy-Jurgenson et al. 2020). En este contexto, la predicción de la expresión génica se formula como una tarea supervisada, donde las imágenes actúan como entrada y los perfiles transcriptómicos como variable objetivo.

Más recientemente, los modelos fundacionales visuales entrenados sobre grandes colecciones de láminas histológicas han mostrado que es posible aprender representaciones morfológicas transferibles a tareas moleculares (R. J. Chen et al. 2024; Zimmermann et al. 2024). Aunque estos modelos no fueron diseñados inicialmente como arquitecturas multimodales, sus embeddings visuales han sido reutilizados para apoyar la predicción de expresión génica en conjuntos con datos co-registrados imagen–transcriptómica.

En contraste con estos enfoques predominantemente visuales, los modelos multimodales buscan alinear explícitamente la información morfológica y molecular en un espacio latente compartido. Inspirados en arquitecturas contrastivas, algunos trabajos entrenan conjuntamente encoders visuales y textuales para aprender correspondencias entre parches histológicos y perfiles de expresión génica (W. Chen et al. 2025; Raza et al. 2025). Este tipo de estrategias permite realizar predicciones y también explorar relaciones entre modalidades.

Figura 8

Arquitectura multimodal del modelo fundacional OmiCLIP.



Nota. Adaptado de W. Chen et al. (2025) y Radford et al. (2021).

En una línea complementaria, las representaciones basadas en grafos han permitido modelar no solo atributos individuales de núcleos o regiones tisulares, sino también sus relaciones espaciales. Mediante grafos celulares o nucleares, es posible capturar interacciones locales y patrones organizacionales del tejido, incorporando información topológica que resulta relevante para caracterizar el microambiente tumoral. Estos enfoques han sido utilizados para explorar asociaciones entre estructura tisular y firmas moleculares, así como para construir modelos supervisados capaces de integrar información morfológica y transcriptómica (Xu et al. 2024). Sin embargo, estos métodos se centran principalmente en modelar relaciones espaciales tradicionales entre células, sin incorporar interacciones entre tipos celulares definidos previamente mediante métodos de segmentación nuclear, ni combinar dichas relaciones con representaciones visuales derivadas de modelos fundacionales, lo que ofrece una oportunidad para explorar y utilizar enfoques novedosos que integren ambas fuentes de información.

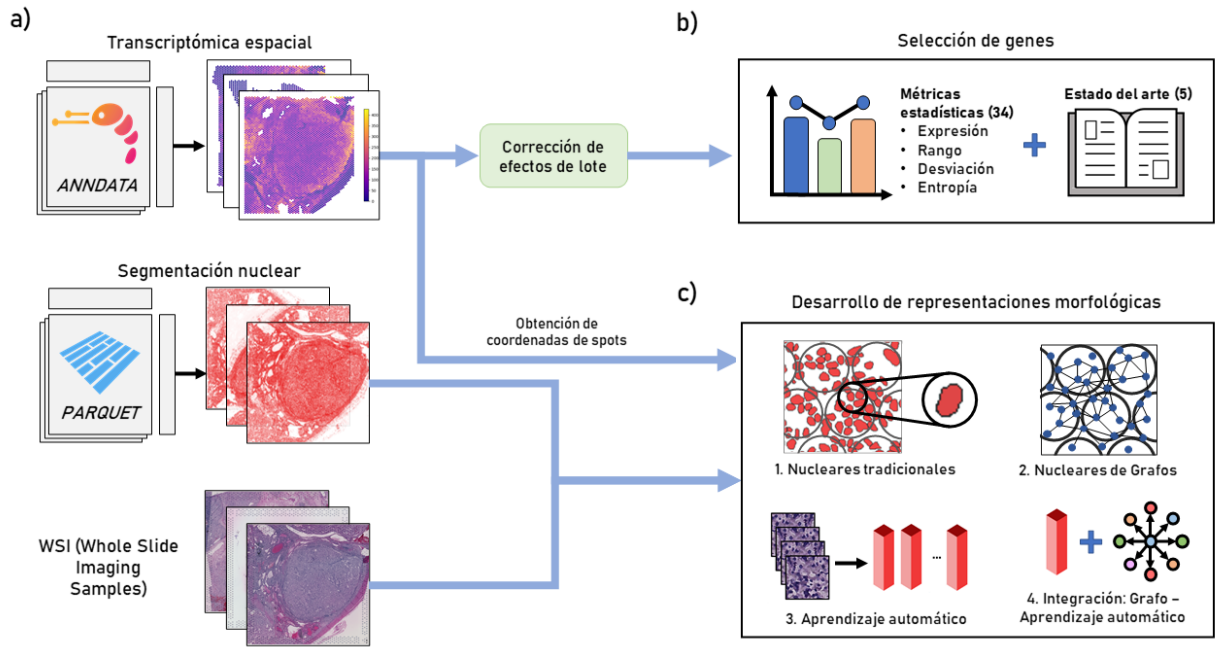
A pesar de la diversidad de representaciones propuestas, persisten desafíos comunes. La alta dimensionalidad de los datos transcriptómicos, combinada con el número reducido de muestras co-registradas, limita la capacidad de generalización de muchos modelos (Zahedi et al. 2024). Asimismo, la interpretabilidad continúa siendo un problema abierto: establecer correlaciones cuantitativas no necesariamente implica comprender los mecanismos biológicos que las sustentan (Rao et al. 2021). En conjunto, el estado del trabajo investigativo en el área muestra una transición progresiva desde descriptores morfológicos individuales hacia representaciones estructuradas y multimodales, en las cuales es cada vez más notoria la necesidad de agregar interpretabilidad como por ejemplo la que se puede tener en interacciones entre diferentes tipos de células/tejidos.

3. Representación Visual Integrada para Muestras de Cáncer Renal

Este capítulo detalla el desarrollo de representaciones morfológicas en muestras de cáncer renal y tejido sano utilizando el conjunto de datos HEST-1K. Inicialmente, la Sección 3.1 expone la organización y selección de las muestras de carcinoma renal de células claras (ccRCC) y tejido sano. La Sección 3.2 describe la identificación de genes relevantes mediante criterios estadísticos y evidencia clínica. Posteriormente, la Sección 3.3 profundiza en la generación de descriptores morfológicos, abarcando desde morfometría nuclear tradicional y grafos celulares hasta embeddings de modelos fundacionales. Finalmente, la Sección 3.4 aborda la integración de estas características a nivel de spot. La Figura 9 ilustra este flujo de trabajo, el cual fundamenta el análisis de correlación con la expresión génica detallado en el capítulo posterior.

Figura 9

Flujo de trabajo para la selección de genes y generación de representaciones morfológicas.



Nota. Flujo de trabajo seguido para el proceso de selección de genes y desarrollo de representaciones morfológicas a partir de la información provista por el dataset HEST-1K. (a) Archivos e información utilizados como base para el flujo de trabajo. (b) Proceso de selección de genes a partir de los objetos de tipo *ANNDATA* y métricas estadísticas. (c) Desarrollo de los cuatro tipos de representaciones a abordar.

3.1. Base de Datos HEST-1K

Para el desarrollo del presente trabajo se utilizó la base de datos HEST-1K, un conjunto de datos de transcriptómica espacial que integra imágenes histológicas completas (WSI)

teñidas con hematoxilina y eosina (H&E). Esta colección reúne 1,276 perfiles de expresión génica alineados con sus imágenes histológicas y la segmentación de núcleos celulares correspondiente para cada muestra, de las cuales 398 corresponden a cáncer perteneciente a 25 tipos tumorales distintos. La colección HEST-1K fue construida a partir de 180 cohortes públicas, incluyendo muestras de 26 órganos y dos especies: *Homo sapiens* (humano) y *Mus musculus* (ratón) (Jaume et al. 2024).

El proyecto HEST incluye una librería de Python llamada *HEST-Library*, la cual está diseñada para facilitar la gestión y el análisis del conjunto de datos explicado previamente. Esta librería utiliza el ecosistema de *scanpy* mediante el formato AnnData, un formato de archivo de Python útil para almacenar datos biológicos de alta dimensión, como matrices de expresión génica de varias células. Entre sus funcionalidades principales se encuentra la descarga selectiva y visualización de muestras del dataset, estandarización de muestras en un formato común y la mitigación de efectos de lote mediante técnicas de reducción de dimensionalidad.

La información relacionada con los perfiles de expresión génica del dataset es generada mediante cuatro tecnologías distintas de transcriptómica espacial, las cuales pueden agruparse en los paradigmas basados en secuenciación y en imágenes, descritos previamente en la subsección 2.5.2. En este contexto, el dataset HEST-1K integra principalmente datos provenientes de tecnologías basadas en secuenciación, como Spatial Transcriptomics v1 (STv1), Visium y Visium HD, así como un menor conjunto de muestras obtenidas mediante tecnologías basadas en imágenes, como Xenium. Un aspecto relevante del dataset es la distribución desigual de muestras entre tecnologías, con una clara predominancia de aquellas basadas en secuenciación, particularmente Visium, lo cual influye tanto en la resolución espacial disponible como en el tipo de análisis que puede realizarse. Esta distribución se detalla a continuación.

1. **Visium:** 602 muestras, lo que representa el 49.0 % del total del dataset.
2. **STv1 (Spatial Transcriptomics v1):** 552 muestras, lo que representa el 44.9 % del total del dataset.
3. **Xenium:** 65 muestras, lo que representa el 5.3 % del total del dataset.
4. **Visium HD:** 10 muestras, lo que representa el 0.8 % del total del dataset.

3.1.1. Datos Asociados a Cáncer Renal

Para el estudio y análisis de cáncer renal se seleccionó un subconjunto de muestras del dataset HEST-1K correspondiente a tejido de riñón humano con su enfermedad clasificada

como cáncer. Asimismo, se utilizaron muestras generadas mediante la tecnología Visium, debido a su mayor disponibilidad dentro del dataset y a su cobertura de transcriptoma completo.

Mediante estos criterios de selección se obtuvieron 24 muestras de carcinoma renal de células claras (ccRCC) provenientes de una de las cohortes internas del proyecto HEST, la cual se trata de un proyecto realizado por Meylan et al. enfocado en el análisis de las estructuras linfoides terciarias (TLS) y su papel en la respuesta inmunológica dentro del tumor de cáncer renal. Las TLS son agregados de células del sistema inmunológico que se forman dentro de algunos tumores o tejidos con inflamación anormal, los cuales se encargan de la producción de células B (un tipo de linfocito del sistema inmunitario encargado de producir anticuerpos) y su maduración hacia células plasmáticas dentro del microambiente tumoral (Meylan et al. 2022).

De las 24 muestras provistas por el estudio mencionado previamente, la mitad son muestras congeladas y la otra mitad son corresponden a muestras FFPE (*Formalin-Fixed Paraffin-Embedded*). Las muestras congeladas se preservan mediante congelación inmediata del tejido para mantener su estado molecular original, mientras que el método mediante FFPE se fijan químicamente con formalina y luego se incluyen en bloques de parafina para su conservación (Gaffney et al. 2018). El estudio incluye ambos métodos con la intención de validar que los patrones espaciales del tejido pueden identificarse de forma consistente independientemente del método de preservación utilizado.

Durante el proceso de recolección de conjuntos de datos para la creación y desarrollo del proyecto HEST-1K se asignaron identificadores únicos a cada muestra, con el propósito de realizar una selección de datos de forma más organizada. En el caso de las 24 muestras de cáncer renal, estos identificadores se definen con las letras INT y los números del 1 al 24; es decir, las muestras corresponden desde INT1 hasta INT24.

3.1.2. Datos Asociados a Riñón Sano

Para el desarrollo de experimentos complementarios se seleccionó otro subconjunto de muestras dentro de la base de datos HEST, utilizando los mismos parámetros de tecnología Visium y la especie *Homo sapiens*, cambiando únicamente el estado de la enfermedad a una clasificación de sano o saludable.

De esta manera, se lograron obtener 12 muestras con los criterios de selección mencionados previamente. Sin embargo, la totalidad de estas muestras no proviene de una única fuente u organización, como sucede con las muestras de cáncer renal. En este caso, se requieren dos fuentes principales, las cuales son:

1. **10x Genomics Portal:** Provee una muestra de riñón sano con su transcriptómica espacial mediante la tecnología Visium. Esta muestra proviene de la publicación titulada *"Human Kidney, 11 mm Capture Area (FFPE)"* (10x Genomics 2022) y su identificador dentro del dataset HEST-1K se define como: TENX71.

2. **NCBI (National Center for Biotechnology Information):** Provee 11 muestras que cumplen con los criterios de selección mencionados. Dichas muestras provienen de cuatro publicaciones diferentes, las cuales son:

   - **Genome-wide spatial expression profiling in formalin-fixed tissues:** Provee tres (3) muestras con los identificadores únicos: MEND54, MEND49 y MEND48 (Gracia Villacampa et al. 2021).
   - **Spatial localization with Spatial Transcriptomics for an atlas of healthy and injured cell states and niches in the human kidney [Visium ST]:** Provee seis (6) muestras con los identificadores únicos: NCBI709, NCBI710, NCBI711, NCBI712, NCBI713 y NCBI714 (Lake et al. 2023).
   - **Integration of spatial and single cell transcriptomics localizes epithelial-immune cross-talk in kidney injury:** Provee una (1) muestra con el identificador único: NCBI599 (R. M. Ferreira et al. 2021).
   - **A Spatial Transcriptomic atlas of the human kidney papilla identifies significant immune injury in patients with stone disease:** Provee una (1) muestra con el identificador único: NCBI568 (Canela et al. 2023).

3.2. Selección de Genes Mediante Análisis Estadísticos y Estado del Arte

Con el propósito de facilitar la identificación de genes que puedan tener relación con el cáncer renal, se realizó un proceso de seleción de genes mediante diferentes análisis estadísticos que sirven para observar el comportamiento de dichos genes dentro del tejido.

3.2.1. Corrección de Efectos de Lote

Previo al desarrollo de los análisis estadísticos de selección, se llevo a cabo la mitigación de efectos de lote (conocidos en inglés como *batch effects*), el cuál hace referencia a variaciones no deseadas introducidas por factores de confusión que no están relacionados con el factor biológico de interés (Yiwen Wang y LêCao 2020). Dentro de las principales causas de este efecto se encuentran:

- **Factores biológicos:** Son variaciones naturales ajenas al objetivo del estudio, como la edad, el sexo o la dieta.
- **Factores técnicos:** Incluyen artefactos introducidos durante la manipulación de las muestras, tales como diferencias en la recolección, el almacenamiento o los protocolos de extracción de información genética.
- **Factores computacionales:** Son aquellos que suceden durante el procesamiento y análisis de datos, tales como el software utilizado o los parámetros seleccionados en los métodos estadísticos.

Para mitigar este problema se realizó un proceso de corrección de efectos de lote utilizando la función `correct_batch_effect` de la librería *HEST-Library* mencionada en la sección 3.1. Este procedimiento se realizó sobre la expresión génica de cada muestra almacenada en formato `AnnData` (`.h5ad`), empleando el método *ComBat*, un algoritmo ampliamente utilizado para la normalización de datos de expresión génica.

El algoritmo ComBat modela los efectos de lote mediante un modelo lineal que estima parámetros para las diferencias en la media y la varianza entre los lotes de datos. Este método ajusta los valores de expresión mediante una estandarización que elimina las variaciones de cada lote, manteniendo al mismo tiempo la variabilidad biológica relevante (Y. Zhang et al. 2020). Luego de aplicar la corrección, la expresión de cada muestra fue almacenada nuevamente en formato `.h5ad`, para ser utilizada en estudios posteriores. Para evaluar visualmente el efecto de la corrección de lote, se generaron representaciones mediante reducción de dimensionalidad utilizando *UMAP* (*Uniform Manifold Approximation and Projection*), la cual es una técnica de reducción de dimensionalidad no lineal que permite proyectar datos de alta dimensión en un espacio bidimensional o tridimensional, preservando en la medida de lo posible la estructura local de los datos.

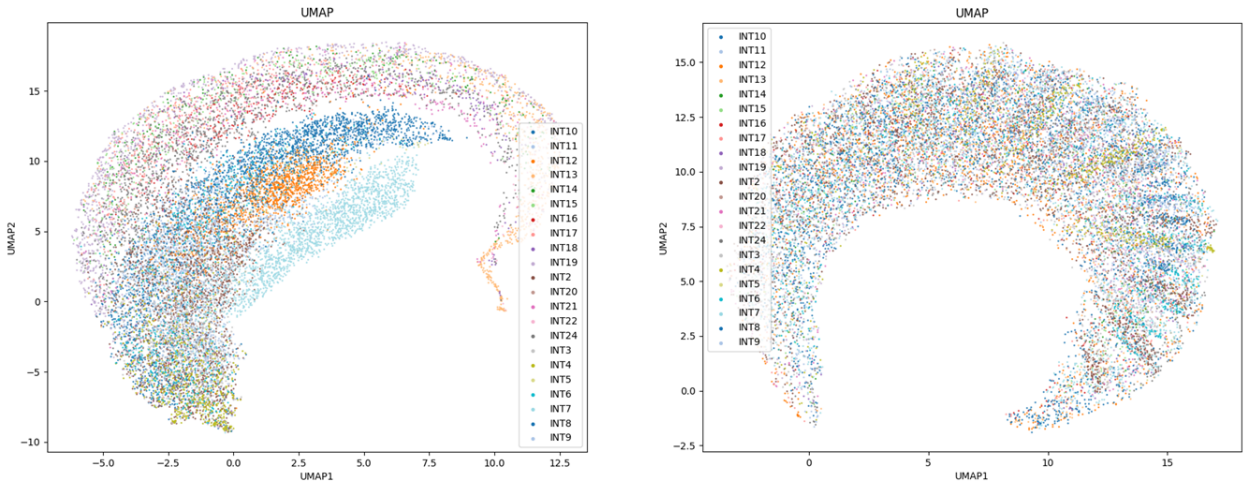
No obstante, durante este proceso se identificaron inconvenientes en dos muestras específicas, `INT1` y `INT23`, las cuales no pudieron ser procesadas correctamente en el proceso de corrección. En el caso de la muestra `INT1`, se presentó una excepción de tipo *TopologyException* asociada a la librería geométrica utilizada internamente, indicando inconsistencias en las geometrías de los contornos del tejido (por ejemplo, intersecciones inválidas o regiones mal definidas). Este error impide realizar operaciones espaciales necesarias para la identificación de regiones válidas dentro del tejido. Por otro lado, la muestra `INT23` no cumplió con el umbral mínimo requerido de regiones estromales, detectándose menos de cinco parches estromales, lo cual activa una restricción interna del método e impide continuar con el procesamiento. Debido a estas limitaciones, ambas muestras fueron excluidas de este subpro-

ceso de corrección de efectos de lote, garantizando así la consistencia y validez de los datos utilizados en los análisis posteriores.

La Figura 10 muestra una visualización mediante UMAP de las muestras de cáncer renal antes y después de aplicar la corrección de efecto de lote mediante la función mencionada previamente. En la subfigura 10a se observa la distribución de los datos antes de la corrección, donde las muestras tienden a agruparse según el lote experimental. En la subfigura 10b se presenta la misma visualización después de aplicar el método ComBat, donde se observa una mayor integración entre las muestras.

Figura 10

Visualización del efecto de lote en muestras de carcinoma renal mediante UMAP.



(a) *Antes de la corrección* **(b)** *Después de la corrección*

Nota. Proyección UMAP de los perfiles de expresión génica de muestras tumorales de ccRCC. Los puntos están coloreados por paciente o lote de secuenciación. La comparación evidencia la mitigación de variaciones técnicas no biológicas tras aplicar el algoritmo de corrección. Datos de HEST-1K (Jaume et al. 2024).

3.2.2. Obtención del Conjunto Común de Genes

Debido a que las distintas muestras pueden contener diferentes subconjuntos de genes medidos durante el procesamiento inicial, se identificó el conjunto de genes presentes en todas las muestras utilizadas. Para ello se calcularon las intersecciones entre los conjuntos de genes para cada archivo `.h5ad`. Este proceso consiste en identificar los genes que aparecen simultáneamente en todas las muestras, garantizando que las comparaciones estadísticas posteriores se realicen sobre un mismo conjunto.

Formalmente, si G_i representa el conjunto de genes presentes en la muestra i, el

conjunto de genes comunes $G_{común}$ se define como:

$$G_{común} = \bigcap_{i=1}^{n} G_i \tag{3.1}$$

donde n corresponde al número total de muestras analizadas. El resultado de este proceso fue un subconjunto de 17.943 genes en común entre las 24 muestras.

3.2.3. Cálculo de Métricas Estadísticas de Expresión Génica

Utilizando el conjunto común de genes, se calcularon diversas métricas estadísticas para caracterizar los patrones de expresión génica a lo largo de todas las muestras y spots espaciales, lo que permite identificar genes que presentan variabilidad o un comportamiento relevante dentro del tejido. Previo al cálculo de las métricas, es importante recordar que la expresión génica se encuentra cuantificada a nivel de *spot* en términos de *conteos crudos de transcritos* (*raw counts*), es decir, el número de moléculas de ARN detectadas por gen en cada ubicación espacial. En base a esto, las métricas descritas a continuación se calculan a partir de estos conteos, considerando inicialmente los valores por *spot* dentro de cada muestra y posteriormente agregándolos a nivel global.

Para cada gen se calcularon las siguientes métricas:

1. **Expresión promedio**

   La expresión promedio de cada gen se calculó como la media de los valores de expresión a lo largo de todos los spots espaciales de una muestra. Sea x_{ij} la expresión del gen j en el spot i, y N el número total de spots. La expresión promedio se define como:

   $$\mu_j = \frac{1}{N} \sum_{i=1}^{N} x_{ij} \tag{3.2}$$

   Mediante esta métrica se puede identificar genes con altos niveles de expresión global dentro del tejido.

2. **Rango dinámico**

   El rango dinámico mide la amplitud de variación en los valores de expresión de un gen dentro del tejido. Se calcula como la diferencia entre el valor máximo y el mínimo observado para cada gen:

   $$RD_j = \text{máx}(x_{ij}) - \text{mín}(x_{ij}) \tag{3.3}$$

Un valor alto de rango dinámico indica que el gen presenta variaciones significativas de expresión entre distintas regiones del tejido.

3. **Desviación estándar** La desviación estándar cuantifica la dispersión de los valores de expresión de un gen alrededor de su media.

   Se define como:

$$\sigma_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{ij} - \mu_j)^2} \tag{3.4}$$

   Los genes con mayor desviación estándar presentan mayor variabilidad espacial en su expresión.

4. **Entropía de Shannon** La entropía de Shannon como una medida de diversidad en la distribución de los niveles de expresión de cada gen. Esta métrica se define como:

$$H = -\sum_{i=1}^{k} p_i \log_2(p_i) \tag{3.5}$$

   donde p_i representa la probabilidad asociada a cada nivel de expresión observado.

   Para facilitar la comparación entre genes, la entropía fue normalizada respecto a su valor máximo posible:

$$H_{norm} = \frac{H}{\log_2(k)} \tag{3.6}$$

   Los genes con mayor entropía presentan distribuciones de expresión más heterogéneas dentro del tejido.

Las métricas se calcularon inicialmente para cada muestra de manera individual y posteriormente se promediaron entre todas las muestras.

3.2.4. Selección de Genes Mediante Ranking Estadístico

A partir de las métricas calculadas se realizó un proceso de ranking para identificar los genes con mayor relevancia estadística. Para cada métrica se ordenaron los genes de mayor a menor valor y se seleccionaron los 20 genes con valores más altos. Posteriormente se combinaron los resultados de las distintas métricas, eliminando los genes duplicados entre las cuatro listas obtenidas para cada métrica.

Como resultado se obtuvo un conjunto final de 34 genes, los cuales presentan altos niveles de expresión, variabilidad o diversidad de distribución dentro de las muestras analizadas. Esta lista de genes se puede apreciar en la tabla 1.

Tabla 1: *Genes de interés seleccionados mediante análisis estadístico*

ACTB	CD74	FTH1	IGKC	RACK1
B2M	COL1A1	FTL	IGLC1	SERF2
CD63	CTSD	HLA-A	IGHM	SPARC
EEF1G	EEF2	HLA-DRA	IGHG1	TMSB4X
ENO1	FN1	HSPB1	IGHG2	UBC
IGFBP7	IGHA1	IGHG3	UBA52	VIM
ITM2B	IGHG4	MT2A	PABPC1	

3.2.5. Selección de Genes Basada en el Estado del Arte

Además del proceso de selección de genes basado en las cuatro métricas estadísticas explicadas previamente, se consideró un segundo conjunto de genes reportados en la literatura científica como relevantes en el contexto del cáncer renal. A partir de una revisión del estado del arte se identificaron cinco genes relacionados con procesos clave dentro del microambiente tumoral renal, particularmente en el carcinoma renal de células claras (ccRCC), los cuales se presentan en la tabla 2.

En particular, el gen *VHL* es ampliamente reconocido como un evento temprano en la tumorigénesis del ccRCC, con mutaciones presentes en la mayoría de los casos y asociadas a la activación de la vía HIF y procesos de angiogénesis (Linehan 2012; Xie et al. 2025). Por su parte, *PBRM1*, *SETD2* y *BAP1* son genes localizados en el cromosoma 3p y frecuentemente co-mutados en ccRCC, desempeñando un papel fundamental en la remodelación de cromatina, regulación epigenética y progresión tumoral (Millan et al. 2025; Yongquan Wang et al. 2023). En particular, mutaciones en *BAP1* se han asociado con fenotipos más agresivos y peor pronóstico, mientras que *SETD2* y *PBRM1* están implicados en la estabilidad genómica y la respuesta a terapias dirigidas (Xie et al. 2025). Finalmente, el gen *MET*, aunque menos frecuente en ccRCC, se encuentra asociado principalmente al carcinoma renal papilar y a procesos de señalización oncogénica relacionados con proliferación celular (Linehan 2012; PDQ Cancer Genetics Editorial Board 2002).

Tabla 2: *Genes de interés seleccionados mediante el estado del arte de cáncer renal*

BAP1	MET	PBRM1	VHL	SETD2

Estos genes fueron incluidos de manera complementaria al conjunto previamente seleccionado mediante análisis estadístico, obteniendo de esta manera un listado final de 39 genes de interés para analizar.

3.3. Desarrollo de Representaciones

En esta sección se describen las estrategias utilizadas en el desarrollo de representaciones a partir de la información histológica obtenida mediante las muestras de tejido renal. Mediante estas representaciones se busca capturar diferentes propiedades del tejido, incluyendo características morfológicas clásicas mediante los núcleos celulares y patrones más complejos mediante estructuras basadas en grafos y herramientas de aprendizaje automático. El objetivo es extraer descriptores que permitan encontrar relaciones significativas con la expresión génica obtenida via transcriptómica espacial.

Para el desarrollo de estas representaciones basadas en morfología nuclear y aprendizaje automático se utilizaron tres tipos de información provistas por el dataset HEST-1K:

- **Segmentación nuclear:** Presente en formato *.parquet* (un formato de almacenamiento columnar), esta información fue generada por los autores del dataset HEST-1K mediante el modelo de aprendizaje profundo CellViT, el cual se basa en una arquitectura *Vision Transformer* para la segmentación automatizada de instancias de núcleos celulares (Hörst et al. 2024). Esta segmentación se utiliza principalmente para el desarrollo de representaciones nucleares tradicionales y basadas en grafos, como se describe en las subsecciones 3.3.1 y 3.3.2.

- **Imágenes de portaobjetos completas (WSIs):** Presentes en formato *.tiff* (Tagged Image File Format), utilizadas principalmente para el desarrollo de representaciones mediante modelos fundacionales, mencionados en la subsección 3.3.3.

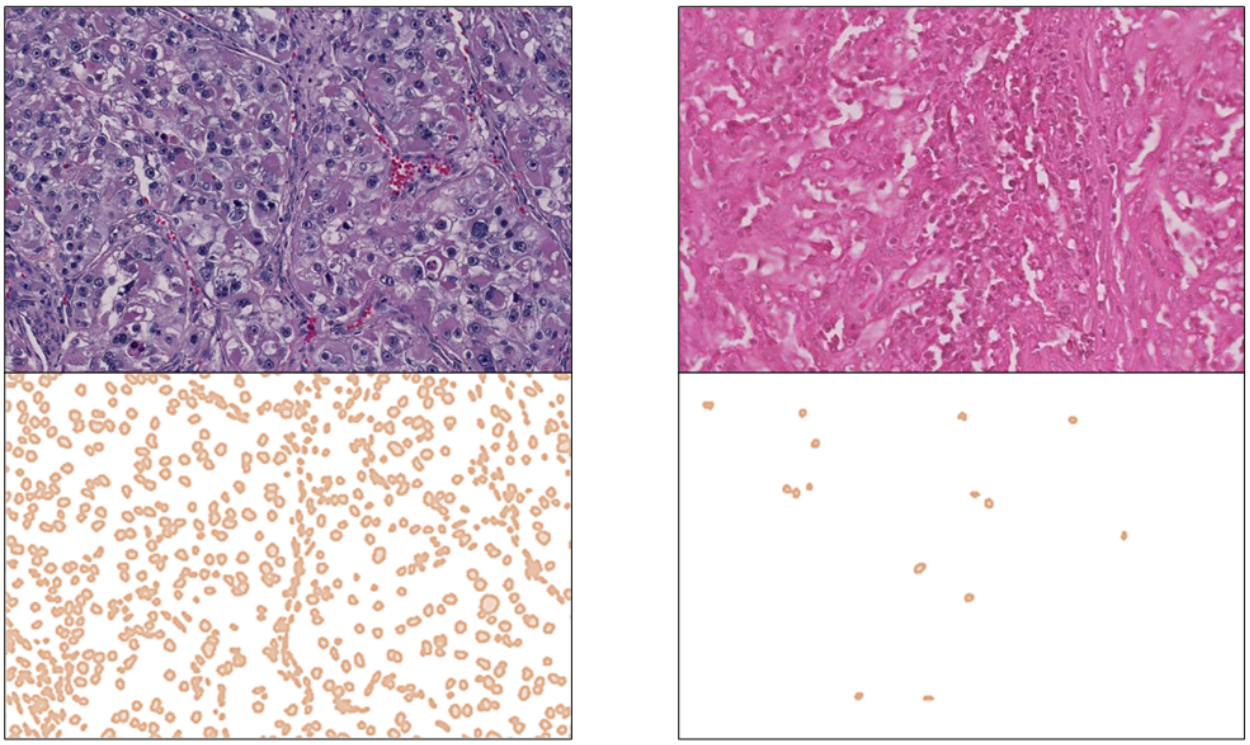
- **Contenedores de parches curados:** Disponibles en formato *.h5* (Hierarchical Data Format), estos archivos almacenan directamente los parches extraídos de las WSIs, evitando su generación manual. Cada contenedor organiza la información en *datasets* clave: el tensor `img` ($N \times 224 \times 224 \times 3$), que contiene los parches en formato RGB compatibles con modelos fundacionales, y la matriz `coords` ($N \times 2$), que registra sus coordenadas espaciales (x, y). Esta estructura sirve como base para las representaciones descritas en la subsección 3.3.3.

- **Expresión génica y transcriptómica espacial:** Mediante la librería Scanpy como base, se emplea el formato AnnData para almacenar matrices de expresión de forma eficiente, el cual se guarda físicamente con la extensión *.h5ad*. Esta información se utiliza en los procesos de correlación y regresión entre las representaciones empleadas y los *spots* de expresión génica, los cuales se describen en el capítulo 4.

Inicialmente se disponía de 24 muestras de tejido de cáncer renal mencionadas en la subsección 3.1.1. Tras evaluar la calidad de las imágenes, se excluyeron INT20, INT22 e INT23 debido a problemas de tinción, especialmente por bajos niveles de hematoxilina, lo que afecta la visibilidad de los núcleos y el desempeño de la segmentación. La Figura 11 ilustra estas diferencias. En consecuencia, el análisis se realizó sobre 21 muestras.

Figura 11

Comparativa de intensidad de tinción en muestras histopatológicas.



(a) *Tinción adecuada* **(b)** *Baja intensidad de tinción*

Nota. La muestra (a) presenta núcleos celulares bien definidos debido a una adecuada intensidad de hematoxilina, mientras que en (b) la degradación visual dificulta la segmentación nuclear y el análisis morfológico. Ambas muestras provienen de HEST-1K (Jaume et al. 2024).

A partir de los diferentes tipos de descriptores morfológicos obtenidos mediante estas representaciones, se lleva a cabo un análisis de correlación entre estas características y la expresión génica, proceso el cual es descrito en el capitulo 4.

3.3.1. Representaciones Nucleares Tradicionales

Las características nucleares extraídas mediante este tipo de representación corresponden a un subconjunto de las descritas en la subsección 2.4.1. En este caso, se utilizaron únicamente características morfológicas, las cuales se explican en dicha sección. En este proceso se utilizó la segmentación nuclear generada por CellViT, previamente descrita en la sección 3.3, donde cada núcleo se representa como una geometría poligonal. Para este trabajo se consideraron siete características principales descritas a continuación e ilustradas visualmente mediante la figura 12:

- **Área (*area*):** Corresponde al área del polígono que representa el núcleo segmentado. Se calcula directamente como:

$$\text{area} = \text{Área}(\text{geometría}) \tag{3.7}$$

  Se expresa en píxeles cuadrados (px^2) y toma valores positivos. Esta característica cuantifica el tamaño del núcleo; valores elevados pueden estar asociados a procesos proliferativos o alteraciones morfológicas.

- **Perímetro (*perimeter*):** Longitud del contorno del núcleo:

$$\text{perimeter} = \text{Longitud}(\text{borde de la geometría}) \tag{3.8}$$

  Se mide en píxeles (px) y está directamente relacionado con el tamaño y la complejidad del contorno. Perímetros relativamente altos, en relación con el área, pueden indicar bordes irregulares o estructuras deformadas.

- **Circularidad (*circularity*):** Mide qué tan cercano es el núcleo a una forma circular perfecta:

$$\text{circularity} = \frac{4\pi \cdot \text{area}}{\text{perimeter}^2} \tag{3.9}$$

  Es una medida adimensional con valores en el rango $(0, 1]$. Un valor cercano a 1 indica una forma circular ideal, mientras que valores menores reflejan elongación o irregularidad en el contorno.

- **Solidez (*solidity*):** Relación entre el área del núcleo y el área de su envolvente convexa:

$$\text{solidity} = \frac{\text{area}}{\text{convex\_area}} \tag{3.10}$$

  También es adimensional y toma valores en el rango $(0, 1]$. Valores cercanos a 1 in-

dican núcleos compactos y sin concavidades, mientras que valores más bajos reflejan contornos con irregularidades, muescas o estructuras fragmentadas.

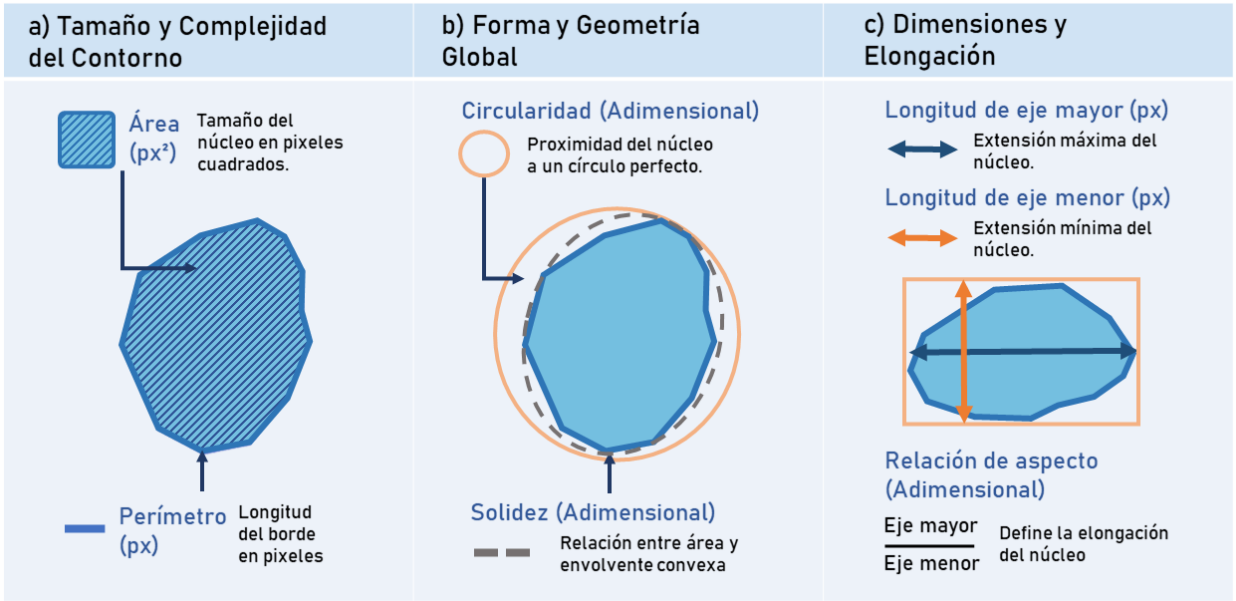
- **Longitud del eje mayor (*major_axis_length*):** Corresponde a la mayor dimensión del rectángulo mínimo rotado que contiene al núcleo, el cual aproxima la orientación principal de la geometría segmentada. Se mide en píxeles y describe la extensión máxima del núcleo.
- **Longitud del eje menor (*minor_axis_length*):** Corresponde a la menor dimensión de dicho rectángulo. También se expresa en píxeles y, en conjunto con el eje mayor, permite caracterizar la forma global del núcleo.
- **Relación de aspecto (*aspect_ratio*):** Define la elongación del núcleo como:

$$\text{aspect\_ratio} = \frac{\text{major\_axis\_length}}{\text{minor\_axis\_length}} \tag{3.11}$$

Es una medida adimensional con valores mayores o iguales a 1. Un valor cercano a 1 indica una forma aproximadamente circular, mientras que valores mayores reflejan núcleos elongados o anisotrópicos.

Figura 12

Clasificación de características morfológicas tradicionales de núcleos celulares.



Nota. Métricas categorizadas en: (a) complejidad del contorno y dimensiones de área; (b) descriptores de geometría global y circularidad; y (c) factores de elongación y excentricidad del núcleo segmentado.

Todas las características geométricas fueron calculadas en el espacio de coordenadas de la imagen, por lo que sus unidades dependen de la resolución espacial del escaneo. En el dataset HEST-1K, esta resolución se encuentra aproximadamente en el orden de 0.45 μm por píxel, lo que permite convertir las magnitudes a unidades físicas. Por ejemplo:

$$\text{Área}(\mu m^2) = \text{Área}(px^2) \times (\text{pixel\_size})^2 \tag{3.12}$$

Por otro lado, las características como la circularidad, la solidez y la relación de aspecto son adimensionales, por lo que no dependen de la escala de la imagen y resultan directamente comparables entre muestras con diferentes resoluciones. Estas características fueron seleccionadas debido a su interpretabilidad y a su amplio uso en el análisis morfológico de tejidos, ya que permiten capturar propiedades geométricas fundamentales de núcleos celulares de forma robusta e independiente de la orientación espacial de los mismos. Cabe resaltar que en casos donde algunas magnitudes geométricas toman valores nulos o generan inestabilidades numéricas (por ejemplo, divisiones por cero), las características correspondientes se tratan como valores no definidos, con el fin de garantizar la estabilidad del proceso de extracción.

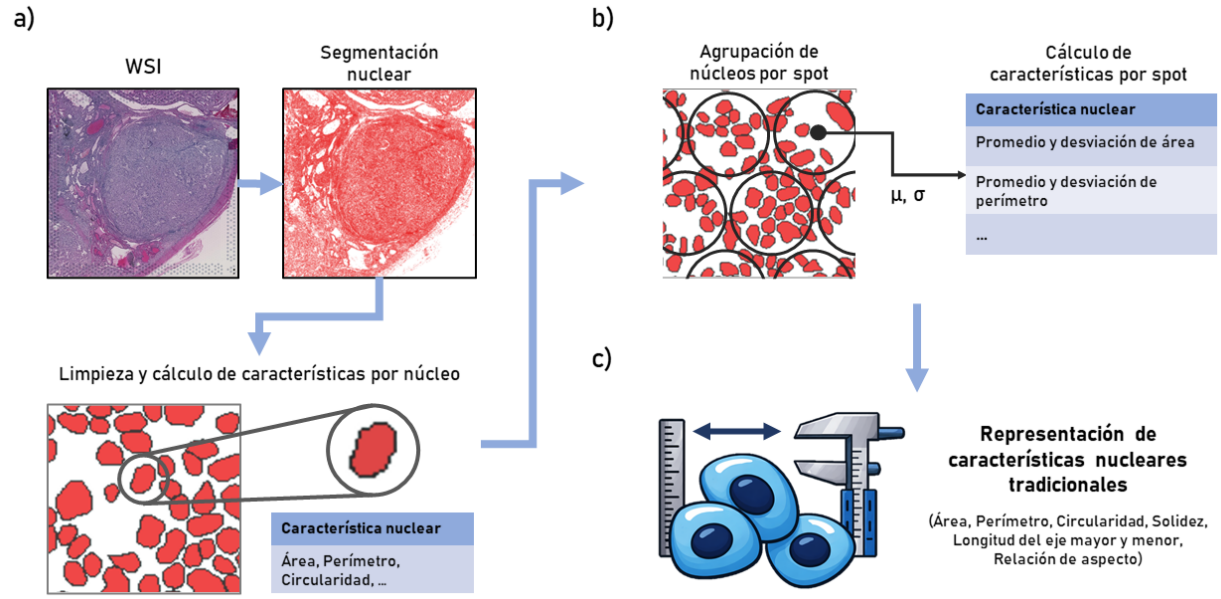
Una vez obtenidas las características a nivel de núcleo individual, se realizó un proceso de agregación espacial con el fin de construir descriptores a nivel de región. Para ello, cada núcleo fue asignado al *spot* más cercano, utilizando como referencia las coordenadas espaciales de los *spots* de transcriptómica contenidas en los archivos *.h5ad*. Esta asignación se llevó a cabo mediante el método de *Nearest Neighbors*, el cual consiste en identificar, para cada punto en el espacio, el elemento más cercano dentro de un conjunto de referencia según una métrica de distancia definida. En este caso, se utilizó la distancia euclidiana bidimensional entre el centroide de cada núcleo y las coordenadas de los *spots*, definida como:

$$d(n_i, s_j) = \sqrt{(x_{n_i} - x_{s_j})^2 + (y_{n_i} - y_{s_j})^2} \tag{3.13}$$

donde (x_{n_i}, y_{n_i}) corresponden a las coordenadas del centroide del núcleo n_i, y (x_{s_j}, y_{s_j}) a las coordenadas del *spot* s_j. De esta forma, para cada núcleo se selecciona el *spot* que minimiza dicha distancia, estableciendo una correspondencia espacial entre la información morfológica y la transcriptómica. Posteriormente, para cada *spot*, se calcularon dos estadísticas resumen por cada característica: la media y la desviación estándar, lo que permite capturar tanto el valor promedio como la variabilidad de las propiedades nucleares dentro de cada región espacial, generando así una representación compacta y robusta de la morfología tisular a nivel de *spot*. El proceso descrito previamente se ilustra mediante la figura 13 mostrada a continuación.

Figura 13

Flujo de trabajo para la obtención de representaciones nucleares tradicionales.



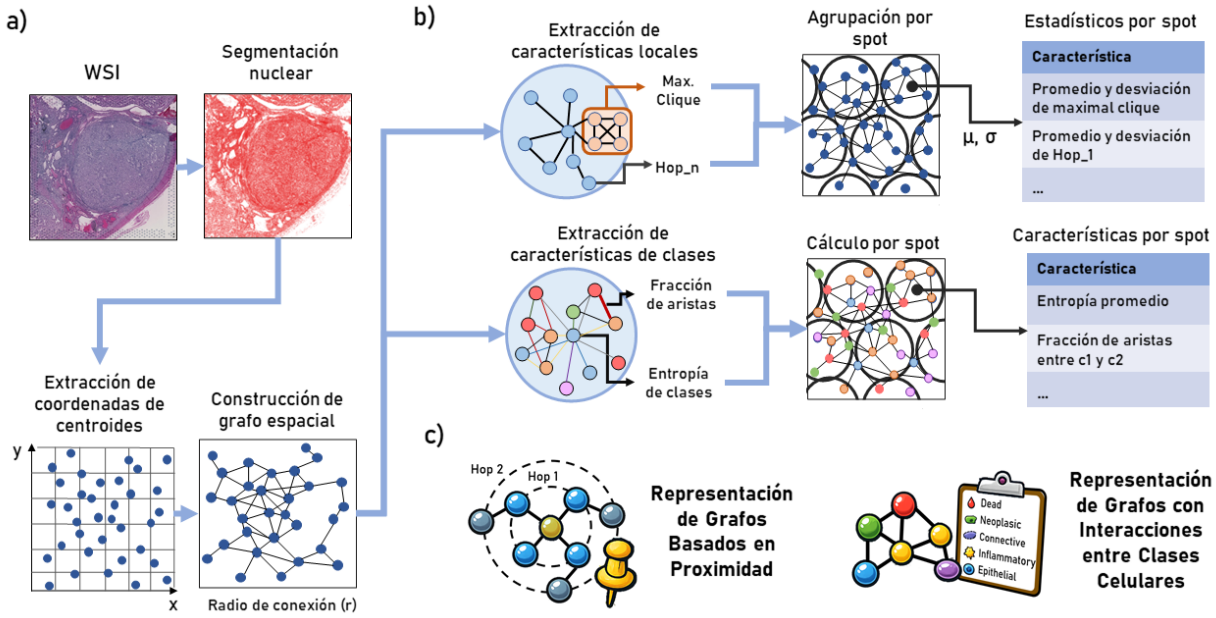
Nota. El proceso integra: (a) segmentación y extracción de descriptores mediante CellViT; (b) agregación espacial de características por *spot* utilizando centroides para el cálculo de métricas de tendencia central y dispersión; y (c) consolidación de estructuras de datos para el análisis de correlación morfogenómica.

3.3.2. Representaciones Nucleares Basadas en Grafos

Para el desarrollo de este tipo de representaciones se abarcaron descriptores presentes en grafos de proximidad y de relaciones espaciales, descritos previamente en la sección 2.4.2, esto con el objetivo de capturar relaciones entre núcleos celulares que no pueden ser descritas únicamente mediante características individuales. En esta representación se abordaron dos enfoques complementarios: (i) grafos simples basados únicamente en la proximidad espacial entre núcleos, y (ii) grafos que incorporan información de clases celulares, permitiendo modelar interacciones entre diferentes tipos de núcleos. La Figura 14 presenta una visión general del proceso llevado a cabo para el desarrollo y análisis de los dos tipos de características basadas en grafos.

Figura 14

Flujo de trabajo para la obtención de representaciones nucleares de grafos.



Nota. El procedimiento incluye: (a) construcción de grafos espaciales mediante centroides segmentados con CellViT; (b) cálculo de métricas de interacción local por *spot*, incluyendo entropía y topología de aristas; y (c) consolidación de datos para la evaluación de correlación morfogenómica.

3.3.2.1. Grafos Basados en Proximidad: Maximal Clique y Hop-N Neighbors. En este enfoque, cada núcleo celular segmentado se modela como un nodo dentro de un grafo no dirigido, donde las aristas representan relaciones de vecindad espacial entre núcleos. La construcción del grafo se realiza a partir de los centroides de los núcleos obtenidos de la segmentación CellViT.

Sea $\{x_i\}_{i=1}^{N}$ el conjunto de centroides de los núcleos en coordenadas espaciales. Se define un grafo $G = (V, E)$ donde cada nodo $v_i \in V$ corresponde a un núcleo, y existe una arista $(v_i, v_j) \in E$ si la distancia euclidiana entre sus centroides es menor que un umbral r:

$$\|x_i - x_j\|_2 \leq r \tag{3.14}$$

En este trabajo se utilizó un radio de conexión $r = 50$, expresado en las mismas unidades espaciales de las coordenadas de la imagen (píxeles en el sistema de referencia original de las WSIs). Este valor fue seleccionado como un compromiso entre capturar interacciones locales relevantes entre núcleos y evitar la sobreconectividad del grafo. Radios menores tienden a generar grafos demasiado dispersos, limitando la capacidad de capturar la estructura del

tejido, mientras que radios mayores pueden inducir conexiones entre núcleos espacialmente no relacionados, diluyendo la información local. De esta manera, $r = 50$ permite modelar adecuadamente la organización espacial a escala celular dentro del tejido.

Para la construcción eficiente de este grafo se utilizó una estructura de tipo *KDTree*, la cual permite identificar pares de nodos vecinos dentro del radio r de manera computacionalmente eficiente. Una vez construido el grafo, se extrajeron características a nivel local (por nodo) enfocadas en capturar la organización espacial y conectividad de los núcleos, las cuales son descritas a continuación e ilustradas en la figura 15:

- **Maximal Clique (*max_clique_size*):** Una clique es un subconjunto de nodos donde todos están conectados entre sí. Para cada nodo, se calcula el tamaño de la clique máxima a la que pertenece. Esta característica captura la densidad local de agrupamiento; valores altos indican regiones con alta conectividad entre núcleos.

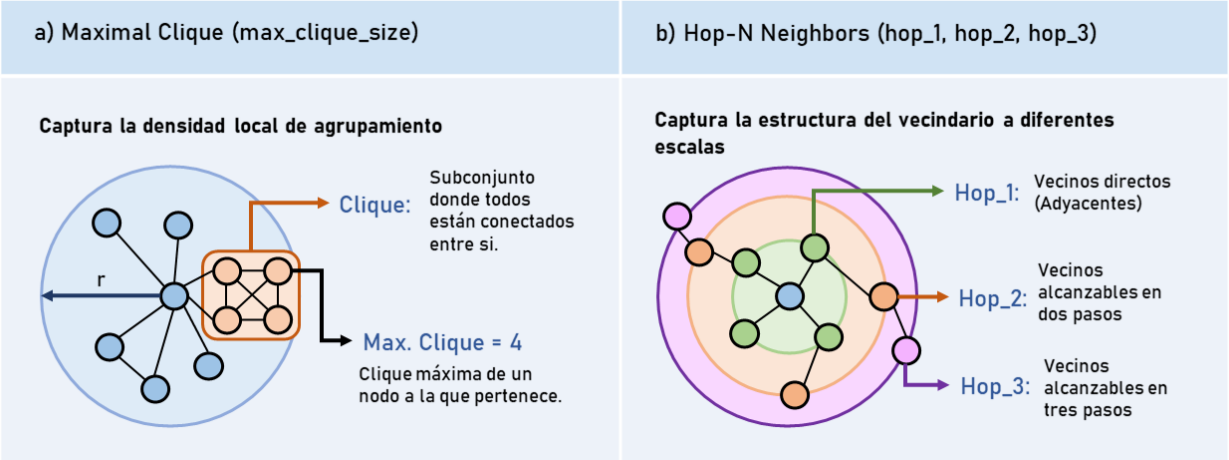
- **Hop-N neighbors (*hop_1, hop_2, hop_3*):** Para cada nodo, se calcula el número de vecinos alcanzables a N saltos en el grafo. En este trabajo se consideraron $N = 1, 2, 3$:

  - *hop_1*: vecinos directos (adyacentes)
  - *hop_2*: vecinos alcanzables en dos pasos
  - *hop_3*: vecinos alcanzables en tres pasos

  Estas métricas permiten capturar la estructura del vecindario a diferentes escalas espaciales.

Figura 15

Representación de características espaciales basadas en grafos de proximidad.



Nota. Los descriptores topológicos incluyen: (a) el cálculo del *maximal clique* por nodo para cuantificar la densidad de agrupamiento celular; y (b) el análisis de vecindarios *Hop-N* para caracterizar la conectividad y estructura local de la red.

De esta manera se logra describir cada nodo del grafo por un conjunto de características locales junto con su posición espacial (centroide).

Posteriormente, de manera análoga a las representaciones nucleares tradicionales, se realizó un proceso de agregación espacial para obtener descriptores a nivel de *spot*. Para ello, cada nodo (núcleo) fue asignado al *spot* más cercano utilizando las coordenadas espaciales provenientes de los datos de transcriptómica (*.h5ad*). Esta asignación se realizó siguiendo el mismo esquema de *Nearest Neighbors* descrito en la subsección 3.3.1, empleando la distancia euclidiana en el espacio bidimensional para identificar el *spot* más cercano a cada nodo. Una vez realizada esta asignación, se calcularon, para cada *spot*, la media y la desviación estándar de cada una de las características de grafo consideradas (*max_clique_size*, *hop_1*, *hop_2*, *hop_3*), permitiendo de esta manera resumir tanto la conectividad promedio como la variabilidad estructural de los núcleos dentro de cada región espacial.

3.3.2.2. Grafos con Interacciones entre Clases Celulares. Además de las características estructurales basadas en la conectividad y proximidad nuclear, se desarrolló una segunda representación de grafos que incorpora información sobre el tipo celular de cada núcleo. En este caso, cada nodo del grafo no solo representa un núcleo, sino que además está asociado a una etiqueta de clase celular, proveniente de la segmentación realizada sobre el dataset HEST-1K mediante el modelo CellViT. En particular, se consideran cinco clases celulares principales:

- **Connective:** Corresponde a células del tejido conectivo, las cuales cumplen funciones de soporte estructural y organización del tejido.
- **Epithelial:** Incluye células epiteliales, responsables de recubrir superficies y formar estructuras glandulares.
- **Neoplastic:** Representa células tumorales o neoplásicas, caracterizadas por un crecimiento descontrolado y alteraciones morfológicas significativas.
- **Inflammatory:** Agrupa células del sistema inmune presentes en procesos inflamatorios, como linfocitos y macrófagos, asociadas a la respuesta inmunitaria del tejido.
- **Dead:** Corresponde a núcleos de células no viables o en proceso de muerte celular, los cuales pueden reflejar necrosis o daño tisular.

La construcción del grafo sigue el mismo principio basado en proximidad espacial descrito anteriormente, donde los nodos corresponden a núcleos y las aristas se definen en función de la distancia entre centroides. Sin embargo, en este caso el interés se centra en identificar cómo interactúan espacialmente las diferentes clases celulares dentro de una región.

Para ello, las características se calculan a nivel de *spot*, considerando el subgrafo inducido por los núcleos asignados a cada uno. A partir de este subgrafo se extraen dos tipos principales de descriptores, detallados a continuación e ilustrados en la figura 16:

- **Fracción de aristas entre clases (*edge_frac_ClassA_ClassB*):** Para cada par de clases celulares (c_1, c_2), se calcula la proporción de aristas del grafo que conectan nodos de dichas clases:

$$\text{edge\_frac}_{c_1,c_2} = \frac{\#\text{ aristas entre } c_1 \text{ y } c_2}{\#\text{ total de aristas}} \tag{3.15}$$

  Dado que se consideran cinco clases celulares y se evalúan todas las combinaciones sin repetición (incluyendo interacciones de una clase consigo misma), en total se obtienen 15 características de este tipo. Esta métrica cuantifica la intensidad de interacción espacial entre diferentes tipos celulares. Por ejemplo, un valor alto en la interacción entre células neoplásicas e inflamatorias puede indicar regiones con infiltración inmune en el tejido tumoral.

- **Entropía de clases en vecindarios (*class_entropy*):** Para cada nodo, se considera la distribución de clases celulares en su vecindario inmediato (nodos adyacentes), y se

calcula la entropía de Shannon:

$$H = -\sum_{i} p_i \log_2 p_i \tag{3.16}$$

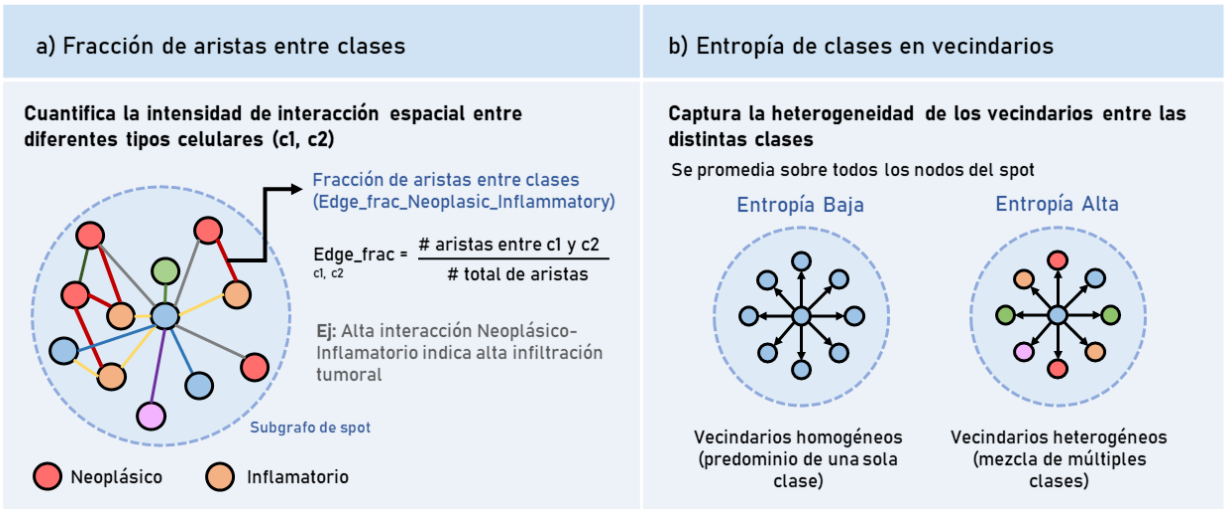
donde p_i es la proporción de vecinos pertenecientes a la clase i. Posteriormente, se promedia esta entropía sobre todos los nodos del grafo del *spot*.

Esta métrica captura el grado de heterogeneidad celular local:

- Entropía baja: vecindarios homogéneos (predominio de una sola clase)
- Entropía alta: vecindarios heterogéneos (mezcla de múltiples clases)

Figura 16

Representación de interacciones espaciales basadas en grafos de clases celulares.



Nota. Representación visual aproximada de las características basadas en grafos de interacciones entre clases celulares. (a) Fracción de aristas entre las cinco clases celulares presentes en la segmentación nuclear. (b) Entropía de clases en vecindarios mediante el promedio de los valores de cada nodo de forma individual.

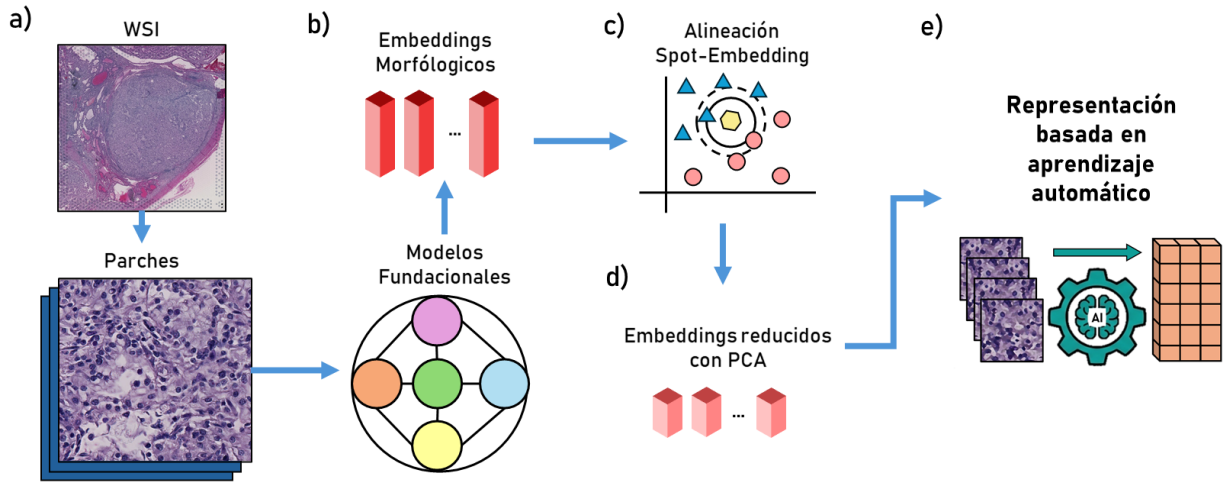
A diferencia del enfoque anterior basado en características locales por nodo, en este caso las métricas se calculan directamente a nivel de *spot*, considerando la estructura del subgrafo correspondiente. De esta forma, se logra resumir tanto la organización espacial como la interacción entre distintos tipos celulares dentro de cada región del tejido. En total, esta representación genera 16 características por *spot*: 15 correspondientes a las fracciones de aristas entre clases y una asociada a la entropía de clases en los vecindarios.

3.3.3. Representaciones de Aprendizaje Automático

En este apartado se describe el paso a paso seguido para la extracción de las representaciones morfológicas a partir de las imágenes WSI de las muestras analizadas. Este proceso, se realizó mediante el uso de modelos fundacionales, previamente entrenados en inmensos conjuntos de datos, los cuales logran transformar regiones visuales específicas de la imagen en vectores numéricos de características que capturan la información morfológica relevante del tejido. Estas representaciones, a las que se les suele llamar *embeddings*, permiten resumir aquellos patrones visuales complejos presentes en las WSIs y facilitar su posterior análisis en conjunto con los datos de transcriptómica espacial. La figura 17 ilustra el proceso llevado a cabo para el desarrollo y análisis de este tipo de representaciones.

Figura 17

Flujo de trabajo para la obtención y análisis de representaciones mediante modelos fundacionales.



Nota. Flujo de trabajo seguido para la extracción y posterior análisis de las representaciones. (a) Obtención de parches; (b) Generación de *embeddings* mediante modelos fundacionales; (c) Asociación espacial parche-*spot* para obtener la expresión génica (Y); (d) Reducción de dimensionalidad mediante PCA (X_{PCA}); (e) Análisis de correlación de Pearson entre la morfología reducida y la expresión génica.

Para este trabajo se utilizaron cuatro modelos fundacionales distintos: UNI, UNI2-h, Virchow y Virchow2, con el fin de extraer las representaciones morfológicas. Estas representaciones serán comparadas en etapas posteriores del análisis. Las arquitecturas y estrategias de entrenamiento de dichos modelos fueron descritas previamente en la sección 2.3.

La etapa inicial consistió en la obtención de los parches histológicos a partir de los archivos estructurados en formato `.h5` del conjunto de datos HEST-1K (descritos en la sección 3.3). Este enfoque evitó la extracción manual desde las imágenes WSI, al utilizar directa-

mente contenedores con información visual previamente curada. En particular, el tensor `img` proporcionó los parches en formato RGB con tamaño estándar (224×224), mientras que la matriz `coords` aportó las coordenadas espaciales (x, y) de cada parche dentro de la imagen original.

A partir de estos datos, se generaron los *embeddings* morfológicos procesando cada parche de forma independiente mediante los cuatro modelos fundacionales. Este proceso permitió transformar la información visual en representaciones vectoriales en el espacio latente. La dimensionalidad de dichos vectores depende de la arquitectura de cada modelo, como se resume en la Tabla 3.

Tabla 3: *Dimensionalidad de los embeddings generados por cada modelo.*

Modelo	**Dimensión del *embedding***
UNI	1024
UNI2-h	1536
Virchow	2560
Virchow2	2560

Para vincular las representaciones morfológicas con la actividad molecular, se realizó una asociación espacial con los *spots* de transcriptómica. Este proceso siguió el esquema de alineación basado en *Nearest Neighbors* ($k = 1$), utilizando distancia euclidiana bidimensional, tal como se describió previamente en la subsección 3.3.1. El emparejamiento se llevó a cabo cruzando las coordenadas de los parches (`coords`) con las posiciones de los *spots* almacenadas en `obsm["spatial"]` del objeto *AnnData*. Además, se definió un umbral de tolerancia para asegurar que cada *embedding* se asociara únicamente con su región anatómica correspondiente. Como resultado, se obtuvieron dos matrices: X, con los descriptores morfológicos, y Y, con los perfiles de expresión génica asociados.

Dada la alta dimensionalidad de los vectores generados (hasta 2560 dimensiones), fue necesario aplicar una técnica de reducción de dimensionalidad para facilitar su manejo. Para ello, se empleó Análisis de Componentes Principales (PCA) sobre la matriz X, proyectando los datos en un espacio latente común. La representación final se redujo a 256 componentes principales (X_{pca}), conservando entre el 89 % y el 99 % de la variabilidad original, dependiendo del modelo. Estas representaciones comprimidas se utilizaron posteriormente en las etapas de regresión y análisis de correlación.

3.4. Integración de Representaciones de Grafos y Modelos Fundacionales

En las secciones anteriores de este capítulo se han abordado tres tipos de representaciones morfológicas: las basadas en características clásicas, las representaciones de grafos celulares y los *embeddings* provenientes de modelos fundacionales. Cada una de estas aproximaciones captura aspectos distintos de la histología, desde descriptores de textura y forma, hasta la organización espacial de los núcleos y patrones visuales complejos de alto nivel.

Dado que estos enfoques ofrecen información complementaria, se planteó integrar las dos representaciones que mostraron una mayor capacidad para capturar la variabilidad de la expresión génica del tejido en las fases preliminares de este trabajo, en función del coeficiente de correlación de *Pearson*, cuyos resultados se presentan en la sección 4.1. En particular, se seleccionaron las representaciones basadas en grafos de clases celulares y los *embeddings* generados por los modelos fundacionales. Mientras que los modelos fundacionales resumen la morfología visual global, las representaciones de grafos de clases celulares incorporan información estructural detallada sobre la disposición y las interacciones de las células en el microambiente tisular.

La integración de ambas representaciones se llevó a cabo a nivel de *spot*. Es importante destacar que este mismo flujo de procesamiento se aplicó de manera uniforme tanto a las muestras de ccRCC como a las muestras de tejido renal sano. Proceder de esta manera garantiza que el espacio de características final capture el espectro completo de variabilidad biológica, permitiendo al modelo contrastar la arquitectura de un tejido normal frente a las alteraciones inducidas por el tumor.

Dado que múltiples parches pueden ser asignados a un mismo *spot*, los *embeddings* correspondientes se agregaron mediante el promedio aritmético, obteniendo así una única representación morfológica por *spot*. De manera análoga, los núcleos celulares segmentados fueron asignados al *spot* más cercano, permitiendo construir un subgrafo por región sobre el cual se calcularon las características de interacción entre clases celulares. Adicionalmente, se consideraron únicamente aquellos *spots* que contenían al menos cinco núcleos celulares, con el fin de garantizar estabilidad en el cálculo de las características basadas en grafos.

Posteriormente, y a partir de las representaciones previamente definidas en las subsecciones 3.3.2.2 y 3.3.3, se construyó una representación integrada a nivel de *spot*. En particular, para cada *spot* i, se dispone de un vector de características morfológicas $\mathbf{x}_i^{(emb)}$ derivado de los *embeddings* y un vector $\mathbf{g}_i$ correspondiente a las características de interacción entre clases celulares obtenidas a partir de grafos.

La integración de ambas fuentes de información se realizó mediante la concatenación de estos vectores, definiendo una representación conjunta:

$$\mathbf{z}_i = [\mathbf{x}_i^{(emb)}, \mathbf{g}_i] \tag{3.17}$$

De esta manera, cada *spot* queda descrito simultáneamente por información morfológica global y por características estructurales del microambiente celular.

A partir de esta representación integrada, se formuló un modelo de regresión *Ridge* independiente para cada gen. Sea $y_i^{(g)}$ el nivel de expresión del gen g en el *spot* i, el modelo se expresa como:

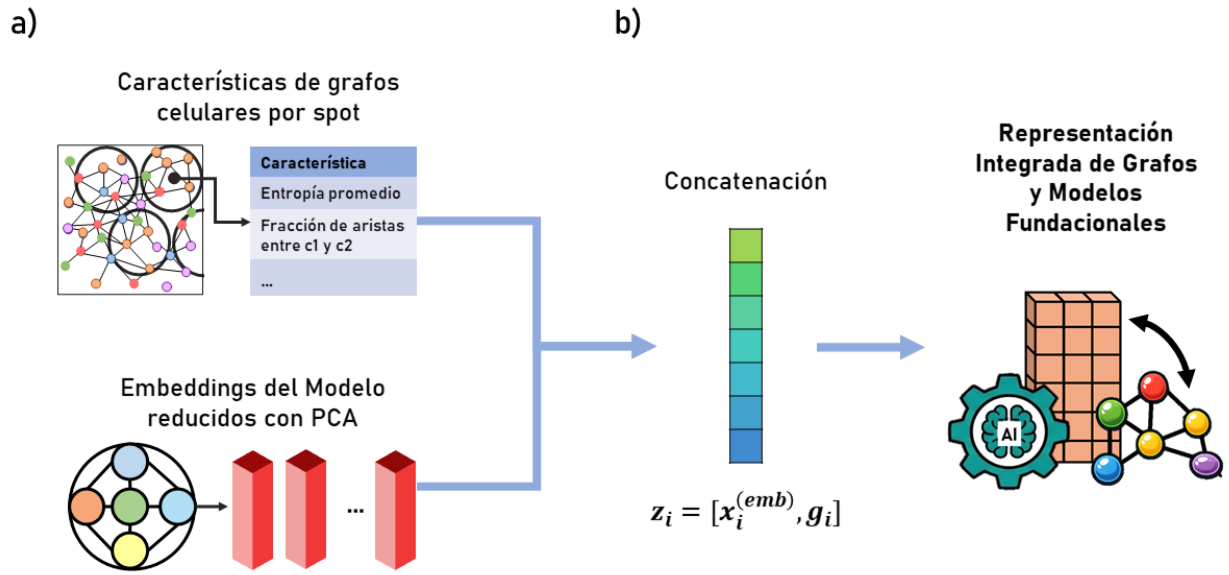
$$y_i^{(g)} = \mathbf{w}_g^\top \mathbf{z}_i + b_g \tag{3.18}$$

donde $\mathbf{w}_g$ representa el vector de coeficientes asociado a las características integradas y b_g es el término independiente.

El modelo se entrena minimizando una función de costo con regularización L2, lo cual permite controlar la complejidad del modelo y mitigar posibles problemas de sobreajuste derivados de la alta dimensionalidad de las representaciones. Cabe resaltar que el proceso detallado de entrenamiento, evaluación y análisis de este modelo de regresión se presenta en el Capítulo 4, específicamente en la sección 4.2. El proceso descrito previamente para el desarrollo de este tipo de representaciones se ilustra visualmente mediante la figura 18.

Figura 18

Flujo de trabajo para la generación de representaciones integradas.



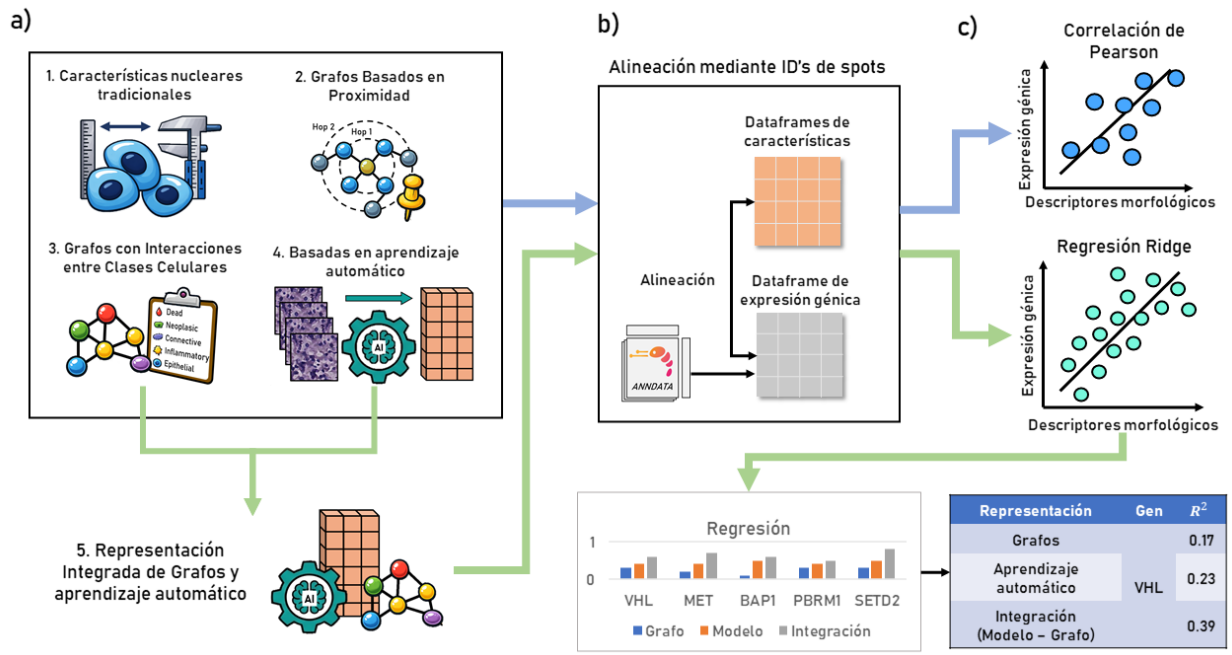
Nota. El proceso de fusión de datos comprende: (a) la selección de descriptores topológicos de grafos celulares y *embeddings* de modelos fundacionales; y (b) concatenación de características y creación de dataframe alineado por spot

4. Correlación y Análisis de Regresión para las Representaciones Histopatológicas y su Transcriptómica Espacial

En este capítulo se presentan los resultados del análisis de correlación entre las representaciones morfológicas extraídas y la expresión génica espacial a nivel de *spot*, con el objetivo de identificar asociaciones directas entre la organización tisular y los patrones transcriptómicos. Este análisis permite caracterizar qué tipo de descriptores morfológicos —tradicionales, basados en grafos o derivados de modelos de aprendizaje automático— capturan mejor la variabilidad de la expresión génica. De manera complementaria, se incorpora un análisis de regresión orientado a evaluar la capacidad predictiva de las representaciones con mejor desempeño en el proceso de correlación, es decir, en qué medida la información morfológica permite estimar cuantitativamente la expresión génica. Para ello, se consideran tanto muestras de tejido renal sano y canceroso , lo que permite analizar el comportamiento de distintas variables bajo distintas condiciones biológicas. La Figura 19 ilustra de forma general el flujo metodológico seguido, integrando los procesos de correlación y regresión entre las representaciones desarrolladas en el capítulo anterior y la expresión génica.

Figura 19

Flujo de trabajo para el análisis de correlación y modelos de regresión.



Representación	Gen	R^2
Grafos	VHL	0.17
Aprendizaje automático		0.23
Integración (Modelo – Grafo)		0.39

Nota. Flujo de trabajo seguido para el proceso de selección de correlación y regresión. (a) Representaciones morfológicas obtenidas previamente; (b) alineación de dataframes para la correspondencia de datos entre spots de características y expresión génica; y (c) Proceso de correlación y regresión.

4.1. Análisis de Correlación

En esta sección se presenta el análisis de correlación entre las representaciones morfológicas obtenidas y la expresión génica a nivel de *spot*. Dicho proceso se realizó considerando el conjunto de genes seleccionados mediante análisis estadísticos y estado del arte, mencionados en la sección 3.2. A partir de este grupo de genes, se evaluó la relación entre cada tipo de representación y la expresión génica mediante el coeficiente de correlación de Pearson, esto con el objetivo de comparar su capacidad para capturar señales biológicas relevantes.

4.1.1. Metodología para la Correlación

El análisis de correlación se aplica de forma consistente a los tres tipos de representaciones descritas en el Capítulo 3: (i) características nucleares tradicionales, (ii) representaciones basadas en grafos y (iii) representaciones derivadas de modelos de aprendizaje automático. El objetivo es evaluar en qué medida las distintas representaciones (X) logran capturar señales asociadas a la expresión génica (Y), estableciendo así un criterio cuantitativo de comparación entre ellas.

Para cada tipo de representación, se construyen dos matrices principales: una matriz X, que contiene las características morfológicas extraídas, y una matriz Y, que almacena los niveles de expresión génica. Ambas matrices se organizan a nivel de *spot*, registrando las observaciones de manera consistente para cada ubicación del tejido. La correspondencia entre X y Y se establece mediante identificadores comunes de los *spots*, lo que permite alinear ambas fuentes de información y garantizar que cada fila en las matrices represente la misma región tisular, habilitando así una comparación directa entre morfología y expresión génica.

Como se describió en las secciones correspondientes a cada tipo de representación, los descriptores morfológicos se obtienen y organizan de manera diferente según el enfoque utilizado. En el caso de las características nucleares tradicionales (3.3.1) y las representaciones basadas en grafos de proximidad (3.3.2.1), las características se agregan a nivel de *spot* mediante estadísticas resumen, específicamente la media y la desviación estándar. Por su parte, en las representaciones basadas en interacciones entre clases celulares, las características corresponden directamente a descriptores globales del grafo calculados para cada *spot*. Finalmente, en las representaciones derivadas de modelos fundacionales, los *embeddings* extraídos a partir de parches de imagen son asociados a *spots* mediante el algoritmo de vecinos más cercanos, construyendo de esta manera matrices comparables entre morfología y expresión génica.

Una vez alineadas las matrices X y Y, la relación entre cada característica morfológica y la expresión génica se evalúa mediante el coeficiente de correlación de *Pearson* (r), el cual

mide la asociación lineal entre dos variables continuas. Su formulación se define como:

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{4.1}$$

donde N corresponde al número de observaciones (parches o *spots*), x_i representa el valor de la característica morfológica para la observación i, y_i el nivel de expresión génica correspondiente, y $\bar{x}$, $\bar{y}$ sus respectivos promedios.

El análisis se realiza de manera independiente para cada gen. Este proceso consiste en tomar su vector de expresión a lo largo de todas las observaciones y realizar su correlación con cada una de las características morfológicas disponibles en X.

En el caso de representaciones de baja dimensionalidad (como las características nucleares o las métricas de grafos), la correlación se calcula directamente entre cada característica y la expresión génica. Por otro lado, en el caso de representaciones de alta dimensionalidad (como los *embeddings* reducidos a 256 componentes PCA), la correlación se evalúa entre la expresión génica y cada una de las dimensiones del espacio reducido. A partir de este conjunto de valores, se selecciona el valor absoluto máximo de correlación (*max_abs_corr*) como una medida representativa de la asociación más fuerte observada para cada gen.

Cabe resaltar que, aunque el procedimiento de correlación es común a todos los tipos de representación, existen diferencias en la naturaleza de las variables analizadas. Mientras que las características tradicionales y de grafos corresponden a descriptores explícitos e interpretables, los *embeddings* representan combinaciones latentes de patrones visuales, lo que implica que la interpretación directa de cada dimensión no es inmediata. Además, previo al cálculo de correlaciones, las características morfológicas son normalizadas a nivel de muestra mediante estandarización (*z-score*), con el fin de reducir efectos de escala y hacer comparables las distintas muestras dentro del análisis global.

En las subsecciones siguientes se presentan los resultados específicos para cada tipo de representación: características nucleares tradicionales, representaciones basadas en grafos y representaciones derivadas de modelos de aprendizaje automático.

4.1.2. Resultado de Correlación entre Expresión Génica y Características Nucleares Tradicionales

En esta subsección se presentan los resultados del análisis de correlación entre las características nucleares tradicionales y la expresión génica, utilizando el conjunto de genes previamente definido. A diferencia del capítulo anterior, donde se describió el proceso de construcción de estas representaciones, en este punto el enfoque se centra en evaluar qué tan informativas resultan dichas características para capturar patrones que se relacionen con la

actividad molecular del tejido.

En la Tabla 4 se presentan los genes que alcanzan los valores más altos de correlación en este tipo de representación. Un primer aspecto relevante es que la mayoría de las correlaciones más altas se asocian con la característica *minor_axis_length* calculada como promedio por *spot*. Este comportamiento sugiere que la dimensión menor de los núcleos, y por tanto su forma, captura variaciones morfológicas que se relacionan con la expresión génica.

Tabla 4: *Top 10 de genes con mayor coeficiente de correlación de Pearson (r) utilizando características nucleares tradicionales.*

Gen	**Característica**	**Estadístico**	r
UBC	minor_axis_length	Media	0.230
RACK1	minor_axis_length	Media	0.211
EEF2	minor_axis_length	Media	0.206
UBC	area	Media	0.200
MET	minor_axis_length	Media	0.196
PABPC1	minor_axis_length	Media	0.194
UBA52	minor_axis_length	Media	0.191
EEF1G	minor_axis_length	Media	0.189
ENO1	minor_axis_length	Media	0.188
CD63	minor_axis_length	Media	0.186

Nota. Todos los valores de correlación de mayor magnitud son positivos. Los resultados corresponden al subconjunto de genes seleccionados mediante análisis estadístico y estado del arte.

Los valores de correlación observados se mantienen en un rango moderado, sin alcanzar magnitudes extremas. Este comportamiento es coherente con lo esperado en datos de transcriptómica espacial, donde factores biológicos y técnicos influyen en la expresión génica, lo que llega a limitar la capacidad de descriptores geométricos para explicar completamente dicha variabilidad. No obstante, el hecho de observar correlaciones positivas consistentes indica que estas características capturan señales relevantes del tejido.

Al analizar los genes presentes en el top, se identifican tanto genes de expresión constitutiva, tales como *UBC*, *EEF2*, *UBA52* y *PABPC1*, así como genes asociados a procesos más específicos, por ejemplo *MET*. La presencia de genes constitutivos sugiere que las ca-

racterísticas nucleares tradicionales logran capturar propiedades globales relacionadas con la actividad celular general, tales como el tamaño nuclear promedio o el estado proliferativo del tejido. Por otro lado, la aparición de genes como el *MET*, conocido en el estado del arte por su rol en proliferación y señalización en cáncer, indica que estas representaciones también reflejan, en cierta medida, alteraciones morfológicas asociadas a procesos patológicos.

Otro aspecto a tener en cuenta es la predominancia del estadístico de la media sobre la desviación estándar en los resultados más relevantes. Esto sugiere que, para este tipo de características, el valor promedio dentro de cada *spot* resulta más informativo que la variabilidad interna, lo cual es coherente con la idea de que la morfología nuclear promedio puede actuar como un descriptor robusto del estado tisular local.

Estos resultados indican que las características nucleares tradicionales, aunque limitadas en su capacidad para capturar relaciones complejas, proporcionan una base interpretable para describir la morfología del tejido. Sin embargo, su capacidad para modelar interacciones espaciales o patrones más complejos es limitada, lo cual justifica la exploración de otros tipos de representaciones más avanzadas, como las basadas en grafos y modelos de aprendizaje automático, abordadas en las siguientes subsecciones.

4.1.3. Correlación entre Expresión Génica y Características Nucleares Basadas en Grafos

En esta subsección se presentan los resultados del análisis de correlación entre la expresión génica y las representaciones morfológicas derivadas de grafos nucleares. A diferencia de las características tradicionales, estas representaciones permiten modelar explícitamente la organización espacial y relacional entre células, ya sea a partir de su proximidad física o de sus interacciones según clases celulares. Con el fin de comparar de manera sistemática estos enfoques, se consideran dos tipos de grafos nucleares: (i) grafos de proximidad espacial y (ii) grafos de interacciones entre clases celulares.

4.1.3.1. Grafos Basados en Proximidad Espacial. En la Tabla 5 se presentan los resultados para el proceso de correlación mediante grafos de proximidad espacial, descritos previamente en la sección 3.3.2.1.

Tabla 5: *Top 10 genes con mayor correlación utilizando características de grafos de proximidad (max_clique y Hop-N neighbors).*

Gen	Característica	Tipo	r
TMSB4X	hop_3	mean	0.204
TMSB4X	hop_2	mean	0.203
TMSB4X	hop_1	mean	0.200
TMSB4X	max_clique_size	mean	0.193
HLA-DRA	hop_3	mean	0.190
HLA-DRA	hop_2	mean	0.189
B2M	hop_3	mean	0.184
B2M	hop_2	mean	0.184
HLA-DRA	hop_1	mean	0.184
B2M	hop_1	mean	0.181

Nota. Todos los valores de correlación de mayor magnitud son positivos. Los resultados corresponden al subconjunto de genes seleccionados mediante análisis estadístico y estado del arte.

Los resultados para los grafos basados en proximidad espacial muestran que las correlaciones más altas alcanzan valores cercanos a $r \approx 0.20$. En este caso, las características que presentan mayor asociación con la expresión génica corresponden principalmente a medidas de vecindad local (*hop-1*, *hop-2*, *hop-3*), así como al tamaño de la máxima clique.

Se observa la presencia recurrente de genes como *TMSB4X*, *HLA-DRA* y *B2M*, lo que sugiere que estas características capturan patrones espaciales locales del tejido. No obstante, la magnitud limitada de los coeficientes indica que este tipo de representación, basada únicamente en proximidad, presenta restricciones para modelar de manera más precisa la variabilidad de la expresión génica.

4.1.3.2. Grafos Basados en Clases Celulares. En la Tabla 6 se presentan los resultados para el proceso de correlación mediante grafos de clases celulares, descritos previamente en la sección 3.3.2.2.

Tabla 6: *Top 10 genes con mayor correlación utilizando características de interacción entre clases celulares, considerando pares únicos de clases y entropía, ordenados por valor absoluto.*

Gen	**Característica**	r	$\|r\|$
MET	edge_frac_Connective_Inflammatory	-0.281	0.281
UBC	edge_frac_Connective_Neoplastic	0.275	0.275
HLA-A	edge_frac_Connective_Epithelial	0.270	0.270
TMSB4X	edge_frac_Inflammatory_Neoplastic	0.201	0.201
MET	class_entropy	-0.172	0.172
ITM2B	edge_frac_Epithelial_Neoplastic	0.127	0.127
PABPC1	edge_frac_Epithelial_Inflammatory	-0.105	0.105
EEF1G	edge_frac_Connective_Dead	-0.015	0.015
EEF1G	edge_frac_Dead_Inflammatory	-0.015	0.015
EEF1G	edge_frac_Dead_Neoplastic	-0.011	0.011

Nota. En los resultados de magnitudes mas altas de correlación se presentan genes con valores negativos, por lo cual se presenta la columna adicional $|r|$. Los resultados corresponden al subconjunto de genes seleccionados mediante análisis estadístico y estado del arte.

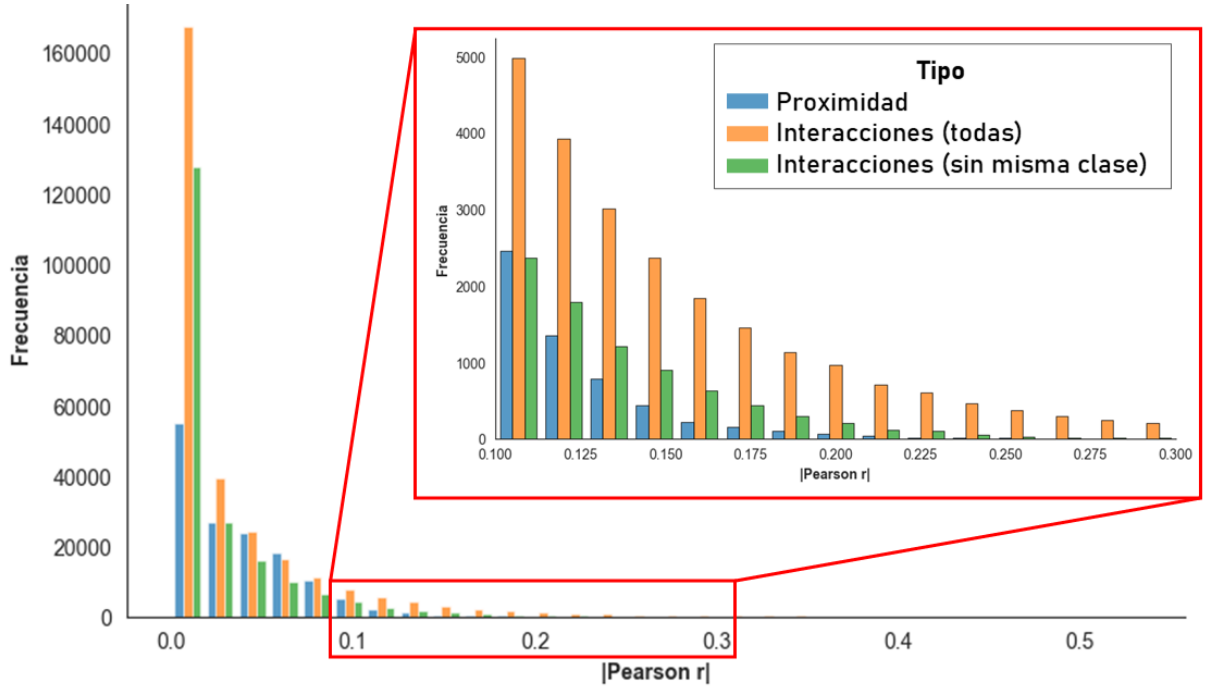
Los grafos basados en clases celulares presentan correlaciones de mayor magnitud, alcanzando valores cercanos a $|r| \approx 0.28$. Estas asociaciones se encuentran relacionadas principalmente con fracciones de interacción entre pares de clases celulares específicas, como *Connective–Inflammatory*, *Connective–Neoplastic* y *Connective–Epithelial*, así como con la entropía de clases. La mayor amplitud observada en la distribución de correlaciones sugiere que este tipo de representación captura información más rica sobre la organización celular del tejido. De manera consistente, en análisis exploratorios adicionales (no mostrados en esta tabla), se identificaron casos puntuales con correlaciones de mayor magnitud (hasta $r \approx 0.5$), asociados principalmente a interacciones entre células neoplásicas. No obstante, estos valores corresponden a situaciones específicas y no representan la tendencia global observada, la cual se mantiene concentrada alrededor de cero. Aun así, su presencia refuerza el potencial de este tipo de descriptores para capturar señales biológicamente relevantes relacionadas con la organización tumoral.

Estos resultados indican que, aunque ambos tipos de grafos aportan información relevante, las representaciones basadas en interacciones entre clases celulares presentan una mayor capacidad para capturar asociaciones significativas con la expresión génica en comparación con los grafos basados únicamente en proximidad espacial. Esta observación se

evidencia en la Figura 20, donde se comparan las distribuciones de correlación correspondientes a los grafos de proximidad y a los grafos basados en clases celulares, considerando tanto interacciones intra-clase como inter-clase.

Figura 20

Histograma comparativo de $|r|$ entre los dos tipos de representaciones basadas en grafos nucleares.



Nota. Se muestran los resultados agrupados en tres categorías: 1) Grafos de proximidad, 2) Grafos de interacciones entre clases teniendo en cuenta aristas intra-clase y 3) Grafos de interacciones entre clases descartando aristas intra-clase.

A nivel global, la Figura 20 muestra que la mayoría de los coeficientes de correlación de Pearson se concentran en valores cercanos a cero para los tres tipos de representaciones, lo que indica que, en términos generales, las asociaciones lineales son débiles. Sin embargo, al analizar la distribución de manera comparativa —representada mediante grupos de tres barras por intervalo— se observa un patrón consistente: las características basadas en interacciones entre clases celulares presentan frecuencias más altas en rangos de correlación moderada en comparación con los grafos de proximidad.

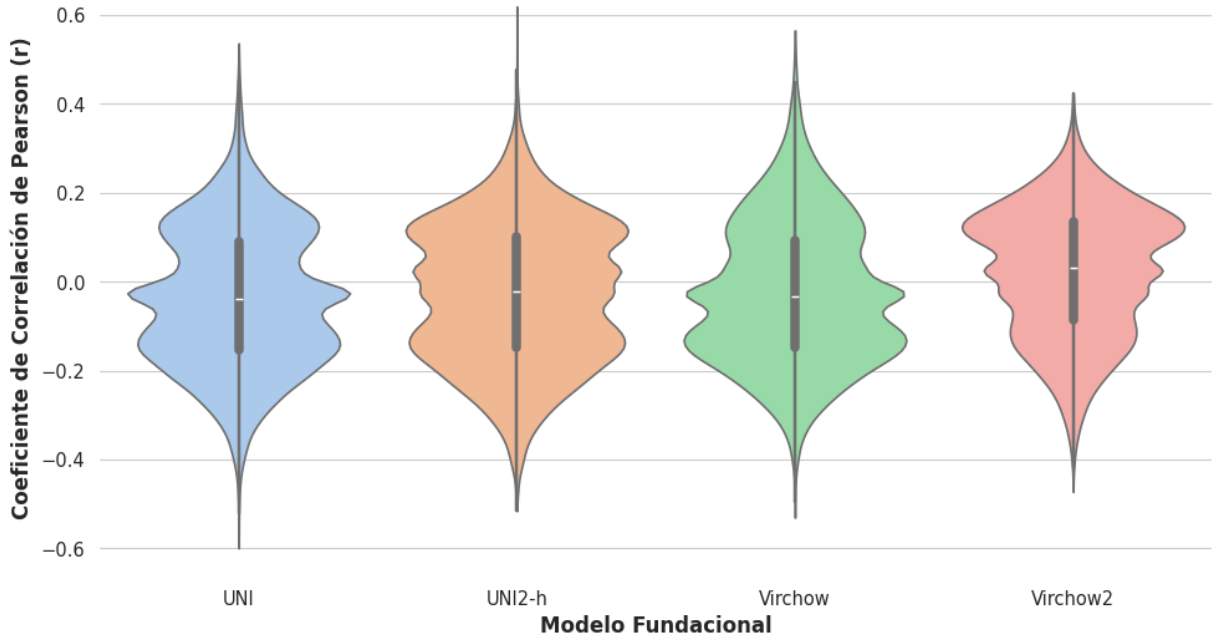
Este comportamiento se aprecia con mayor claridad en el acercamiento al intervalo 0.1 a 0.3, donde tanto las interacciones considerando todas las clases como aquellas que excluyen conexiones intra-clase superan a las representaciones basadas únicamente en proximidad. Esto sugiere que la incorporación de información semántica sobre el tipo celular introduce señales adicionales que fortalecen la relación con la expresión génica.

4.1.4. Correlación entre Expresión Génica y Características de Aprendizaje Automático

En esta subsección se analizan las correlaciones entre los *embeddings* morfológicos generados por los modelos fundacionales y la expresión génica de los genes. El interés principal no es únicamente describir estas relaciones, sino evaluar si las representaciones aprendidas capturan patrones que realmente puedan asociarse con la actividad molecular del tejido.

Figura 21

Distribución del coeficiente de correlación de Pearson por modelo fundacional.



Nota. Gráficos de violín que ilustran el coeficiente de correlación de Pearson (r) para el conjunto de genes comunes evaluados con UNI, UNI2-h, Virchow y Virchow2.

Tabla 7: *Resumen comparativo de métricas de correlación por modelo fundacional.*

Modelo	**Máx.** $\lvert r \rvert$		**Mejor** r		
	Media	**Mediana**	**Media**	**Mediana**	**D.E.**
UNI	0.135	0.127	-0.037	-0.039	0.158
UNI2-h	0.133	0.124	-0.027	-0.022	0.157
Virchow	0.136	0.126	-0.026	-0.033	0.161
Virchow2	0.124	0.118	0.023	0.032	0.145

En la Tabla 7 y la Figura 21 se presenta el comportamiento de los cuatro modelos evaluados, donde se incorpora en cada diagrama de violín un boxplot interno, lo que permite visualizar de forma más clara la mediana y la dispersión de los datos. En términos generales, la mayoría de los genes exhibe correlaciones moderadas, siendo UNI2-h el que tiene los picos más altos positivos.

Sin embargo, se observan diferencias marcadas en la dirección de las distribuciones. En los modelos UNI, UNI2-h y Virchow, la mediana de la mejor correlación se sitúa en valores negativos (-0.039, -0.022 y -0.033, respectivamente), mientras que Virchow2 muestra una tendencia opuesta, con una mediana positiva de 0.032.

Considerando esta volubilidad en los resultados, se optó por conservar los cuatro modelos en los análisis posteriores, evitando así introducir sesgos asociados a la selección de una única arquitectura. En este sentido, modelos como Virchow y UNI destacan por su capacidad para capturar los valores máximos de correlación, con medias de correlación absoluta de 0.136 y 0.135, respectivamente, lo que sugiere una mayor sensibilidad frente a variaciones biológicas extremas. Por otro lado, Virchow2 presenta la mayor estabilidad estadística del conjunto, evidenciada en una menor dispersión de los datos (0.079). A partir de esta evaluación global, se identificaron los genes con mayor correlación por cada modelo, cuyos resultados se muestran en la Tabla 8.

Tabla 8: *Top 10 de genes con mayor coeficiente correlación (r) morfológica entre modelos.*

Gen	$\|r\|$	r	**Comp.**
UBC	0.599	-0.599	PCA_1
C1QL1	0.556	-0.556	PCA_4
MT2A	0.549	-0.549	PCA_4
MT1E	0.547	-0.547	PCA_4
HLA-A	0.536	0.536	PCA_1
MRC2	0.520	-0.520	PCA_4
NDUFA4L2	0.505	0.505	PCA_1
ENPP3	0.503	-0.503	PCA_2
SLC38A5	0.503	-0.503	PCA_4
NDUFA4	0.502	0.502	PCA_1

(a) *Modelo UNI*

Gen	$\|r\|$	r	**Comp.**
UBC	0.619	0.619	PCA_1
PLOD2	0.515	-0.515	PCA_3
MATR3	0.505	0.505	PCA_1
HLA-A	0.490	-0.490	PCA_1
S100A10	0.481	-0.481	PCA_4
ANXA2	0.478	-0.478	PCA_4
PTMA	0.477	-0.477	PCA_1
HAVCR1	0.476	0.476	PCA_1
MT2A	0.475	0.475	PCA_6
IGFBP3	0.474	-0.474	PCA_4

(b) *Modelo UNI2-h*

Gen	$\|r\|$	r	**Comp.**
NDUFA4L2	0.565	0.565	PCA_1
HLA-A	0.558	0.558	PCA_1
LAPTM4A	0.546	0.546	PCA_1
PTMA	0.538	0.538	PCA_1
NDUFA4	0.536	0.536	PCA_1
UBC	0.530	-0.530	PCA_1
EIF3F	0.524	0.524	PCA_1
YBX1	0.521	0.521	PCA_1
BTF3	0.520	0.520	PCA_1
ABHD17A	0.516	0.516	PCA_1

(c) *Modelo Virchow*

Gen	$\|r\|$	r	**Comp.**
UBC	0.472	-0.472	PCA_1
C1QL1	0.453	-0.453	PCA_4
HLA-A	0.445	-0.445	PCA_2
SAA1	0.432	-0.432	PCA_6
AK4	0.424	0.424	PCA_3
HLA-DQB1	0.422	-0.422	PCA_2
PTMA	0.421	-0.421	PCA_2
AGRN	0.416	-0.416	PCA_4
EIF3F	0.416	-0.416	PCA_2
MRC2	0.414	-0.414	PCA_4

(d) *Modelo Virchow2*

Nota. Los resultados corresponden al conjunto principal de todos los genes en común en las muestras.

4.2. Análisis de Regresión

Tras la extracción de características de aprendizaje automático mediante los modelos fundacionales mencionados y la extracción de características topológicas mediante grafos

basados en clases celulares, se procede a evaluar si la integración de ambos enfoques permite modelar con mayor precisión el microambiente tisular. Este análisis utiliza como base el modelo de representación integrada propuesto en la sección 3.4. De manera consistente con el análisis de correlación mostrado previamente, la comparación entre las características morfológicas y la expresión génica se realiza a nivel de spot. Para ello, se emplean modelos de regresión Ridge, los cuales permiten cuantificar la proporción de la varianza explicada (R^2) aportada por cada conjunto de características (Hoerl y Kennard 1970).

4.2.1. Desempeño de Regresión con Características Integradas

Para analizar el impacto de la integración entre características visuales y estructurales, se comparó el desempeño predictivo del modelo de regresión bajo tres configuraciones: los *embeddings* de los modelos fundacionales, las características derivadas de los grafos celulares y el vector resultante de concatenar ambas representaciones. La métrica utilizada para realizar dicha comparación fue el coeficiente de determinación (R^2), donde un valor cercano a uno (1) indica una mayor capacidad explicativa o predictiva de la expresión génica a partir de las características implementadas. Los resultados obtenidos mediante la implementación de las características previamente mencionadas para el subconjunto de genes seleccionados (mencionado en la sección 3.2) se muestran en la Tabla 9.

Tabla 9: *Top 10 de genes con mayor desempeño predictivo (R^2) utilizando regresión Ridge sobre características morfológicas y su integración para los distintos modelos evaluados.*

Gen	R^2 G	R^2 M	R^2 Int
MT2A	0.171	0.704	0.714
UBC	0.304	0.688	0.700
HLA-A	0.194	0.666	0.676
FTH1	0.091	0.620	0.631
ACTB	0.150	0.625	0.630
HSPB1	0.131	0.619	0.628
VIM	0.113	0.616	0.624
PABPC1	0.296	0.589	0.598
RACK1	0.102	0.578	0.590
CD74	0.106	0.571	0.586

(a) *UNI*

Gen	R^2 G	R^2 M	R^2 Int
MT2A	0.171	0.734	0.741
UBC	0.304	0.698	0.709
HLA-A	0.194	0.676	0.686
VIM	0.113	0.644	0.652
FTH1	0.091	0.640	0.650
ACTB	0.150	0.639	0.643
HSPB1	0.131	0.631	0.639
CD74	0.106	0.598	0.613
PABPC1	0.296	0.599	0.607
ENO1	0.100	0.593	0.605

(b) *UNI2-h*

Gen	R^2 G	R^2 M	R^2 Int
MT2A	0.171	0.714	0.724
UBC	0.304	0.683	0.697
HLA-A	0.194	0.676	0.689
FTH1	0.091	0.639	0.651
ACTB	0.150	0.634	0.638
HSPB1	0.131	0.627	0.635
VIM	0.113	0.623	0.634
CD74	0.106	0.598	0.615
ENO1	0.100	0.589	0.602
RACK1	0.102	0.587	0.601

(c) *Virchow*

Gen	R^2 G	R^2 M	R^2 Int
MT2A	0.171	0.722	0.731
UBC	0.304	0.694	0.705
HLA-A	0.194	0.684	0.693
FTH1	0.091	0.641	0.652
ACTB	0.150	0.644	0.647
HSPB1	0.131	0.637	0.644
VIM	0.113	0.635	0.642
CD74	0.106	0.599	0.613
RACK1	0.102	0.597	0.608
ENO1	0.100	0.595	0.607

(d) *Virchow2*

Nota. R^2 G: características basadas en grafos celulares; R^2 M: *embeddings* morfológicos; R^2 Int: representación integrada. Los valores corresponden al coeficiente de determinación obtenido mediante regresión Ridge. Los resultados corresponden al subconjunto de genes seleccionados mediante análisis estadístico y estado del arte.

Al evaluar de manera individual los *embeddings* visuales provenientes de los modelos

considerados (UNI, UNI2-h, Virchow y Virchow2), se obtienen valores de R^2 consistentemente altos en comparación con las características basadas en grafos, en particular para genes como MT2A, UBC y HLA-A, que alcanzan valores cercanos o superiores a 0.70 en varios modelos, lo que evidencia una fuerte capacidad predictiva de las representaciones morfológicas profundas. Por otro lado, las características de grafos presentan valores considerablemente menores (principalmente por debajo de 0.30), lo cual sugiere que, aunque capturan información estructural relevante, su poder explicativo individual es limitado.

El resultado más relevante se observa al utilizar la representación integrada. De forma consistente en los modelos evaluados, la combinación de *embeddings* visuales con características de grafos produce mejoras en el coeficiente de determinación (R^2) respecto a cada representación por separado. Este comportamiento se mantiene a lo largo de todos los genes analizados, aunque con incrementos moderados (típicamente entre 0.005 y 0.02). En promedio, estas mejoras corresponden a 0.010 en UNI, 0.009 en UNI2-h, 0.012 en Virchow y 0.009 en Virchow2, lo que indica que el aporte de la información estructural es complementario más que dominante.

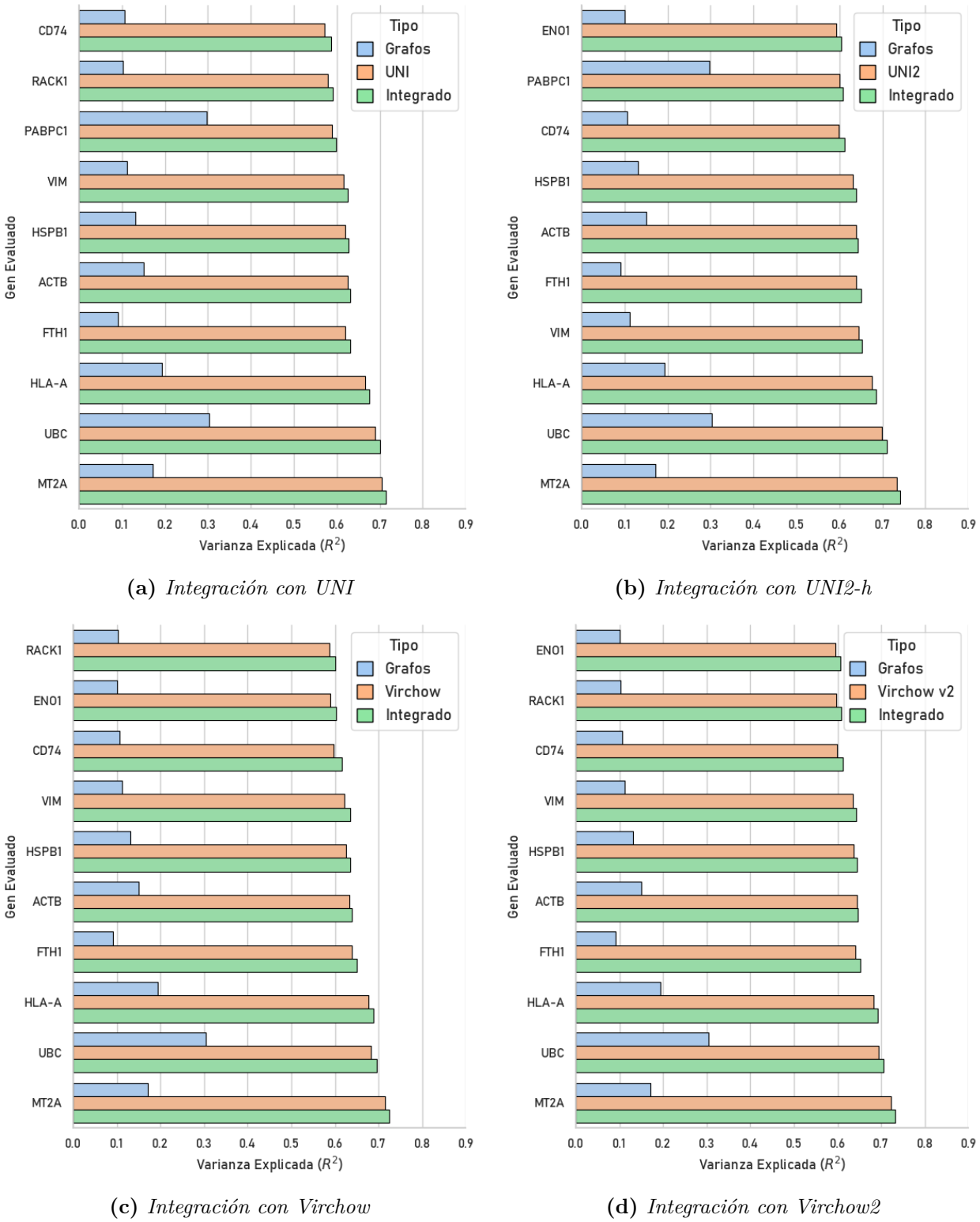
Mediante una comparación entre los distintos modelos, se observa que UNI2-h y Virchow2 tienden a presentar los valores más altos de R^2 en la representación integrada, con MT2A alcanzando hasta 0.741 en UNI2-h y 0.731 en Virchow2. Por su parte, UNI y Virchow mantienen un comportamiento estable, con valores altos y consistentes, pero sin superar a sus variantes más recientes.

Un aspecto destacable es la estabilidad de ciertos genes a través de todos los modelos, donde genes como MT2A, UBC y HLA-A aparecen de forma recurrente entre los mejores resultados, lo que sugiere que su expresión génica está fuertemente asociada con patrones morfológicos y estructurales del tejido. Asimismo, genes como FTH1, VIM y ACTB muestran un comportamiento consistente, lo que refuerza la robustez de los hallazgos.

Para ilustrar de una manera más clara el desempeño mediante los distintos tipos de representaciones, los resultados son mostrados mediante la Figura 22, donde se evidencia que la representación integrada (barras correspondientes) supera de manera sistemática tanto a los *embeddings* visuales como a las características de grafos en todos los modelos considerados. Esta tendencia confirma que, aunque los modelos fundacionales capturan la mayor parte de la señal predictiva, la incorporación de información estructural por medio de grafos celulares permite refinar dicha predicción, lo que sugiere que ambas fuentes de información son complementarias y que su integración permite construir descriptores más completos del microambiente tisular, mejorando de forma consistente la predicción de la expresión génica.

Figura 22

Comparación de la varianza explicada (R^2) entre representaciones basadas en grafos, embeddings morfológicos e integración de ambas para los distintos modelos evaluados.



(a) *Integración con UNI*

(b) *Integración con UNI2-h*

(c) *Integración con Virchow*

(d) *Integración con Virchow2*

Nota. Se observa cómo la integración de características morfológicas y grafos (verde) supera consistentemente a las representaciones aisladas (naranja y azul).

4.2.2. Comparación entre Tejido Sano y Tejido Canceroso

Con el objetivo de analizar cómo varía la relación entre la morfología del tejido y la expresión génica en diferentes condiciones, se realizó una comparación entre muestras de tejido renal canceroso y tejido renal sano, este último descrito previamente en la subsección 3.1.2. Para ello, se evalúa el desempeño predictivo mediante un modelo de regresión basado en características integradas de modelos fundacionales y grafos celulares, descrito en la sección 4.2.1.

Bajo este enfoque, se entrenan modelos de manera separada para cada condición, lo que permite contrastar directamente el desempeño predictivo entre muestras sanas y cancerosas. Este análisis busca determinar si la relación entre morfología y expresión génica resulta más predecible en presencia de alteraciones tumorales o en condiciones de tejido sano. En particular, valores más altos de R^2 en tejido canceroso podrían indicar una asociación más estrecha entre ambas dimensiones en contextos tumorales, mientras que valores más bajos en tejido sano podrían reflejar una menor dependencia o una mayor variabilidad intrínseca.

A continuación se presentan los resultados obtenidos tras evaluar la capacidad predictiva de las dos versiones del modelo integrado. A diferencia de la sección anterior, donde se presentan los resultados para el subconjunto de genes seleccionados, en esta subsección se consideran los genes con mayor capacidad predictiva a nivel global, con el fin de identificar patrones emergentes no restringidos por conocimiento previo. Estos resultados se aprecian en las siguientes tablas:

Tabla 10: *Genes globales con mayor coeficiente de determinación (R^2) en tejido sano utilizando representaciones integradas.*

Gen	R^2
MT-ND6	0.874
MT-ND4L	0.854
MT-ND5	0.817
LENG8	0.731
CCNL2	0.729
FOSB	0.728
ZFP36	0.724
PATJ	0.716
MAT2A	0.712
RBM33	0.708

(a) *Integrado con UNI*

Gen	R^2
MT-ND6	0.871
MT-ND4L	0.852
MT-ND5	0.809
CCNL2	0.723
LENG8	0.722
PATJ	0.715
FOSB	0.708
RBM33	0.702
MAT2A	0.701
ZFP36	0.700

(b) *Integrado con UNI2-h*

Gen	R^2
MT-ND6	0.870
MT-ND4L	0.853
MT-ND5	0.811
CCNL2	0.725
LENG8	0.725
FOSB	0.718
PATJ	0.713
MAT2A	0.709
ZFP36	0.708
RBM33	0.706

(c) *Integrado con Virchow*

Gen	R^2
MT-ND6	0.880
MT-ND4L	0.866
MT-ND5	0.826
FOSB	0.790
ZFP36	0.788
JUNB	0.760
SOCS3	0.758
JUN	0.757
MCL1	0.757
LENG8	0.752

(d) *Integrado con Virchow2*

Nota. Los resultados corresponden al conjunto principal de todos los genes en común en las muestras.

Tabla 11: *Genes globales con mayor coeficiente de determinación (R^2) en tejido canceroso utilizando representaciones integradas.*

Gen	R^2
PPP1R1A	0.721
S100A10	0.720
ITGA3	0.720
S100A1	0.718
MT2A	0.714
ANXA2	0.708
VDAC1	0.701
ITGB4	0.701
UBC	0.700
IGFBP6	0.698

(a) *Integrado con UNI*

Gen	R^2
S100A1	0.742
MT2A	0.741
S100A10	0.741
ITGA3	0.734
PPP1R1A	0.733
MT1E	0.724
ANXA2	0.718
VDAC1	0.711
UBC	0.709
IGFBP6	0.709

(b) *Integrado con UNI2-h*

Gen	R^2
S100A10	0.728
MT2A	0.724
S100A1	0.719
PPP1R1A	0.719
ITGA3	0.716
MT1E	0.714
PTMA	0.710
ANXA2	0.705
VDAC1	0.704
UBC	0.697

(c) *Integrado con Virchow*

Gen	R^2
S100A10	0.735
S100A1	0.734
ITGA3	0.733
MT2A	0.731
PPP1R1A	0.731
ANXA2	0.722
MT1E	0.718
ITGB4	0.710
VDAC1	0.709
IGFBP6	0.709

(d) *Integrado con Virchow2*

Nota. Los resultados corresponden al conjunto principal de todos los genes en común en las muestras.

Al comparar los resultados de las Tablas 10 y 11, se observan patrones consistentes en los modelos evaluados. En primer lugar, destaca que el desempeño predictivo máximo es, en general, más alto en tejido sano, donde múltiples genes alcanzan valores de R^2 superiores a

0.85. Este comportamiento es consistente entre modelos, con genes mitocondriales encargados de la producción de energía celular mediante fosforilación oxidativa, tales como *MT-ND6*, *MT-ND4L* y *MT-ND5* (T. Ferreira y Rodriguez 2024), ocupando las primeras posiciones en todas las configuraciones. En conjunto, esto sugiere una fuerte relación entre la morfología tisular y procesos metabólicos fundamentales en condiciones no patológicas.

Adicionalmente, en tejido sano se observa una alta estabilidad entre modelos, ya que los conjuntos de genes mejor predichos son prácticamente idénticos entre UNI, UNI2-h y Virchow, con ligeras variaciones en el orden e incrementos leves en Virchow2. Esto indica que, independientemente del modelo fundacional utilizado, la señal morfológica en tejido sano es consistente, permitiendo una predicción confiable de la expresión génica. En contraste, en tejido canceroso los valores máximos de R^2 tienden a ser ligeramente menores y más variables entre modelos. Mientras que UNI, UNI2-h y Virchow presentan valores máximos cercanos a 0.72–0.74, Virchow2 mantiene un comportamiento similar pero con mejoras en algunos genes específicos.

Otro aspecto relevante es la divergencia en los genes mejor predichos en cáncer. A diferencia del tejido sano, donde predominan genes mitocondriales, en tejido canceroso aparecen genes asociados a procesos como adhesión celular (*ITGA3*, *ITGB4*) (D'Arcy y Kiel 2021), respuesta al estrés (*S100A1*, *S100A10*) (Okura et al. 2023), y regulación metabólica (*MT2A*, *MT1E*) (Miyazaki y Asanuma 2023). Esta variabilidad refleja la heterogeneidad del microambiente tumoral y sugiere que la relación entre morfología y expresión génica depende más del contexto biológico. Asimismo, se identifican genes recurrentes entre modelos en cáncer, como *MT2A* y *UBC*, los cuales pertenecen al subconjunto de genes previamente seleccionados con base en criterios estadísticos y evidencia reportada en el estado del arte. Esto sugiere que ciertos procesos biológicos mantienen una asociación consistente con la morfología incluso en condiciones tumorales. Sin embargo, la menor coincidencia global entre listas en comparación con el tejido sano evidencia una mayor dispersión en la señal predictiva.

4.2.3. Método Comparativo de Modelos Lineales

Con el objetivo de cuantificar el aporte de la información asociada al estado patológico en la predicción de la expresión génica, se plantean dos modelos de regresión Ridge que difieren únicamente en la inclusión de la variable de condición biológica. En particular, se define un modelo base M_A, que utiliza la representación de características morfológicas integradas ya definida y utilizada en secciones anteriores, y un modelo completo M_B, que incorpora adicionalmente una variable indicadora que distingue entre tejido sano y canceroso. La comparación entre ambos modelos permite observar el efecto de la etiqueta de cáncer sobre el desempeño predictivo. Es decir, cualquier mejora observada al pasar de M_A a M_B puede

atribuirse directamente a la información adicional contenida en la variable de condición. De esta manera, el experimento no solo evalúa la capacidad de las representaciones morfológicas para explicar la expresión génica, sino también en qué medida dicha explicación depende de conocer el estado patológico del tejido.

En particular, se introduce una variable binaria z definida como:

$$z = \begin{cases} 1 & \text{si el } \textit{spot} \text{ pertenece a tejido canceroso} \\ 0 & \text{si el } \textit{spot} \text{ pertenece a tejido sano} \end{cases}$$

De esta manera, los modelos pueden expresarse como:

$$M_A : \quad y_g \sim X \tag{4.2}$$

$$M_B : \quad y_g \sim X + z \tag{4.3}$$

donde y_g representa la expresión del gen g, y X corresponde al conjunto de características morfológicas integradas, que incluyen tanto representaciones basadas en *embeddings* como descriptores derivados de grafos celulares. Dado que se emplea un modelo lineal Ridge, la formulación explícita del modelo completo puede escribirse como:

$$y_g = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_i + \alpha_z z \tag{4.4}$$

donde X_i representa cada una de las características morfológicas (componentes de PCA y características de grafos), y α_z es el coeficiente asociado a la condición biológica. Este coeficiente captura el cambio promedio en la expresión génica atribuible al estado tumoral, una vez controlado el efecto de la morfología. De forma equivalente, el modelo puede interpretarse como:

$$y_g = \begin{cases} \alpha_0 + \sum_{i=1}^{p} \alpha_i X_i & \text{si } z = 0 \text{ (tejido sano)} \\ \alpha_0 + \sum_{i=1}^{p} \alpha_i X_i + \alpha_z & \text{si } z = 1 \text{ (tejido canceroso)} \end{cases} \tag{4.5}$$

lo que permite entender z como un desplazamiento global en la predicción de la expresión génica entre condiciones, condicionado a una misma estructura morfológica.

El desempeño de ambos modelos se evalúa mediante validación cruzada (*cross-validation*), una técnica que consiste en dividir el conjunto de datos en múltiples subconjuntos o *folds*. En cada iteración, el modelo se entrena sobre un subconjunto de los datos y se evalúa sobre el subconjunto restante, repitiendo este proceso de forma rotativa. En este trabajo se emplea

un esquema de 3 *folds*, lo que permite obtener estimaciones más robustas y generalizables del desempeño, reduciendo el riesgo de sobreajuste.

A partir de este procedimiento se obtienen las métricas:

$$R^2_{\mathrm{CV}}(M_A), \quad R^2_{\mathrm{CV}}(M_B)$$

donde el subíndice CV indica que los valores corresponden al promedio del desempeño en los distintos *folds*. A partir de estos valores, se define la mejora atribuible a la inclusión de la variable de condición como:

$$\Delta R^2_{\mathrm{CV},g} = R^2_{\mathrm{CV}}(M_B) - R^2_{\mathrm{CV}}(M_A) \tag{4.6}$$

Esta diferencia cuantifica el aporte adicional de la información de cáncer en la predicción de la expresión génica. Para caracterizar la dependencia de cada gen respecto a esta variable, se introduce la métrica:

$$T_g = -\log\left(|\Delta R^2_{\mathrm{CV},g}| + \varepsilon\right) \tag{4.7}$$

donde ε es un término pequeño que garantiza estabilidad numérica. Valores altos de T_g indican que la inclusión de la etiqueta de cáncer no mejora significativamente el desempeño, lo que sugiere que la expresión del gen puede explicarse principalmente a partir de la morfología.

Estrategia de estratificación. Un aspecto clave del enfoque es el esquema de validación cruzada, el cual se implementa mediante la clase `StratifiedGroupKFold` de la biblioteca `scikit-learn`. Esta estrategia combina dos criterios fundamentales:

- **Agrupamiento por paciente:** todas las observaciones (spots) provenientes de una misma muestra o paciente se mantienen en el mismo fold. Esto evita fuga de información (*data leakage*), ya que características específicas de un paciente no se comparten entre entrenamiento y validación.
- **Estratificación por condición:** se preserva la proporción de muestras sanas y cancerosas en cada fold, garantizando que el modelo sea evaluado bajo distribuciones balanceadas de clases.

En la práctica, la partición se realiza a nivel de pacientes y posteriormente se proyecta a nivel de spots. De este modo, cada fold contiene un subconjunto disjunto de pacientes, manteniendo simultáneamente el balance entre condiciones biológicas. Esto permite obtener estimaciones más robustas y generalizables del desempeño predictivo, evitando sobreajuste a patrones específicos de una sola muestra.

Finalmente, se plantea una optimización multiobjetivo mediante el frente de Pareto, considerando dos criterios: maximizar $R^2_{\mathrm{CV}}(M_A)$ y minimizar T_g. Esto permite identificar genes cuya expresión es bien explicada por la estructura tisular y que presentan baja dependencia de la etiqueta de condición patológica, es decir, genes cuya señal está fuertemente asociada a la morfología intrínseca del tejido. El proceso completo de construcción del conjunto de datos y validación cruzada se ilustra en el Algoritmo 1, mientras que la función para la integración de características a nivel de *spot* se ilustra en el Algoritmo 2. Por último, el ajuste de modelos por gen se detalla en el Algoritmo 3.

Algorithm 1 Algoritmo de construcción del conjunto de datos y esquema de validación

Entrada: Muestras $\mathcal{S}$, embeddings, segmentación celular, expresión génica

Salida: Matrices X_{base}, X_{full}, Y y particiones $\mathcal{F}$

1. Selección de genes:

Obtener el conjunto de genes comunes $\mathcal{G}$ en todas las muestras.

2. Construcción de observaciones a nivel de *spot*:

Ejecutar Algoritmo 2.

3. Preprocesamiento:

Aplicar PCA y estandarización sobre X_{embed}.

Normalizar X_{graph}.

Definir:

$$X_{\mathrm{base}} = [X_{\mathrm{embed}}, X_{\mathrm{graph}}], \quad X_{\mathrm{full}} = [X_{\mathrm{base}}, z]$$

4. Validación cruzada estratificada:

Construir particiones $\mathcal{F}$ usando **StratifiedGroupKFold**, asegurando:

- Separación por paciente
- Balance entre clases (cáncer/sano)

5. Ajuste de modelos por gen:

Ejecutar Algoritmo 3.

6. Selección de genes:

Filtrar genes con $R^2_{\mathrm{CV}}(M_A) > 0$.

Construir frente de Pareto maximizando $R^2_{\mathrm{CV}}(M_A)$ y T_g.

7. Resultado:

Retornar genes no dominados y métricas asociadas.

Algorithm 2 Algoritmo de construcción de observaciones a nivel de *spot*

Entrada: Muestras $\mathcal{S}$, expresión génica, embeddings, segmentación celular

Salida: X_{embed}, X_{graph}, y, z, grupo

Para cada *muestra* $s \in \mathcal{S}$ **hacer**

  Cargar embeddings, expresión génica y segmentación celular.

  Construir el grafo global G.

  Asociar *patches* a *spots* y nodos del grafo a *spots*.

  Para cada *spot válido* **hacer**

    Si el número de nodos es menor a 5, omitir el spot.

    Calcular X_{embed} como el promedio de embeddings.

    Extraer X_{graph} a partir de interacciones entre clases celulares.

    Obtener y (expresión génica) y z (cáncer/sano).

    Almacenar $(X_{\text{embed}}, X_{\text{graph}}, y, z, \text{grupo})$.

  fin

fin

Algorithm 3 Algoritmo de ajuste de modelos por gen y validación cruzada

Entrada: X_{base}, X_{full}, Y, particiones $\mathcal{F}$, genes $\mathcal{G}$

Salida: $R^2_{\text{CV}}(M_A)$, $R^2_{\text{CV}}(M_B)$, ΔR^2_g, T_g

Para cada *gen* $g \in \mathcal{G}$ **hacer**

  Obtener y_g.

  Si $\text{Var}(y_g) = 0$, omitir el gen.

  Inicializar listas de desempeño R^2_A y R^2_B.

  Para cada *fold* $(\mathcal{D}_{train}, \mathcal{D}_{test}) \in \mathcal{F}$ **hacer**

    Ajustar modelo M_A: $y_g \sim X_{\text{base}}$.

    Evaluar R^2_{base} en $\mathcal{D}_{\text{test}}$ y almacenar.

    Ajustar modelo M_B: $y_g \sim X_{\text{full}}$.

    Evaluar R^2_{full} en $\mathcal{D}_{\text{test}}$ y almacenar.

  fin

  Calcular:

$$R^2_{\text{CV}}(M_A), \quad R^2_{\text{CV}}(M_B)$$

$$\Delta R^2_g = R^2_{\text{CV}}(M_B) - R^2_{\text{CV}}(M_A)$$

$$T_g = -\log\left(|\Delta R^2_g| + \varepsilon\right)$$

fin

Los resultados obtenidos evidencian el impacto del esquema de validación cruzada estratificada en la estimación del desempeño predictivo y, particularmente, en la cuantificación del efecto de la variable de condición z. A diferencia de análisis previos sin validación

cruzada, este enfoque introduce una evaluación más estricta, reduciendo la sobreestimación del desempeño y atenuando el efecto aparente de la variable de cáncer sobre la predicción. Todos los resultados correspondientes al presente proceso de regresión son mostrados en las tablas a continuación.

Tabla 12: *Genes con mayor coeficiente de determinación ($R^2_{CV}(M_A)$) utilizando el modelo de regresión Ridge con representaciones integradas.*

Gen	$R^2_{\mathbf{CV}}(M_A)$	$R^2_{\mathbf{CV}}(M_B)$	T_g
UBC	0.461	0.471	4.564
HIST1H1E	0.296	0.282	4.279
DEK	0.286	0.280	5.168
TMED2	0.285	0.267	4.019
HIST1H4C	0.273	0.271	6.476
VWF	0.262	0.216	3.085
SMARCE1	0.261	0.250	4.449
HAVCR1	0.253	0.253	8.062
HINT1	0.246	0.205	3.197
PRRC2C	0.240	0.201	3.262

(a) *UNI*

Gen	$R^2_{\mathbf{CV}}(M_A)$	$R^2_{\mathbf{CV}}(M_B)$	T_g
HIST1H1E	0.299	0.304	5.421
UBC	0.291	0.308	4.043
HAVCR1	0.274	0.270	5.789
RBM3	0.269	0.279	4.552
VWF	0.266	0.273	4.914
FXYD2	0.249	0.246	5.807
GPX3	0.243	0.243	9.602
ENPP3	0.240	0.239	6.752
SMARCE1	0.223	0.235	4.458
PRRC2C	0.219	0.225	5.140

(b) *UNI2-h*

Gen	$R^2_{\mathbf{CV}}(M_A)$	$R^2_{\mathbf{CV}}(M_B)$	T_g
HAVCR1	0.310	0.308	5.998
UBC	0.280	0.247	3.431
VWF	0.255	0.221	3.387
BTF3	0.218	0.204	4.269
SMARCE1	0.216	0.198	3.996
FXYD2	0.209	0.205	5.657
HIST1H1E	0.208	0.205	5.744
NDUFA4L2	0.187	0.166	3.882
NUTF2	0.168	0.158	4.609
LAPTM4A	0.165	0.151	4.311

(c) *Virchow*

Gen	$R^2_{\mathbf{CV}}(M_A)$	$R^2_{\mathbf{CV}}(M_B)$	T_g
UBC	0.415	0.415	10.538
HIST1H1E	0.299	0.288	4.524
SMARCE1	0.274	0.267	4.917
ENPP3	0.273	0.277	5.393
TMED2	0.264	0.262	6.183
DEK	0.253	0.247	5.033
PRRC2C	0.251	0.211	3.199
SND1	0.246	0.228	3.983
PLOD2	0.239	0.215	3.728
NUTF2	0.239	0.235	5.457

(d) *Virchow2*

Nota. Los resultados corresponden al conjunto principal de todos los genes en común.

Los resultados presentados en la Tabla 12 muestran los genes con mayor capacidad predictiva basada exclusivamente en características morfológicas. En todos los modelos fundacionales se observa que los valores de $R^2_{\mathrm{CV}}(M_A)$ se sitúan típicamente en un rango aproximado de 0.24 a 0.46, siendo *UBC* el gen más consistente —y además parte del subconjunto de genes previamente seleccionados con base en criterios estadísticos y evidencia del estado del arte—, alcanzando valores cercanos a 0.461 en UNI y 0.415 en Virchow2.

Un aspecto clave es que, tras la implementación de validación cruzada estratificada por pacientes, la diferencia entre $R^2_{\mathrm{CV}}(M_A)$ y $R^2_{\mathrm{CV}}(M_B)$ se reduce considerablemente. En la mayoría de los casos, la mejora $\Delta R^2_{\mathrm{CV},g}$ es inferior a 0.02, e incluso negativa para varios genes (por ejemplo, *VWF*, *HINT1* o *PRRC2C* en UNI). Esto indica que la inclusión de la variable de condición biológica no aporta una mejora significativa en el desempeño predictivo bajo un esquema de evaluación más robusto.

Este comportamiento se refleja directamente en los valores de T_g, que para estos genes oscilan principalmente entre 3 y 6, alcanzando valores más altos cuando la diferencia entre modelos es prácticamente nula. Por ejemplo, *HAVCR1* en UNI presenta $T_g \approx 8.06$, indicando alta independencia respecto a la variable z.

La Tabla 13 mostrada a continuación, presenta los genes con mayor valor de T_g, es decir, aquellos cuya predicción es prácticamente independiente de la etiqueta de cáncer.

Tabla 13: *Genes con mayor estadístico T_g utilizando el modelo de regresión Ridge con representaciones integradas.*

Gen	$R^2_{\mathbf{CV}}(M_A)$	$R^2_{\mathbf{CV}}(M_B)$	T_g
RAB19	0.012	0.012	10.178
C19orf71	0.017	0.017	9.724
INHBA	0.078	0.078	9.649
SLC39A4	0.026	0.026	9.585
ABI3BP	0.152	0.152	8.993
PROSER3	0.020	0.020	8.926
CLDN11	0.065	0.064	8.843
C16orf87	0.006	0.006	8.703
RGS22	0.009	0.009	8.511
KCNK17	0.005	0.006	8.444

(a) *UNI*

Gen	$R^2_{\mathbf{CV}}(M_A)$	$R^2_{\mathbf{CV}}(M_B)$	T_g
TTC30A	0.005	0.005	13.175
BCL6B	0.007	0.007	12.329
LGALS9C	0.011	0.011	11.199
CENPN	0.063	0.063	10.879
FLII	0.020	0.020	10.514
MCM10	0.003	0.003	10.318
HIST1H2A	0.021	0.021	10.290
KIAA1211	0.015	0.015	10.288
TRIP13	0.035	0.035	10.195
CAMK1G	0.007	0.007	10.116

(b) *UNI2-h*

Gen	$R^2_{\mathbf{CV}}(M_A)$	$R^2_{\mathbf{CV}}(M_B)$	T_g
SLC26A10	0.006	0.006	10.985
ADAMTS12	0.047	0.047	10.360
HSD17B12	0.034	0.034	9.966
ACER3	0.005	0.005	9.452
PDGFRA	0.004	0.004	8.847
ITGB3	0.024	0.024	8.668
SLC12A8	0.005	0.004	8.577
TRIP13	0.013	0.012	8.319
CORIN	0.010	0.010	8.184
RFTN2	0.000	0.001	8.124

(c) *Virchow*

Gen	$R^2_{\mathbf{CV}}(M_A)$	$R^2_{\mathbf{CV}}(M_B)$	T_g
SLC25A46	0.055	0.055	11.399
UBC	0.415	0.415	10.538
DHCR7	0.016	0.016	10.423
SLC5A4	0.013	0.013	10.231
ABI3BP	0.085	0.085	10.198
GNAZ	0.019	0.019	9.910
PFDN1	0.024	0.024	9.892
ADAM20	0.001	0.001	9.667
CDKL4	0.002	0.002	9.633
TBC1D8	0.060	0.060	9.598

(d) *Virchow2*

Nota. Los resultados corresponden al conjunto principal de todos los genes en común en las muestras.

En este caso, se observan valores elevados de T_g, principalmente en el rango de 8 a 13, lo cual corresponde a diferencias $\Delta R^2_{\mathrm{CV},g}$ cercanas a cero (del orden de 10^{-4} o menores). Esto implica que $R^2_{\mathrm{CV}}(M_A) \approx R^2_{\mathrm{CV}}(M_B)$ con gran precisión. Sin embargo, estos genes presentan

en su mayoría valores muy bajos de capacidad predictiva, con $R^2_{\text{CV}}(M_A)$ frecuentemente inferiores a 0.05, e incluso cercanos a cero. Ejemplos claros incluyen *RAB19*, *KCNK17* o *TTC30A*. Esto indica que, aunque son independientes de la condición biológica, tampoco están fuertemente asociados a la morfología.

Una excepción relevante ocurre en Virchow2, donde *UBC* aparece simultáneamente con un alto $R^2_{\text{CV}}(M_A) \approx 0.415$ y un $T_g \approx 10.538$, lo que lo posiciona como un caso ideal: altamente predecible y prácticamente independiente de la variable de cáncer. Esta tabla evidencia que la independencia respecto a z no implica necesariamente relevancia biológica o capacidad predictiva, y resalta la importancia de considerar simultáneamente ambas métricas.

Los resultados demuestran que la implementación de validación cruzada estratificada por pacientes (3 folds) reduce significativamente el efecto aparente de la variable de condición biológica sobre el desempeño predictivo. Mientras que en enfoques sin estratificación la inclusión de z puede parecer relevante, en este esquema más riguroso su contribución se vuelve marginal, con $\Delta R^2_{\text{CV},g}$ cercano a cero para la mayoría de los genes.

Frente de Pareto

Con el fin de analizar la capacidad predictiva basada en morfología y la independencia respecto a la condición biológica, se construyó un frente de Pareto considerando dos criterios: el coeficiente de determinación $R^2_{\text{CV}}(M_A)$ y la métrica T_g.

En este contexto, $R^2_{\text{CV}}(M_A)$ cuantifica qué tan bien puede explicarse la expresión génica a partir de la morfología del tejido, mientras que T_g mide el grado de independencia respecto a la variable de condición (sano/cáncer). Valores altos en ambas métricas indican genes cuya expresión es altamente predecible a partir de la estructura del tejido y, al mismo tiempo, poco dependiente del estado patológico.

De manera formal, un gen pertenece al frente de Pareto si no existe otro gen que sea simultáneamente mejor en ambos criterios. Es decir, un gen i es no dominado si no existe un gen j tal que:

$$R^2_{\text{CV},j}(M_A) \geq R^2_{\text{CV},i}(M_A) \quad \text{y} \quad T_{g,j} \geq T_{g,i} \tag{4.8}$$

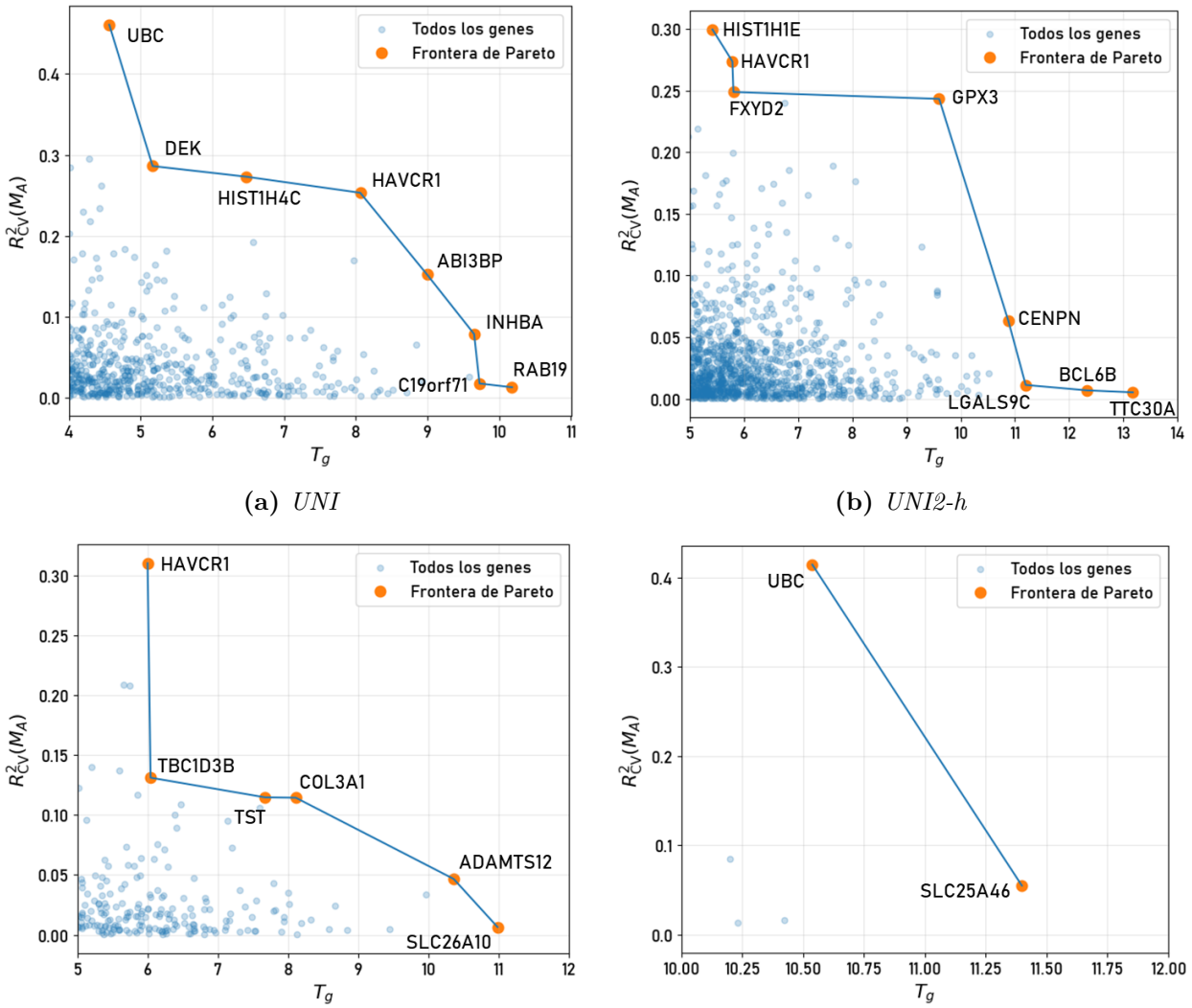
con al menos una de las desigualdades estricta.

Bajo esta definición, el frente de Pareto se compone por aquellos genes que representan soluciones óptimas en términos de compromiso: no es posible mejorar su capacidad predictiva sin disminuir su independencia respecto a la condición biológica, o viceversa. En otras palabras, estos genes maximizan simultáneamente la información morfológica relevante

y minimizan la influencia de la variable de cáncer. De esta manera, se logra identificar un subconjunto de genes robustos, cuya expresión está fuertemente determinada por la organización tisular y no por diferencias globales entre condiciones, lo que resulta clave para estudiar asociaciones intrínsecas entre morfología y expresión génica.

Figura 23

Comparación entre frentes de pareto de regresión para los modelos fundacionales utilizados.



(a) *UNI* **(b)** *UNI2-h*

(c) *Virchow* **(d)** *Virchow2*

En la Figura 23 se observa que los frentes de Pareto presentan un comportamiento consistente entre los distintos modelos, caracterizado por una relación inversa entre T_g y $R^2_{\text{CV}}(M_A)$. En general, los genes con valores muy altos de T_g (entre 10 y 13) tienden a presentar capacidades predictivas muy bajas ($R^2 < 0.05$), lo que indica que, aunque son independientes de la condición biológica, no están fuertemente asociados a la morfología. Este

patrón se repite en todos los modelos, con ejemplos como *RAB19*, *TTC30A* o *SLC26A10*.

A medida que se avanza a lo largo del frente, se observa un incremento progresivo en $R^2_{\text{CV}}(M_A)$ acompañado de una disminución en T_g. En este rango intermedio aparecen genes como *ABI3BP*, *GPX3* o *COL3A1*, que alcanzan valores de R^2 entre 0.10 y 0.25, manteniendo aún niveles relativamente altos de independencia ($T_g \approx 6$–9). Estos genes representan un balance entre ambas métricas, donde la señal morfológica comienza a ser relevante sin una fuerte dependencia de la condición. En el extremo de mayor capacidad predictiva se encuentran genes como *UBC*, *HAVCR1* o *HIST1H1E*, que alcanzan valores de R^2 entre 0.30 y 0.46. Aunque sus valores de T_g son menores en comparación con los extremos de alta independencia, se mantienen en rangos moderados (aproximadamente entre 4 y 6), lo que indica que la influencia de la variable de condición sigue siendo limitada incluso en los genes más predictivos.

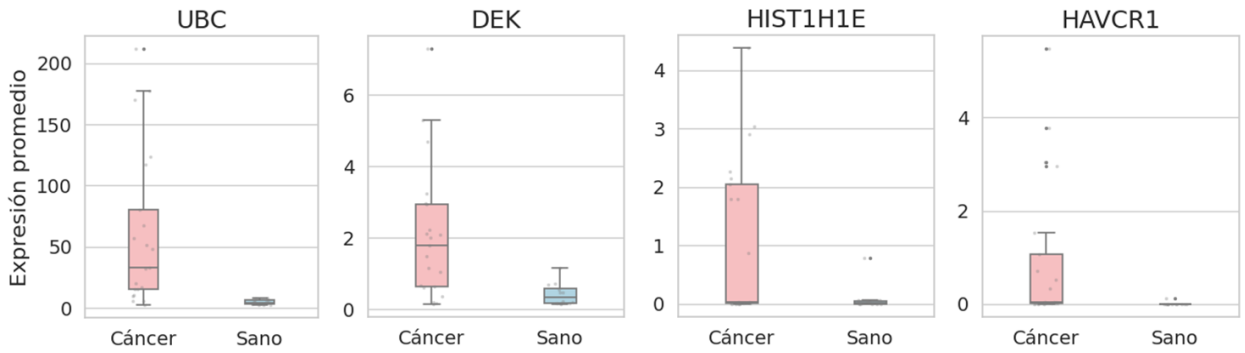
Un resultado particularmente consistente es la aparición de *UBC* en múltiples frentes de Pareto, destacando especialmente en UNI y Virchow2, donde combina valores altos de R^2 (> 0.4) con niveles elevados de T_g, lo que lo posiciona como uno de los genes más robustos en términos de asociación morfológica e independencia respecto a la condición. De manera similar, *HAVCR1* aparece de forma recurrente en varios modelos con valores intermedios-altos en ambas métricas.

Gráficos de caja de expresión promedio

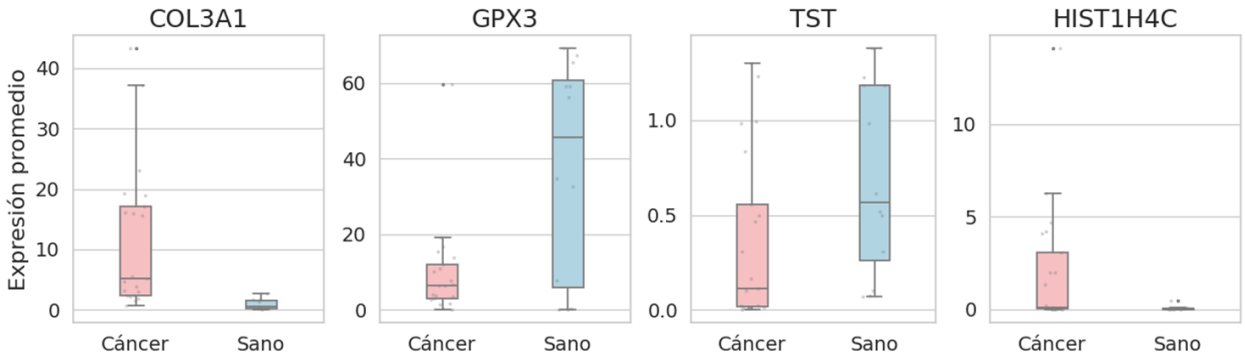
Los gráficos de caja presentados en la Figura 24 permiten complementar el análisis del frente de Pareto al visualizar directamente la distribución de la expresión génica entre las muestras de tejido sano y canceroso para genes representativos de distintos rangos de T_g. En particular, se seleccionaron genes con valores bajos, medios y altos de esta métrica, con el objetivo de contrastar empíricamente el grado de dependencia respecto a la variable de condición biológica. Dado que T_g cuantifica la independencia frente a la *etiqueta de cáncer*, se espera que genes con valores bajos presenten diferencias marcadas en expresión entre condiciones, mientras que genes con valores altos exhiban distribuciones similares, independientemente del estado patológico.

Figura 24

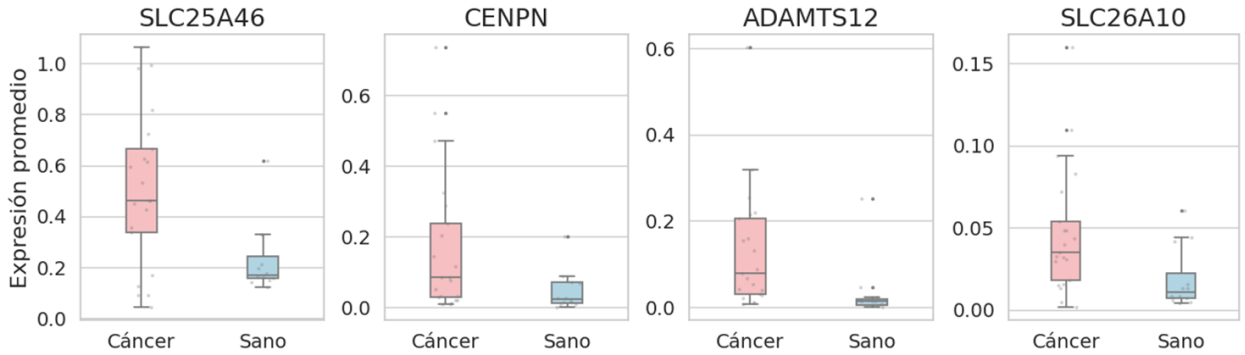
Comparación entre expresiones promedio de genes destacados del frente de pareto en tejido canceroso y sano.



(a) *Valores bajos de T_g y altos de R^2*



(b) *Valores medios de T_g y R^2*



(c) *Valores altos de T_g y bajos de R^2*

Nota. Para cada gen se realiza el gráfico de caja correspondiente a su expresión promedio por paciente tanto en muestras cancerosas ($21_{cancer} = X$) como tejido sano ($12_{sano} = Y$).

Cabe recordar que los valores de expresión mostrados en estos gráficos corresponden a conteos crudos de transcritos promediados a nivel de muestra. En particular, para cada

gen se calcula inicialmente la expresión promedio por *spot* dentro de cada muestra, y posteriormente estos valores se promedian para obtener una medida por paciente. Por tanto, las diferencias observadas en las distribuciones reflejan variaciones en la cantidad de transcritos entre regiones tisulares y condiciones biológicas.

En concordancia con el comportamiento observado en el frente de Pareto, los genes con valores bajos de T_g (Figura 24a) presentan diferencias claras en la expresión entre tejido canceroso y sano, acompañadas de valores altos de R^2. Esto indica una fuerte asociación con la morfología y una mayor dependencia de la información proporcionada por la etiqueta de condición (T_g). Por ejemplo, *UBC* alcanza medias de expresión de 60.59 en cáncer frente a 4.86 en tejido sano, con una mayor dispersión en el grupo de cáncer, coherente con su alto poder predictivo ($R^2 \approx 0.41$–0.46) y valores moderados de T_g. De forma similar, *HIST1H1E*, *DEK* y *HAVCR1* presentan distribuciones claramente separadas entre las dos condiciones, lo que confirma que, en este régimen del frente, la mejora en capacidad predictiva está asociada a la inclusión de información de la condición biológica.

Para los genes con valores intermedios de T_g (Figura 24b), se observa un comportamiento más heterogéneo que muestra el compromiso entre ambas métricas. En este caso, los valores de R^2 son moderados (≈ 0.10–0.25), lo que indica que la morfología comienza a capturar parte de la variabilidad en la expresión génica, mientras que la dependencia de la etiqueta de cáncer es menor que en el grupo anterior. Por ejemplo, *COL3A1* mantiene diferencias claras entre condiciones (media de 11.58 en cáncer frente a 0.88 en sano), coherente con su capacidad predictiva intermedia ($R^2 \approx 0.11$). En contraste, *TST* y *GPX3* presentan distribuciones más solapadas e incluso comportamientos opuestos, con mayor expresión promedio en tejido sano en el caso de *GPX3* (media de 37.66 frente a 9.82 en cáncer), lo que evidencia que la relación entre morfología y expresión es más variable en este rango del frente.

Finalmente, los genes con valores altos de T_g (Figura 24c) muestran distribuciones de expresión más similares entre tejido canceroso y sano, lo que es consistente con su independencia frente a la información aportada por la etiqueta de condición. No obstante, estos genes también presentan valores muy bajos de R^2 (< 0.06), como en *SLC26A10*, *ADAMTS12*, *CENPN* o *SLC25A46*, lo que indica que la morfología no proporciona suficiente información para estimar su expresión génica. Por ejemplo, aunque *SLC25A46* muestra diferencias moderadas en las medias (0.51 en cáncer frente a 0.23 en sano), estas significa que exista una relación predictiva fuerte con las características morfológicas. Este comportamiento confirma que valores altos de T_g no implican necesariamente una buena capacidad de modelado, sino únicamente independencia respecto a la información asociada a la condición (canceroso/sano), reforzando la relación inversa entre T_g y R^2 observada a lo largo del frente de

Pareto.

Cabe señalar que la menor cantidad de muestras de tejido sano (12 frente a 21 cancerosas) puede influir en la estabilidad de las distribuciones observadas, particularmente en la estimación de la dispersión en este grupo, aunque sin alterar las tendencias generales descritas.

5. Conclusiones y Trabajo Futuro

El presente trabajo propone un marco metodológico que permite comparar e integrar distintas representaciones morfológicas en un mismo entorno experimental, contribuyendo a la comprensión de la relación entre estructura tisular y actividad molecular. Este enfoque contribuye al desarrollo de herramientas en patología computacional orientadas a la identificación de biomarcadores y al apoyo en el análisis del microambiente tumoral.

Los resultados obtenidos en el presente trabajo evidencian que la integración de *embeddings* morfológicos con representaciones basadas en grafos mejora de manera consistente la predicción de la expresión génica. Este hallazgo confirma que la información presente en imágenes histopatológicas no se puede caracterizar únicamente a partir de la apariencia visual, sino también es ventajoso el uso de representaciones que incluyan la organización espacial y las interacciones entre células dentro del tejido.

También se resalta la importancia de conjuntos de datos abiertos como HEST-1k en metodologías que buscan asociaciones entre la información histopatológica y su transcriptómica espacial. En este trabajo se abordó tanto el tejido renal sano como el carcinoma de células renales de células claras (ccRCC) mediante la construcción de representaciones basadas en modelos fundacionales y la extracción de características nucleares tradicionales y basadas en grafos. Estas representaciones permitieron establecer un marco experimental integral para analizar la relación entre morfología y la expresión génica.

Desde una perspectiva biológica, los resultados evidencian que la relación entre morfología y expresión génica es significativa pero parcial. Es decir, si bien ciertas características visuales permiten aproximar la actividad molecular del tejido, no toda la variabilidad puede ser explicada únicamente mediante la morfología. Adicionalmente, nuestros resultados indican que la morfología puede explicar ciertos genes de una manera más clara, mientras que otros genes requieren de información adicional.

El presente estudio presenta limitaciones significativas. En primer lugar, el número reducido de muestras con datos de transcriptómica espacial, junto con la alta dimensionalidad de los perfiles génicos, puede favorecer la aparición de correlaciones espurias y limitar la capacidad de generalización de los modelos. Asimismo, la resolución de los datos de transcriptómica espacial a nivel de *spots* implica que cada medición puede contener múltiples

células, lo que introduce ambigüedad en la relación directa entre características nucleares y expresión génica. No obstante, con el fin de mitigar estos efectos, se emplearon estrategias como la validación cruzada estratificada a nivel de paciente y la comparación sistemática entre distintas representaciones, buscando reducir el sobreajuste y obtener estimaciones más robustas del desempeño.

A partir de los resultados obtenidos, surgen diversas líneas de trabajo futuro que permiten expandir y consolidar lo desarrollado en este estudio.

En primer lugar, resulta pertinente realizar una validación de la metodología propuesta utilizando conjuntos de datos adicionales, dado que la tecnología de transcriptómica espacial empleada (Visium estándar) promedia la expresión génica en *spots* multicelulares, lo cual puede introducir ruido técnico y limitar la resolución espacial. Una transición hacia tecnologías emergentes de resolución unicelular, tales como Xenium, MERSCOPE o Visium HD, representaría un avance significativo, lo que permitiría refinar la integración multimodal y evaluar la relación entre morfología y expresión génica a nivel de célula individual, reduciendo la ambigüedad presente en las mediciones actuales.

En segundo lugar, es fundamental evaluar la capacidad de generalización del enfoque propuesto. Se sugiere investigar si la superioridad observada en la integración entre modelos fundacionales y grafos celulares se mantiene en otros tipos de tejido y contextos patológicos, tales como como cáncer de mama o pulmón, lo que permitiría validar la robustez del enfoque y determinar su aplicabilidad en escenarios clínicos más amplios.

Finalmente, como proyección aplicada, se sugiere integrar este tipo de enfoques en herramientas de apoyo para patología computacional, orientadas a la identificación de biomarcadores o a la caracterización del microambiente tumoral. No obstante, esto requerirá validaciones adicionales en entornos clínicos y el uso de datos más diversos antes de su posible adopción en la práctica médica.

Referencias bibliográficas

10X GENOMICS. *Human Kidney, 11 mm Capture Area (FFPE): Spatial Gene Expression Dataset*. Ver. Space Ranger 2.0.1. Spatial Gene Expression dataset analyzed using Space Ranger 2.0.1. 2022.

AHMEDT-ARISTIZABAL, David; ARMIN, Mohammad Ali; DENMAN, Simon; FOOKES, Clinton y PETERSSON, Lars. «A survey on graph-based deep learning for computational histopathology». En: *Computerized Medical Imaging and Graphics* 95 (ene. de 2022), pág. 102027. ISSN: 08956111. DOI: 10.1016/j.compmedimag.2021.102027.

ALHALMI, Abdulsalam; ALZOBAIDI, Nafaa y ABDULRAHMAN, Amer. «Intracellular Protein Biosynthesis: A Review». En: *Asian Journal of Biochemistry, Genetics and Molecular Biology* (8 de sep. de 2020), págs. 10-18. ISSN: 2582-3698. DOI: 10.9734/ajbgmb/2020/v5i230125.

AMERICAN COLLEGE OF SURGEONS. y AMERICAN JOINT COMMITTEE ON CANCER. *AJCC cancer staging system*. Ed. por Mahul B. AMIN y Stephen B. EDGE. 8th edition, Version 9. OCLC: 1388367660. Chicago, IL: American College of Surgeons, 2024. 1 pág. ISBN: 9783319406183.

AWAIS, Muhammad; NASEER, Muzammal; KHAN, Salman; ANWER, Rao Muhammad; CHOLAKKAL, Hisham; SHAH, Mubarak; YANG, Ming-Hsuan y KHAN, Fahad Shahbaz. «Foundation Models Defining a New Era in Vision: A Survey and Outlook». En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47.4 (2025), págs. 2245-2264. DOI: 10.1109/TPAMI.2024.3506283.

BANOON, Shaima; SABHAN, Talal y GHASEMIAN, Abdolmajid. «Genetic Mutations and Major Human Disorders: A Review». En: *Egyptian Journal of Chemistry* 65 (feb. de 2022), págs. 571-589. DOI: 10.21608/EJCHEM.2021.98178.4575.

BAXI, Vipul; EDWARDS, Robin; MONTALTO, Michael y SAHA, Saurabh. «Digital pathology and artificial intelligence in translational medicine and clinical practice». En: *Modern Pathology* 35 (2022), págs. 23-32. DOI: 10.1038/s41379-021-00919-2.

BRAY, Freddie; LAVERSANNE, Mathieu; SUNG, Hyuna; FERLAY, Jacques; SIEGEL, Rebecca L.; SOERJOMATARAM, Isabelle y JEMAL, Ahmedin. «Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries». En: *CA: A Cancer Journal for Clinicians* 74 (2024), págs. 229-263. DOI: 10.3322/caac.21834.

BRUSSEE, Siemen; BUZZANCA, Giorgio; SCHRADER, Anne M.R. y KERS, Jesper. «Graph neural networks in histopathology: Emerging trends and future directions». En: *Medical Image Analysis* 101 (abr. de 2025), pág. 103444. ISSN: 13618415. DOI: 10.1016/j.media.2024.103444.

CANELA, Victor Hugo et al. «A spatially anchored transcriptomic atlas of the human kidney papilla identifies significant immune injury in patients with stone disease». En: *Nature Communications* 14.1 (19 de jul. de 2023), pág. 4140. ISSN: 2041-1723. DOI: 10.1038/s41467-023-38975-8.

CHEN, Richard J et al. «Towards a General-Purpose Foundation Model for Computational Pathology». En: *Nature Medicine* (2024).

CHEN, Tsai-Ying; YOU, Li; HARDILLO, Jose Angelito U. y CHIEN, Miao-Ping. «Spatial Transcriptomic Technologies». En: *Cells* 12 (2023), pág. 2042. DOI: 10.3390/cells12162042.

CHEN, Weiqing et al. «A visual–omics foundation model to bridge histopathology with spatial transcriptomics». En: *Nature Methods* 22 (2025), págs. 1568-1582. DOI: 10.1038/s41592-025-02707-1.

CHEN, Zheng; YANG, Ziwei; ZHU, Lingwei; SHI, Guang; YUE, Kun; MATSUBARA, Takashi; KANAYA, Shigehiko y ALTAF-UL-AMIN, M. D. *Cancer Subtyping by Improved Transcriptomic Features Using Vector Quantized Variational Autoencoder*. 20 de jul. de 2022. DOI: 10.48550/arXiv.2207.09783.

CROSS, Simon S. y UNDERWOOD, J. C. E. *Underwood's pathology : a clinical approach*. eng. Sixth edition. Edinburgh: Churchill Livingstone, 2013. ISBN: 9780702046728.

D'ARCY, Cian y KIEL, Christina. «Cell Adhesion Molecules in Normal Skin and Melanoma». En: *Biomolecules* 11.8 (15 de ago. de 2021), pág. 1213. ISSN: 2218-273X. DOI: 10.3390/biom11081213.

DELAHUNT, Brett; SRIGLEY, John R.; EGEVAD, Lars y MONTIRONI, Rodolfo. «International Society of Urological Pathology Grading and Other Prognostic Factors for Renal Neoplasia». En: *European Urology* 66.5 (nov. de 2014), págs. 795-798. ISSN: 03022838. DOI: 10.1016/j.eururo.2014.05.027.

DIESTEL, Reinhard. *Graph Theory*. Vol. 173. Graduate Texts in Mathematics. Berlin, Heidelberg: Springer Berlin Heidelberg, 2025. ISBN: 9783662701065 9783662701072. DOI: 10.1007/978-3-662-70107-2.

DONG, ZhiCheng y CHEN, Yan. «Transcriptomics: Advances and approaches». En: *Science China Life Sciences* 56 (2013), págs. 960-967. DOI: 10.1007/s11427-013-4557-2.

DOSOVITSKIY, Alexey et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV].

EVANS, D Gareth y WOODWARD, Emma R. «Genetic predisposition to cancer». En: *Medicine* 51.1 (ene. de 2023), págs. 75-79. ISSN: 13573039. DOI: 10.1016/j.mpmed.2022.10.011.

FERREIRA, Ricardo Melo et al. «Integration of spatial and single-cell transcriptomics localizes epithelial cell–immune cross-talk in kidney injury». En: *JCI Insight* 6.12 (22 de jun. de 2021), e147703. ISSN: 2379-3708. DOI: 10.1172/jci.insight.147703.

FERREIRA, Tomas y RODRIGUEZ, Santiago. «Mitochondrial DNA: Inherent Complexities Relevant to Genetic Analyses». En: *Genes* 15.5 (12 de mayo de 2024), pág. 617. ISSN: 2073-4425. DOI: 10.3390/genes15050617.

FISCHER, Andrew H.; JACOBSON, Kenneth A.; ROSE, Jack y ZELLER, Rolf. «Hematoxylin and Eosin Staining of Tissue and Cell Sections». En: *Cold Spring Harbor Protocols* 2008 (2008), pdb.prot4986. DOI: 10.1101/pdb.prot4986.

FUHRMAN, Susan A.; LASKY, Larry C. y LIMAS, Catherine. «Prognostic significance of morphologic parameters in renal cell carcinoma:» en: *The American Journal of Surgical Pathology* 6.7 (oct. de 1982), págs. 655-664. ISSN: 0147-5185. DOI: 10.1097/00000478-198210000-00007.

GADIYA, Shrey; ANAND, Deepak y SETHI, Amit. *Histographs: Graphs in Histopathology*. 2019. DOI: 10.48550/ARXIV.1908.05020.

GAFFNEY, Ef; RIEGMAN, Ph; GRIZZLE, We y WATSON, Ph. «Factors that drive the increasing use of FFPE tissue in basic and translational cancer research». En: *Biotechnic & Histochemistry* 93.5 (4 de jul. de 2018), págs. 373-386. ISSN: 1052-0295, 1473-7760. DOI: 10.1080/10520295.2018.1446101.

GAO, Yi; LIANG, Jianwen; TIAN, Mu; DENG, Wenjiang; WANG, Lei; TAO, Siyuan y MOU, Tian. «Histology image analysis of 13 healthy tissues reveals molecular-histological correlations». En: *Scientific Reports* 15.1 (23 de jul. de 2025), pág. 26812. ISSN: 2045-2322. DOI: 10.1038/s41598-025-11853-7.

GHAFOOR, M.; PARKINSON, Je.; PHAM, T.; SUTHERLAND, Te. y RATTRAY, M. *Cell–ECM Graphs: A Graph-Based Method for Joint Analysis of Cells and the Extracellular Matrix*. 4 de jun. de 2025. DOI: 10.1101/2025.06.04.657781.

GRACIA VILLACAMPA, Eva et al. «Genome-wide spatial expression profiling in formalin-fixed tissues». En: *Cell Genomics* 1.3 (dic. de 2021), pág. 100065. ISSN: 2666979X. DOI: 10.1016/j.xgen.2021.100065.

GU, Zhujie; BOUHADDANI, Said el; PEI, Jiayi; HOUWING-DUISTERMAAT, Jeanine y UH, Hae-Won. «Statistical integration of two omics datasets using GO2PLS». En: *BMC Bioinformatics* 22 (2021). DOI: 10.1186/s12859-021-03958-3.

GUTIÉRREZ GARCÍA, Ismael. *Introducción a la Teoría de Grafos: Conceptos, Algoritmos y Aplicaciones*. Col. de Yesneri Maider Zuleta SALDARRIAGA. 1st ed. Bogotá: Uninorte, Ediciones, 2024. 1 pág. ISBN: 9789587896312.

HANAHAN, Douglas. «Hallmarks of Cancer: New Dimensions». En: *Cancer Discovery* 12 (2022), págs. 31-46. DOI: 10.1158/2159-8290.CD-21-1059.

HASSAN, Taimur; LI, Zhu; JAVED, Sajid; DIAS, Jorge y WERGHI, Naoufel. «Neural Graph Refinement for Robust Recognition of Nuclei Communities in Histopathological Landscape». En: *IEEE Transactions on Image Processing* 33 (2024), págs. 241-256. ISSN: 1057-7149, 1941-0042. DOI: 10.1109/TIP.2023.3337666.

HOERL, Arthur E y KENNARD, Robert W. «Ridge regression: Biased estimation for nonorthogonal problems». En: *Technometrics* 12.1 (1970), págs. 55-67. DOI: 10.1080/00401706.1970.10488634.

HÖRST, Fabian et al. «CellViT: Vision Transformers for precise cell segmentation and classification». En: *Medical Image Analysis* 94 (mayo de 2024), pág. 103143. ISSN: 13618415. DOI: 10.1016/j.media.2024.103143.

HSIEH, James J.; PURDUE, Mark P.; SIGNORETTI, Sabina; SWANTON, Charles; ALBIGES, Laurence; SCHMIDINGER, Manuela; HENG, Daniel Y.; LARKIN, James y FICARRA, Vincenzo. «Renal cell carcinoma». En: *Nature Reviews Disease Primers* 3.1 (9 de mar. de 2017), pág. 17009. ISSN: 2056-676X. DOI: 10.1038/nrdp.2017.9.

HUANG, Zhi; BIANCHI, Federico; YUKSEKGONUL, Mert; MONTINE, Thomas J. y ZOU, James. «A visual–language foundation model for pathology image analysis using medical Twitter». En: *Nature Medicine* 29 (2023), págs. 2307-2316. DOI: 10.1038/s41591-023-02504-3.

IKEZOGWO, Wisdom Oluchi; SEYFIOGLU, Mehmet Saygin; GHEZLOO, Fatemeh; GEVA, Dylan Stefan Chan; MOHAMMED, Fatwir Sheikh; ANAND, Pavan Kumar; KRISHNA, Ranjay y SHAPIRO, Linda. *Quilt-1M: One Million Image-Text Pairs for Histopathology.* 2025. arXiv: 2306.11207 [cs.CV].

JAUME, Guillaume et al. *HEST-1k: A Dataset for Spatial Transcriptomics and Histology Image Analysis.* 2024. arXiv: 2406.16192 [cs.LG].

JI, Meng-Yao; YUAN, Lei; JIANG, Xiao-Da; ZENG, Zhi; ZHAN, Na; HUANG, Ping-Xiao; LU, Cheng y DONG, Wei-Guo. «Nuclear shape, architecture and orientation features from H&E images are able to predict recurrence in node-negative gastric adenocarcinoma». En: *Journal of Translational Medicine* 17.1 (dic. de 2019), pág. 92. ISSN: 1479-5876. DOI: 10.1186/s12967-019-1839-x.

JIA, Yuran; LIU, Junliang; CHEN, Li; ZHAO, Tianyi y WANG, Yadong. «THItoGene: a deep learning method for predicting spatial transcriptomics from histological images». En: *Briefings in Bioinformatics* 25 (2023). DOI: 10.1093/bib/bbad464.

JIANG, Lai et al. «Mitophagy and clear cell renal cell carcinoma: insights from single-cell and spatial transcriptomics analysis». En: *Frontiers in Immunology* 15 (27 de jun. de 2024), pág. 1400431. ISSN: 1664-3224. DOI: 10.3389/fimmu.2024.1400431.

JUWAYRIA; SHRIVASTAVA, Priyansh; YADAV, Kaustar; DAS, Sourabh; MITTAL, Shubham; KUMAR, Sunil; JAIN, Deepali; MALIK, Prabhat Singh y GUPTA, Ishaan. *Microarray Integrated Spatial Transcriptomics (MIST) for Affordable, Robust, and Comprehensive Digital Pathology.* 3 de jun. de 2024. DOI: 10.1101/2024.05.31.596759.

KASHIF, Muhammad Nasim; AHMED RAZA, Shan E.; SIRINUKUNWATTANA, Korsuk; ARIF, Muhammmad y RAJPOOT, Nasir. «Handcrafted features with convolutional neural networks for detection of tumor cells in histology images». En: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI).* 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI). Prague: IEEE, abr. de 2016, págs. 1029-1032. ISBN: 9781479923496. DOI: 10.1109/ISBI.2016.7493441.

KATHER, Jakob Nikolas et al. «Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study». En: *PLOS Medicine* 16 (2019), e1002730. DOI: 10.1371/journal.pmed.1002730.

KIM, Jong Wook et al. *Learning Transferable Visual Models From Natural Language Supervision.* 2021. arXiv: 2103.00020 [cs.CV].

KONSTANDINOU, Christos; GLOTSOS, Dimitris; KOSTOPOULOS, Spiros; KALATZIS, Ioannis; RAVAZOULA, Panagiota; MICHAIL, George; LAVDAS, Eleftherios; CAVOURAS, Dionisis y SAKELLAROPOULOS, George. «Multifeature Quantification of Nuclear Properties from Images of H&E-Stained Biopsy Material for Investigating Changes in Nuclear Structure with Advancing CIN Grade». En: *Journal of Healthcare Engineering* 2018 (5 de jul. de 2018), págs. 1-11. ISSN: 2040-2295, 2040-2309. DOI: 10.1155/2018/6358189.

KOUL, Bhupendra. «Types of Cancer». En: *Herbs for Cancer Treatment*. Springer Singapore, 2019, págs. 53-150. DOI: 10.1007/978-981-32-9147-8_2.

KUKURBA, Kimberly R. y MONTGOMERY, Stephen B. «RNA Sequencing and Analysis». En: *Cold Spring Harbor Protocols* 2015.11 (nov. de 2015), pdb.top084970. ISSN: 1940-3402, 1559-6095. DOI: 10.1101/pdb.top084970.

KUMAR, Neeraj; VERMA, Ruchika; CHEN, Chuheng; LU, Cheng; FU, Pingfu; WILLIS, Joseph y MADABHUSHI, Anant. «Computer-extracted features of nuclear morphology in hematoxylin and eosin images distinguish stage II and IV colon tumors». En: *The Journal of Pathology* 257.1 (mayo de 2022), págs. 17-28. ISSN: 0022-3417, 1096-9896. DOI: 10.1002/path.5864.

LAKE, Blue B. et al. «An atlas of healthy and injured cell states and niches in the human kidney». En: *Nature* 619.7970 (20 de jul. de 2023), págs. 585-594. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-023-05769-3.

LEVY-JURGENSON, Alona; TEKPLI, Xavier; KRISTENSEN, Vessela N. y YAKHINI, Zohar. «Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer». En: *Scientific Reports* 10 (2020). DOI: 10.1038/s41598-020-75708-z.

LINEHAN, W. Marston. «Genetic basis of kidney cancer: Role of genomics for the development of disease-based therapeutics». En: *Genome Research* 22.11 (nov. de 2012), págs. 2089-2100. ISSN: 1088-9051. DOI: 10.1101/gr.131110.111.

LITJENS, Geert; KOOI, Thijs; BEJNORDI, Babak Ehteshami; SETIO, Arnaud Arindra Adiyoso; CIOMPI, Francesco; GHAFOORIAN, Mohsen; LAAK, Jeroen A.W.M. van der; GINNEKEN, Bram van y SÁNCHEZ, Clara I. «A survey on deep learning in medical image analysis». En: *Medical Image Analysis* 42 (dic. de 2017), págs. 60-88. ISSN: 1361-8415. DOI: 10.1016/j.media.2017.07.005.

LONSDALE, John et al. «The Genotype-Tissue Expression (GTEx) project». En: *Nature Genetics* 45.6 (jun. de 2013). Publisher: Springer Science and Business Media LLC, págs. 580-585. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.2653.

LOU, Wei; WAN, Xiang; LI, Guanbin; LOU, Xiaoying; LI, Chenghang; GAO, Feng y LI, Haofeng. «Structure Embedded Nucleus Classification for Histopathology Images». En: *IEEE Transactions on Medical Imaging* 43.9 (sep. de 2024), págs. 3149-3160. ISSN: 0278-0062, 1558-254X. DOI: 10.1109/TMI.2024.3388328.

LU, Wenqi; TOSS, Michael; DAWOOD, Muhammad; RAKHA, Emad; RAJPOOT, Nasir y MINHAS, Fayyaz. «SlideGraph + : Whole slide image level graphs to predict HER2 status in breast cancer». En: *Medical Image Analysis* 80 (ago. de 2022), pág. 102486. ISSN: 13618415. DOI: 10.1016/j.media.2022.102486.

MEYLAN, Maxime et al. «Tertiary lymphoid structures generate and propagate anti-tumor antibody-producing plasma cells in renal cell cancer». En: *Immunity* 55.3 (mar. de 2022), 527-541.e5. ISSN: 10747613. DOI: 10.1016/j.immuni.2022.02.001.

MILLAN, Braden; LOEBACH, Lauren; BLACHMAN-BRAUN, Ruben; PATEL, Milan H.; SAINI, Jaskirat; LINEHAN, W. Marston y BALL, Mark W. «Molecular Genetics of Renal Cell Carcinoma: A Narrative Review Focused on Clinical Relevance». En: *Current Oncology* 32.6 (18 de jun. de 2025), pág. 359. ISSN: 1718-7729. DOI: 10.3390/curroncol32060359.

MIYAZAKI, Ikuko y ASANUMA, Masato. «Multifunctional Metallothioneins as a Target for Neuroprotection in Parkinson's Disease». En: *Antioxidants* 12.4 (6 de abr. de 2023), pág. 894. ISSN: 2076-3921. DOI: 10.3390/antiox12040894.

MOCH, Holger; CUBILLA, Antonio L.; HUMPHREY, Peter A.; REUTER, Victor E. y ULBRIGHT, Thomas M. «The 2016 WHO Classification of Tumours of the Urinary System and Male Genital Organs—Part A: Renal, Penile, and Testicular Tumours». En: *European Urology* 70.1 (2016), págs. 93-105. ISSN: 0302-2838. DOI: 10.1016/j.eururo.2016.02.029.

OKURA, Gillian C.; BHARADWAJ, Alamelu G. y WAISMAN, David M. «Recent Advances in Molecular and Cellular Functions of S100A10». En: *Biomolecules* 13.10 (26 de sep. de 2023), pág. 1450. ISSN: 2218-273X. DOI: 10.3390/biom13101450.

OQUAB, Maxime et al. *DINOv2: Learning Robust Visual Features without Supervision*. 2024. arXiv: 2304.07193 [cs.CV].

PDQ CANCER GENETICS EDITORIAL BOARD. «Genetics of Renal Cell Carcinoma (PDQ®): Health Professional Version». En: *PDQ Cancer Information Summaries*. Bethesda (MD): National Cancer Institute (US), 2002.

POREBSKI, Alice; VANDENBROUCKE, Nicolas y MACAIRE, Ludovic. «Haralick feature extraction from LBP images for color texture classification». En: *2008 First Workshops on Image Processing Theory, Tools and Applications*. 2008 First Workshops on Image Processing Theory, Tools and Applications. ISSN: 2154-512X. Nov. de 2008, págs. 1-8. DOI: 10.1109/IPTA.2008.4743780.

RADFORD, Alec et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV].

RAO, Anjali; BARKLEY, Dalia; FRANÇA, Gustavo S. y YANAI, Itai. «Exploring tissue architecture using spatial transcriptomics». En: *Nature* 596 (2021), págs. 211-220. DOI: 10.1038/s41586-021-03634-9.

RAZA, Manahil; AZAM, Ayesha; QAISER, Talha y RAJPOOT, Nasir. *PS3: A Multimodal Transformer Integrating Pathology Reports with Histology Images and Biological Pathways for Cancer Survival Prediction*. 2025. arXiv: 2509.20022 [cs.CV].

ROMO-BUCHELI, David; CORREDOR, Germán; GARCÍA-ARTEAGA, Juan D.; ARIAS, Viviana y ROMERO, Eduardo. «Nuclei graph local features for basal cell carcinoma classification in whole slide images». En: 12th International Symposium on Medical Information Processing and Analysis. Ed. por Eduardo ROMERO; Natasha LEPORE; Jorge BRIEVA e Ignacio LARRABIDE. Tandil, Argentina, 27 de ene. de 2017, 101600Q. DOI: 10.1117/12.2257386.

SHULMAN, Eldad D. et al. «AI-Driven Spatial Transcriptomics Unlocks Large-Scale Breast Cancer Biomarker Discovery from Histopathology». En: (2024). DOI: 10.1101/2024.10.16.618609.

SMIT, Paul. *Kidney Cancer – Dr Paul Smit*.

SONG, Andrew H.; JAUME, Guillaume; WILLIAMSON, Drew F. K.; LU, Ming Y.; VAIDYA, Anurag; MILLER, Tiffany R. y MAHMOOD, Faisal. «Artificial intelligence for digital and computational pathology». En: *Nature Reviews Bioengineering* 1.12 (oct. de 2023), págs. 930-949. ISSN: 2731-6092. DOI: 10.1038/s44222-023-00096-8.

STÅHL, Patrik L. et al. «Visualization and analysis of gene expression in tissue sections by spatial transcriptomics». En: *Science* 353 (2016), págs. 78-82. DOI: 10.1126/science.aaf2403.

STARK, Rory; GRZELAK, Marta y HADFIELD, James. «RNA sequencing: the teenage years». En: *Nature Reviews Genetics* 20 (2019), págs. 631-656. DOI: 10.1038/s41576-019-0150-2.

SUMON, Rashadul Islam; MOZUMDAR, Md Ariful Islam; AKTER, Salma; UDDIN, Shah Muhammad Imtiyaj; AL-ONAIZAN, Mohammad Hassan Ali; ALKANHEL, Reem Ibrahim y MUTHANNA, Mohammed Saleh Ali. «Comparative Study of Cell Nuclei Segmentation Based on Computational and Handcrafted Features Using Machine Learning Algorithms». En: *Diagnostics* 15 (2025), pág. 1271. DOI: 10.3390/diagnostics15101271.

TRICHOPOULOS, Dimitrios; LI, Frederick P. y HUNTER, David J. «What Causes Cancer?» En: *Scientific American* 275.3 (1996), págs. 80-87. ISSN: 00368733, 19467087.

UNGER, Michaela y KATHER, Jakob Nikolas. «Deep learning in cancer genomics and histopathology». En: *Genome Medicine* 16 (2024). DOI: 10.1186/s13073-024-01315-6.

VILLA DÍAZ, Matías Joaquín. «Diseño de soluciones de visión computacional y aprendizaje profundo para la detección de cáncer mediante imágenes histopatológicas». En: (2025). DOI: 10.58011/sxeq-0c18.

VILLANUEVA, Lorea; ÁLVAREZ-ERRICO, Damiana y ESTELLER, Manel. «The Contribution of Epigenetics to Cancer Immunotherapy». En: *Trends in Immunology* 41 (2020), págs. 676-691. DOI: 10.1016/j.it.2020.06.002.

VORONTSOV, Eugene et al. «A foundation model for clinical-grade computational pathology and rare cancers detection». En: *Nature Medicine* (2024).

WANG, Jianjun. «Biostatistical Challenges in High-Dimensional Data Analysis: Strategies and Innovations». En: *Computational Molecular Biology* (2024). DOI: 10.5376/cmb.2024.14.0019.

WANG, Yiwen y LÊCAO, Kim-Anh. «Managing batch effects in microbiome data». En: *Briefings in Bioinformatics* 21.6 (1 de dic. de 2020), págs. 1954-1970. ISSN: 1477-4054. DOI: 10.1093/bib/bbz105.

WANG, Yongquan et al. «Gene mutation profiling and clinical significances in patients with renal cell carcinoma». En: *Clinics* 78 (1 de ene. de 2023). ISSN: 1807-5932. DOI: 10.1016/j.clinsp.2023.100259.

WEN, Si; KURC, Tahsin M.; GAO, Yi; ZHAO, Tianhao; SALTZ, Joel H. y ZHU, Wei. «A Methodology for Texture Feature-based Quality Assessment in Nucleus Segmentation of

Histopathology Image». En: *Journal of Pathology Informatics* 8.1 (ene. de 2017), pág. 38. ISSN: 21533539. DOI: 10.4103/jpi.jpi_43_17.

XIE, Deqian; LI, Guandu; ZHENG, Zunwen; ZHANG, Xiaoman; WANG, Shijin; JIANG, Bowen; LI, Xiaorui; WANG, Xiaoxi y WU, Guangzhen. «The molecular code of kidney cancer: A path of discovery for gene mutation and precision therapy». En: *Molecular Aspects of Medicine* 101 (feb. de 2025), pág. 101335. ISSN: 00982997. DOI: 10.1016/j.mam.2024.101335.

XU, Hang et al. «Unsupervised spatially embedded deep representation of spatial transcriptomics». En: *Genome Medicine* 16 (2024). DOI: 10.1186/s13073-024-01283-x.

ZAHEDI, Roxana; GHAMSARI, Reza; ARGHA, Ahmadreza; MACPHILLAMY, Callum; BEHESHTI, Amin; ALIZADEHSANI, Roohallah; LOVELL, Nigel H; LOTFOLLAHI, Mohammad y ALINEJAD-ROKNY, Hamid. «Deep learning in spatially resolved transcriptomics: a comprehensive technical view». En: *Briefings in Bioinformatics* 25 (2024). DOI: 10.1093/bib/bbae082.

ZHANG, Yuqing; PARMIGIANI, Giovanni y JOHNSON, W Evan. «ComBat-seq: batch effect adjustment for RNA-seq count data». En: *NAR Genomics and Bioinformatics* 2.3 (1 de sep. de 2020), lqaa078. ISSN: 2631-9268. DOI: 10.1093/nargab/lqaa078.

ZHONG, Ziming et al. «The distinct roles of genome, methylation, transcription, and translation on protein expression in Arabidopsis thaliana resolve the Central Dogma's information flow». En: *Genome Biology* 26 (2025). DOI: 10.1186/s13059-025-03741-0.

ZHOU, Yanning; GRAHAM, Simon; KOOHBANANI, Navid Alemi; SHABAN, Muhammad; HENG, Pheng-Ann y RAJPOOT, Nasir. *CGC-Net: Cell Graph Convolutional Network for Grading of Colorectal Cancer Histology Images*. 2019. DOI: 10.48550/ARXIV.1909.01068.

ZIMMERMANN, Eric et al. *Virchow2: Scaling Self-Supervised Mixed Magnification Models in Pathology*. 2024. arXiv: 2408.00738 [cs.CV].

Apéndices

Apéndice A
Artículo científico: Integrating nuclear graph interaction features and foundation model's embeddings in histopathological images

Abstract

In recent years, spatial transcriptomics has emerged as a novel technology that enables the examination of gene expression within tissue, taking into account its original structure and spatial context. This combination of visual and molecular data offers new possibilities for understanding tumor heterogeneity and advancing personalized medicine methods. On the other hand, foundation deep learning models are becoming a highly precise representation computational tools that capture the characteristics of the histological images. In this work, we investigate how cellular shapes and spatial patterns observable in hematoxylin and eosin images, represented through foundation model embeddings and nuclei-based graphs, are associated with underlying genetic processes. These associations are examined on a subset of the HEST-1K dataset, which integrates morphological features and spatial transcriptomic profiles across multiple tissue types. Specifically for this work, we conducted a case study with 20 samples (patients) with cell renal cell carcinoma (kidney cancer, ccRCC). The consistency between morphological patterns and gene expression profiles is evaluated aiming to provide evidence of the relationship between tissue structure and molecular activity. Our results indicate that incorporating nuclei interaction graphs enhance the ability of foundation models to find associations with the genomic expression, both across the full set of genes and within selected gene subsets.

Referencias del producto científico

ROBLES-ARDILA, Jorge; JAIMES-RODRIGUEZ, Deciré y ROMO-BUCHELI, David. «Integrating nuclear graph interaction features and foundation model's embeddings in histopathological images: Regularized association analysis with spatial gene expression». En: (2026). Sometido a 48th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2026).

Apéndice B
Herramienta de software para visualización de transcriptómica espacial

Como complemento al presente trabajo, se desarrolló una herramienta web interactiva para la visualización y exploración de imágenes histopatológicas tipo Whole Slide Image (WSI) y sus datos asociados. Esta herramienta fue implementada utilizando el framework *Flask* en Python, y permite integrar en una misma interfaz distintos resultados generados a lo largo del estudio.

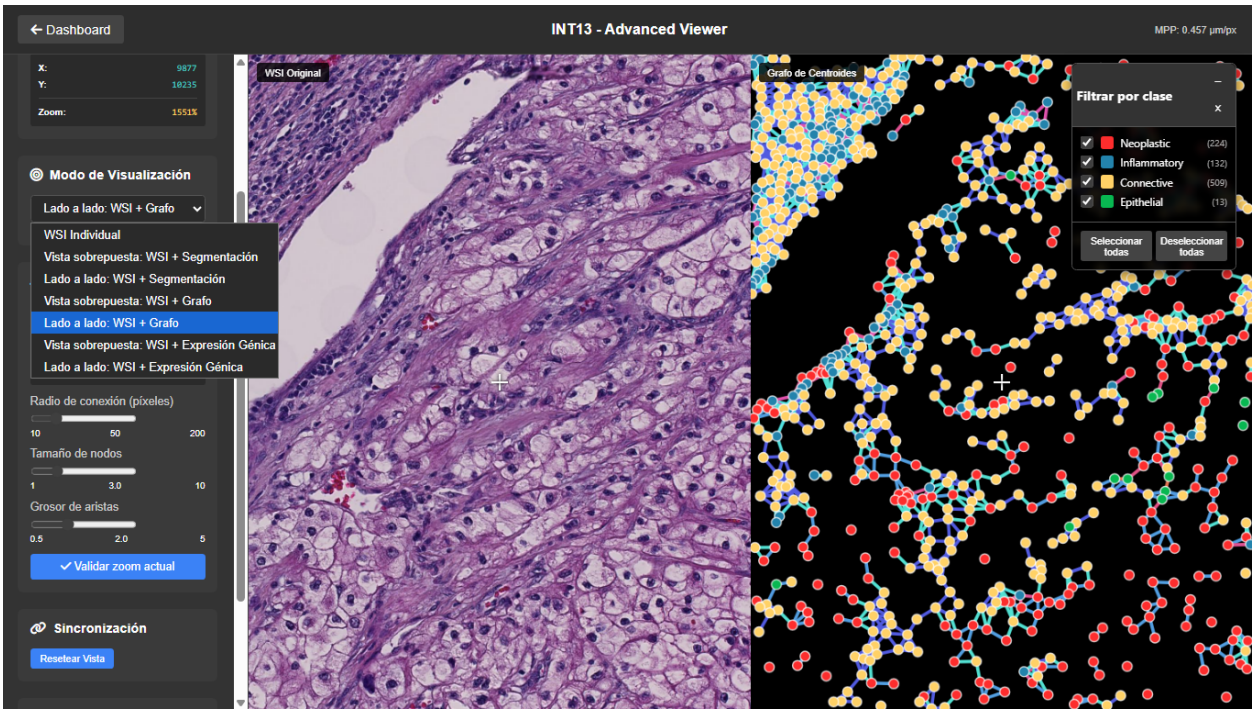
Entre las funcionalidades principales del sistema se incluyen:

- Visualización de imágenes WSI con capacidad de navegación a múltiples niveles de zoom.
- Visualización de segmentaciones nucleares del tejido histológico.
- Visualización de la representación de grafos de interacción celular derivados de las estructuras nucleares, con opciones para la variación en el tamaño de nodos, aristas y el radio de conexión entre núcleos. También se ofrecen filtros para elegir qué tipos de células observar en el grafo construido.
- Visualización de la expresión génica espacial asociada a regiones específicas del tejido, mediante la graficación de *spots* sobre el tejido y su coloreado a modo de mapa de calor para cuantificar la expresión génica de un gen previamente seleccionado.

La herramienta ofrece dos modos de visualización de las funcionalidades mencionadas previamente: vista sobrepuesta (se muestra la segmentación nuclear, el grafo celular o la expresión génica sobre el tejido original) o vista lado a lado (se ubica la muestra original a la izquierda y la función elegida a la derecha). Ambos modos de visualización se sincronizan en tiempo real a medida que se navega en la imagen histológica original, haciendo uso de las segmentaciones nucleares en formato *.parquet* y las matrices de expresión génica en formato *.h5ad* provistas por el dataset HEST-1K. A continuación se muestra en la Figura 25 la funcionalidad de vista lado a lado para la representación basada en grafos celulares.

Figura 25

Herramienta de visualización de transcriptómica espacial, segmentación nuclear y grafos celulares.



Esta herramienta permite una mejor interpretación de los resultados obtenidos, al proporcionar un entorno interactivo donde se pueden explorar las relaciones entre la estructura tisular y la actividad molecular.

Tecnologías utilizadas: Python, Flask, OpenSlide y librerías de visualización.