



UNIVERSITAT DE  
BARCELONA

## Desarrollo de herramientas para el análisis y predicción patogénica de las variantes *missense* de *ATM* en el entorno clínico

Luz Marina Porras Monroy

**ADVERTIMENT.** La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX ([www.tdx.cat](http://www.tdx.cat)) i a través del Dipòsit Digital de la UB ([diposit.ub.edu](http://diposit.ub.edu)) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX ni al Dipòsit Digital de la UB. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX o al Dipòsit Digital de la UB (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

**ADVERTENCIA.** La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR ([www.tdx.cat](http://www.tdx.cat)) y a través del Repositorio Digital de la UB ([diposit.ub.edu](http://diposit.ub.edu)) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR o al Repositorio Digital de la UB. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR o al Repositorio Digital de la UB (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

**WARNING.** On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX ([www.tdx.cat](http://www.tdx.cat)) service and by the UB Digital Repository ([diposit.ub.edu](http://diposit.ub.edu)) has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized nor its spreading and availability from a site foreign to the TDX service or to the UB Digital Repository. Introducing its content in a window or frame foreign to the TDX service or to the UB Digital Repository is not authorized (framing). Those rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.



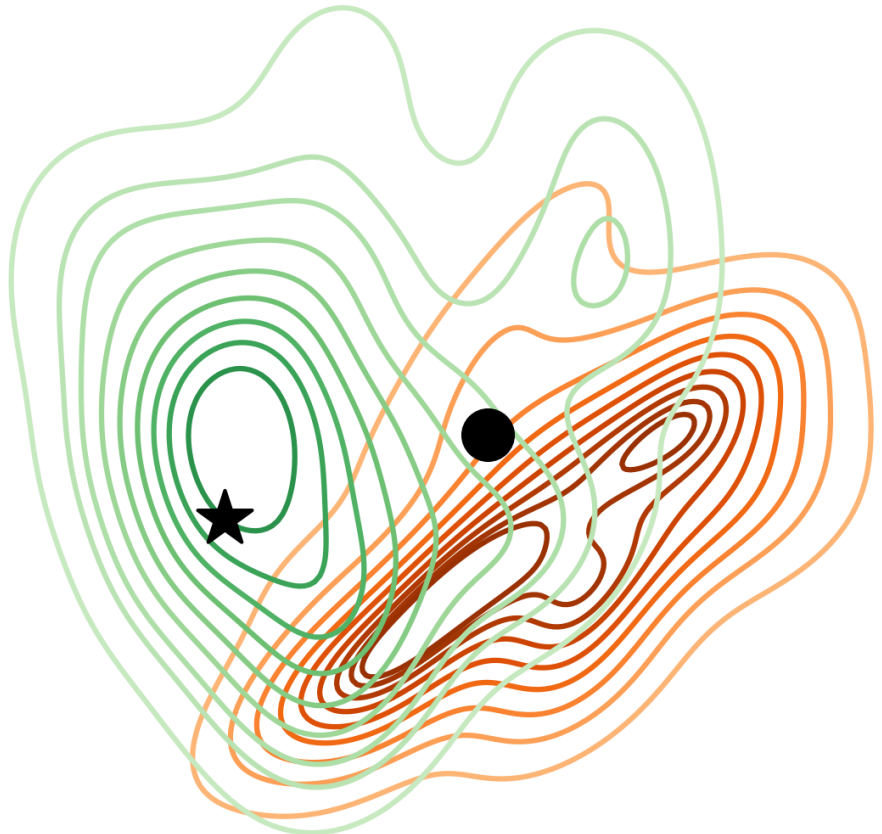
UNIVERSITAT DE  
BARCELONA

DESARROLLO DE HERRAMIENTAS PARA EL  
ANÁLISIS Y PREDICCIÓN PATOGENICA DE LAS  
VARIANTES *MISSENSE* DE *ATM* EN EL  
ENTORNO CLÍNICO

**LUZ MARINA PORRAS MONROY**

*Tesis de Doctorado 2022*

*Director, Xavier del Cruz*

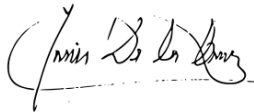


Tesi Doctoral  
Universitat de Barcelona

DESARROLLO DE HERRAMIENTAS PARA EL  
ANÁLISIS Y PREDICCIÓN PATOGENICA DE LAS  
VARIANTES *MISSENSE* DE *ATM* EN EL  
ENTORNO CLÍNICO

Memòria presentada per  
**LUZ MARINA PORRAS MONROY**

Per optar al grau de  
*Doctora per la Universitat de Barcelona*  
*Programa de Genètica*  
*Vall d'Hebron Institut de Recerca*

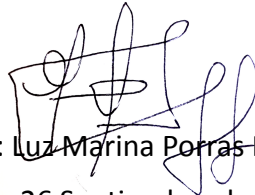


Director: Xavier del Cruz

MARTA  
PASCUAL  
BERNIOLA - DNI  
78069688Y  
78069688Y  
Fecha: 2022.09.27  
10:04:09 +02'00'

Firmado digitalmente  
por MARTA PASCUAL  
BERNIOLA - DNI  
78069688Y  
Fecha: 2022.09.27  
10:04:09 +02'00'

Tutora: Marta Pascual Berniola



Doctoranda: Luz Marina Porras Monroy

Barcelona, 26 Septiembre de 2022



UNIVERSITAT DE  
BARCELONA

# DECLARACIÓN

Declaro que la tesis titulada “DESARROLLO DE HERRAMIENTAS PARA EL ANÁLISIS Y PREDICCIÓN PATOGENICA DE LAS VARIANTES MISSENSE DE ATM EN EL ENTORNO CLÍNICO” ha sido desarrollada completamente por mí. En esta tesis se encuentran claramente referenciadas las figuras y afirmaciones que han sido obtenidas de otros autores.

Este trabajo se ha realizado en el grupo de investigación Bioinformática Clínica y Traslacional en el Vall d'Hebron Institut de Recerca bajo la dirección del Dr. Xavier de la Cruz Montserrat. No ha sido y no será presentado, como parte, para cualquier otro grado.

Barcelona, 25 de Septiembre 2022



LUZ MARINA PORRAS MONROY

*A mi estrella mi hija*

## AGRADECIMIENTOS

A mi hija Luz Helena por toda la paciencia y adaptación en este proyecto. A mi hermano Luis Carlos por ser mi primer tutor en la academia, mi guía y mi mejor amigo, este es un producto más de tu presencia en mi vida. A mis padres por todo su apoyo incondicional, no importa a donde vaya ustedes siempre están a mi lado.

A mi director de tesis, Dr. Xavier de la Cruz, por guiarme en este proceso de aprendizaje, admiro la calidad humana con la cual motivas y enseñas. Por aportar todo tu conocimiento sin restricción alguna, infinitas gracias.

A mi tutora Dra. Marta Pascual, muchas gracias por estar pendiente en todo momento de mi evolución académica y por recordarme cada paso necesario para llegar a esta meta.

A mis compañeras, inolvidables y buenas amistades que he logrado establecer en el grupo de investigación. Natàlia una mujer brillante tanto en la academia como en lo personal, gracias por todo el apoyo, tiempo y enseñanzas. Selen, gracias por los consejos, por escucharme y subir mi ánimo cuando estaba en momentos difíciles. En especial gracias a las dos por seguir las locuras y juegos a mi hija.

Al Consorcio español para la interpretación de las variantes genéticas en el gen *ATM* por permitirme participar en sus proyectos de investigación y sus reuniones llenas de conocimiento que aportaron a mi formación académica.

A Colciencias y el Departamento de Santander Colombia, por la financiación por medio de la beca : «formación de profesionales de alto nivel, Doctorado en el exterior. Convocatoria 771».

# RESUMEN

Establecer la naturaleza patogénica de las variantes de secuencia en *ATM*, un gen asociado con el cáncer de mama y otros cánceres hereditarios, es crucial para brindar una atención adecuada a los pacientes. Sin embargo, lograr buenas clasificaciones de estas variantes sigue siendo un problema sin resolver. Aquí, abordamos este problema mejorando la contribución de las herramientas *in silico* a la clasificación de variantes missense.

Un problema importante en el uso de herramientas *in silico* es su baja interpretabilidad. Nos acercamos a esta limitación explorando primero el desarrollo de predictores de patogenicidad específicos de proteínas que incorporan medidas interpretables del impacto de una variante. Paralelamente, desarrollamos una familia de representaciones gráficas rápidas e intuitivas en las que se considera el impacto de una variante en relación al de variantes patogénicas y benignas ya conocidas. Entre los resultados obtenidos destaca un sistema de clasificación para el criterio de predicción *in silico* de la adaptación de las guías ACMG/AMP a *ATM* y tres predictores específicos de proteína con diferentes grados de interpretabilidad. Dos tienen capacidades predictivas similares a las de los predictores de patogenicidad de alto rango. Curiosamente, aunque menos preciso, nuestro predictor



biofísico alcanza un rendimiento que abre el camino para usar evidencia biofísica para completar la anotación de variantes de *ATM*.

En lo que respecta a las representaciones gráficas, las hemos aplicado a tres problemas de clasificación de variantes: evaluación de predicciones, caracterización de VUS y comparación de las dos versiones de las guías ACMG/AMP adaptadas a *ATM*. En estas aplicaciones, nuestros gráficos muestran sus virtudes como herramientas complementarias para los procesos de clasificación *in silico*: son rápidos, fáciles de usar y casi no requieren capacitación.

En resumen, en esta tesis presentamos una familia de herramientas *in silico* para mejorar la anotación de variantes missense en *ATM* y facilitar el papel de los profesionales en este proceso.

# ABSTRACT

Establishing the pathogenic nature of sequence variants in *ATM*, a gene associated with breast cancer and other hereditary cancers, is crucial to providing adequate patient care. However, achieving good rankings for these variants remains an unresolved issue. Here, we address this problem by improving the contribution of *in silico* tools to missense variant classification. A major problem in the use of *in silico* tools is their low interpretability. We approach this limitation by first exploring the development of protein-specific pathogenicity predictors that incorporate interpretable measures of the impact of a variant. In parallel, we developed a family of fast and intuitive graphical representations in which the impact of a variant is considered in relation to that of known pathogenic and benign variants. Among the results obtained, a classification system stands out for the *in silico* prediction criterion of the adaptation of the ACMG/AMP guides to *ATM* and three protein-specific predictors with different degrees of interpretability. Two have predictive abilities similar to those of high-rank pathogenicity predictors. Interestingly, although less accurate, our biophysical predictor achieves a performance that paves the way for using biophysical evidence to complete *ATM* variant annotation. Regarding the graphical representations, we have applied them to three variant classification problems: prediction evaluation, VUS characterization and

comparison of the two versions of the ATM-adapted ACMG/AMP guides. In these applications, our graphs show their virtues as complementary tools for in silico classification processes: they are fast, easy to use and require almost no training. In summary, in this thesis we present a family of in silico tools to improve the annotation of missense variants in ATM and facilitate the role of professionals in this process.



patogenicidad	59
3.1.1. MSA	60
3.1.2. Comparación con el modelo de frecuencias de variantes benignas	62
3.1.3. Variantes de entrenamiento	64
<b>4. DESARROLLO DE UNA METODOLOGÍA <i>in silico</i> PARA LA ADAPTACIÓN DE LAS GUÍAS ACMG/AMP AL GEN <i>ATM</i></b>	<b>66</b>
4.1. Introducción	67
4.2. Materiales y métodos específicos	69
4.2.1. Metodología adaptada de Hart <i>et al.</i>	69
4.2.1.1. Predictores primarios	69
4.2.1.2. Optimización de los puntos de corte	70
4.2.1.3. Metapredictor Naïve Voting Method (NVM)	71
4.2.1.4. Metapredictor Random Forest (RF)	72
4.2.1.5. Variantes	72
4.2.2. Metodología final aplicada al caso de ATM	72
4.2.2.1. Predictores primarios analizados	73
4.2.2.2. Variantes	73
4.3. Resultados	74
4.3.1. Adaptación Hart <i>et al.</i>	74
4.3.1.1. Metapredictores NVM Y RF	74
4.3.1.2. Predicción sistemática de variantes missense con NVM	77
4.3.1.3. Visualización estructural	80

4.3.1.4. Posible existencia de un sesgo de dominios en NVM	81
4.3.2. Combinación de predictores	83
4.3.2.1. N-terminal	84
4.3.2.2. C-terminal	88
4.4. Discusión	91
4.4.1 Metodología Hart	91
4.4.2 Combinación de predictores	91
<b>5. HACIA UNA MEJOR CLASIFICACIÓN <i>IN SILICO</i> DE LAS VARIANTES MISSENSE DE ATM UTILIZANDO HERRAMIENTAS DE ANÁLISIS VISUAL</b>	<b>94</b>
5.1. Introducción	95
5.2. Materiales y métodos específicos	98
5.2.1. Características de entrenamiento	98
5.2.1.1. RF_Biophys	98
5.2.1.2. RF_Bioinf	99
5.2.1.1. RF_Metap	100
5.2.2. Algoritmo Random Forest	100
5.2.3. Evaluación comparativa de la capacidad predictiva	102
5.2.4. Cálculo de Fiabilidad	103
5.2.5. Análisis de Componentes Principales (PCA)	106
5.2.6. Representaciones gráficas	107
5.3. Resultados	107
5.3.1. Predictores <i>in silico</i> para ATM	107
5.3.1.1. Capacidad predictiva de nuestros predictores	

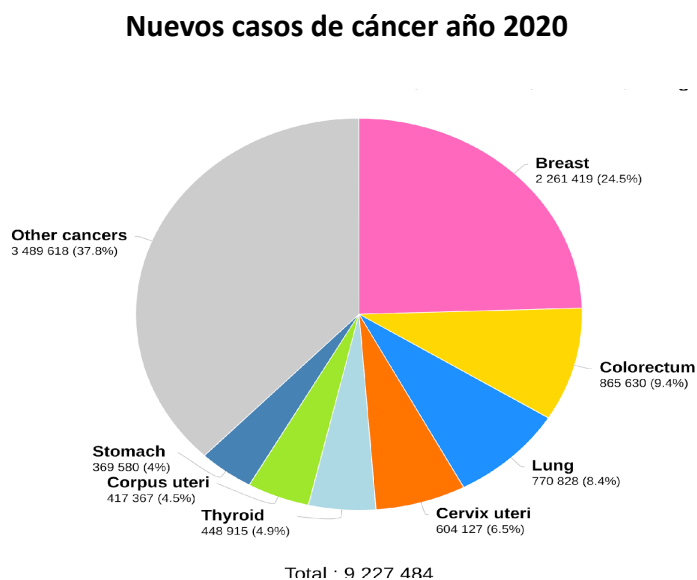
RF	108
5.3.1.2. Comparación de predictores específicos de <i>ATM</i> con predictores generales	112
5.3.1.3. Análisis de fiabilidad de las predicciones	115
5.3.2. Análisis gráfico de variantes	116
5.3.2.1. Gráficas unidimensionales	117
5.3.2.2. Gráficos de contorno como herramientas complementarias para predictores de patogenicidad	119
5.3.2.3. Priorización de VUS	123
5.3.2.4. Comparación de las guías ACMG/AMP adaptadas a <i>ATM</i>	126
5.4. Discusión	130
<b>6.DISCUSIÓN GENERAL</b>	<b>135</b>
<b>7.CONCLUSIONES</b>	<b>141</b>
<b>8. REFERENCIAS</b>	<b>144</b>
<b>9. ANEXOS</b>	<b>156</b>

# 1 INTRODUCCIÓN



## 1.1 Cáncer de Mama y ATM

El cáncer de mama es el tipo de cáncer con mayor prevalencia en el mundo, tan solo en el 2020 se estimó que 7.8 millones de mujeres se encontraban diagnosticadas con este tipo de cáncer según la Organización Mundial de la Salud (OMS). El cáncer de mama se ha convertido en el cáncer más diagnosticado a nivel mundial (Figura 1.1) (Zhu SY. *et al.*, 2022) y se ha mantenido por más de dos décadas como la principal causa de muerte por cáncer en las mujeres, sumando, tan solo para el año 2020, 685.000 muertes en el mundo (Piruzan E. *et al.*, 2021).

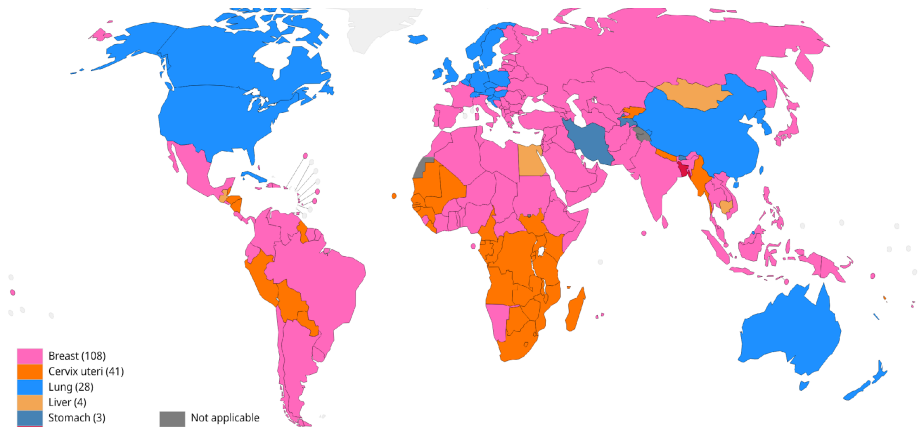


**Figura 1.1. Estadística de nuevos casos de cáncer en el año 2020 en mujeres.** Tomada de Global Cancer Observatory <http://gco.iarc.fr>

A pesar del panorama desalentador ante esta enfermedad, el esfuerzo en investigación a nivel mundial que se ha llevado a cabo en las últimas décadas ha tenido un impacto positivo en las posibilidades de sobrevivir a esta enfermedad. Los avances generados han reflejado una disminución en la mortalidad por cáncer de mama; sin embargo, son mejoras cuya implementación requiere inversión económica y sistemas de salud bien estructurados. Por tal razón, el acceso a muchos de los nuevos avances que se han establecido no ha sido igual en todos los países del mundo (Wojtyla *et al.*, 2021). La mortalidad por cáncer de mama ha disminuido sustancialmente en los países de altos ingresos como América del Norte y Australia, pero las tendencias han sido menos consistentes en América Latina y Asia, esta situación se ve reflejada en la Figura 1.2 (Wojtyla C. *et al.*, 2021).

Ante este inconveniente, se han establecido estudios de coste para el control del cáncer de mama donde se priorizan los países en desarrollo y de bajos ingresos, y cuyo objetivo es que muchos de estos países logren ampliar los programas nacionales de atención del cáncer (Wojtyla C. *et al.*, 2017). Por consiguiente, investigaciones que permitan priorizar los estudios, variantes genéticas, y tratamientos que se deben realizar y considerar en esta enfermedad son de vital importancia para mejorar la atención de las pacientes diagnosticadas con cáncer de mama en el mundo.

## Tasa de mortalidad por cáncer en el mundo



**Figura 1.2. Cáncer que presenta la mayor mortalidad por país en mujeres.** Tomada de Global Cancer Observatory <http://gco.iarc.fr>

### 1.1.1 Genética y su relación con cáncer de mama

El diagnóstico temprano y preciso son los principales hechos por los cuales se ha logrado una reducción en la tasa de mortalidad por cáncer de mama. Ello ha sido posible gracias a los recientes avances en genómica que han permitido establecer una nueva clasificación molecular de este cáncer, posibilitando un diagnóstico basado en el tipo de mutación, lo que ha hecho de la genómica una herramienta esencial para decidir el tratamiento y el seguimiento de los pacientes (Merino Bonilla JA. *et al.* 2017).

Durante la última década, las mutaciones en los genes *BRCA1* y *BRCA2* (*BRCA1/2*) se han considerado y examinado para la detección

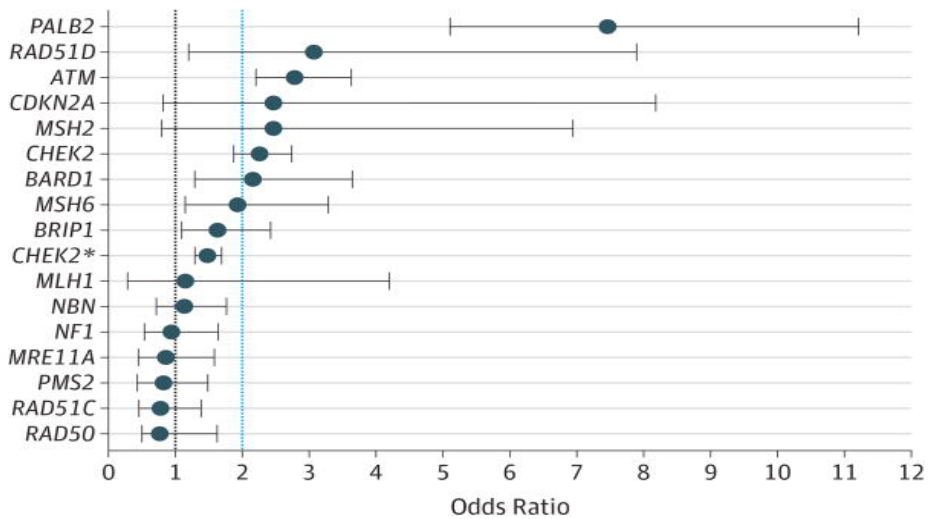
temprana de cáncer de mama hereditario, con unos resultados valiosos, ya que se ha mostrado que las mutaciones en estos genes representan el 5 % del cáncer de mama hereditario.(Moslemi *et al.*, 2021).

Sin embargo, estudios poblacionales han mostrado que variantes patogénicas en otros genes, como *CHEK2*, *ATM*, *BRIP1 (FANCI)*, *PALB2 (FANCN)*, y *RAD51C (FANCO)*, están relacionadas con el desarrollo de cáncer de mama (Filippini *et al.*, 2013).

En esta tesis centraremos nuestros esfuerzos en las variantes del gen Ataxia-Telangiectasia Mutated (*ATM*) que se encuentran con frecuencia entre las mujeres con este tipo de cáncer (Moslemi *et al.*, 2021). Desde hace más de cuarenta años se reconoció por primera vez que las mujeres en familias que presentaban variantes patogénicas de *ATM* tenían una incidencia elevada de cáncer de mama (Bernstein JL. *et al.*, 2017 ). Esta contribución fue confirmada en análisis posteriores realizados con paneles multigénicos centrados en cáncer de mama hereditario. En dichos análisis se hallaron que un 6.2% de mujeres diagnosticadas con este cáncer presentaban variantes patogénicas en genes diferentes a *BRCA1* y *BRCA2*, destacando *ATM* como uno de los genes que aumenta el riesgo a padecer cáncer de mama (Figura 1.3) (Couch *et al.*, 2017).

Otra de las evidencias que relaciona *ATM* con cáncer de mama son los

estudios en familias doble negativo para variantes de riesgo de los genes *BRCA1-BRCA2* y con alta probabilidad a padecer cáncer de mama hereditario, en las que se encontró que *ATM* es el segundo gen más frecuente en la población de alto riesgo después de *CHEK2*. En este estudio también se estableció que la prevalencia de variantes patogénicas de *ATM* en pacientes diagnosticadas con este cáncer es del 0.9% (Jerzak *et al.*, 2018) mientras que la prevalencia de variantes patogénicas de *ATM* en la población adulta es del 1-2 %. Lo cual quiere decir que más del 50 % de personas que presentan este tipo de variantes terminaron desarrollando cáncer de mama (Jerzak *et al.*,2018).



**Figura 1.3. Grado de asociación de variantes patogénicas en cada gen y cáncer de mama en mujeres con cáncer de mama.**

Tomado de: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5599323/>

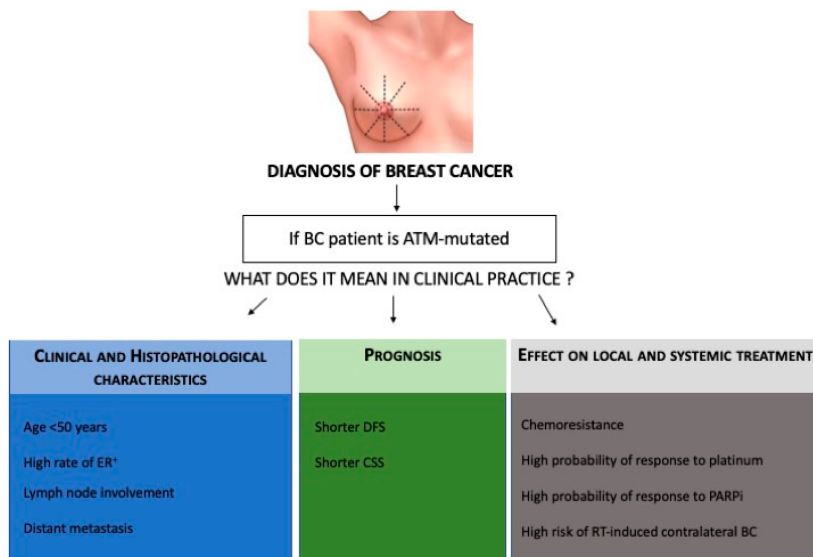
En paralelo, los análisis epidemiológicos han estimado que los portadores heterocigotos de una variante patogénica *ATM* tienen un riesgo de 2 a 3 veces mayor a padecer cáncer de mama. Asumiendo que el riesgo inicial a sufrir esta enfermedad es de aproximadamente un 10 %, el aumento del riesgo por presentar las variantes patogénicas de *ATM* significa un riesgo de por vida a presentar cáncer de mama del 20 % al 30 % (Petracci E. *et al.*, 2011).

En el escenario de las pacientes ya diagnosticadas, aquellas mujeres que padecen cáncer de mama y son portadoras de variantes patogénicas en *ATM* son las que exhiben los peores pronósticos y mayor dificultad para el tratamiento. De hecho, estudios de seguimiento a pacientes sometidas a radioterapia han mostrado que

estas mujeres desarrollaron su segundo tumor antes que el grupo sin tratamiento de radiación y/o sin estas variantes de *ATM* (Stucci LS. *et al.*, 2021).

Todo lo anterior muestra una asociación de *ATM* con un cáncer de mama más agresivo, en la que la ausencia de la expresión de proteína *ATM* está fuertemente relacionada con la existencia de metástasis distantes (Bernstein *et al.*, 2010). Además, los efectos colaterales del tratamiento con radioactividad serían mayores en casos con variantes missense de *ATM* que reducen el nivel de actividad de la proteína, aumentando así la susceptibilidad de inducción de tumores por radiación (Bernstein *et al.*, 2010).

Debido a características específicas como la susceptibilidad a la radiación que se presentan en las pacientes portadoras de variantes patogénicas de *ATM*, se ha sugerido una guía especial para su tratamiento que resumimos en la Figura 4. En dicha guía se describen las singularidades de la historia clínica que se comparten entre las pacientes, su pronóstico y los tratamientos en los cuales hay una respuesta favorable en estas pacientes (Bernstein *et al.*, 2010).



**Figura 1.4.** Diagrama de flujo de las características clínicas de pacientes con cáncer de mama que presentan variantes *ATM*. Tomado de [.https://doi.org/10.3390/genes12050727](https://doi.org/10.3390/genes12050727)

### 1.1.2 Otros tipos de Cáncer asociados a *ATM*

La razón por la cual variantes patogénicas de *ATM* generan una sensibilidad a la radiación ionizante, es la función celular que realiza este gen. La proteína *ATM* cumple una función como regulador que detecta y señala la respuesta celular a daños del ADN por roturas de la doble cadena (Bernstein *et al.*, 2010). Consistente con esta función oncosupresora, se ha visto que las variantes patogénicas en *ATM* también estarían asociadas a un aumento en los riesgos de desarrollar



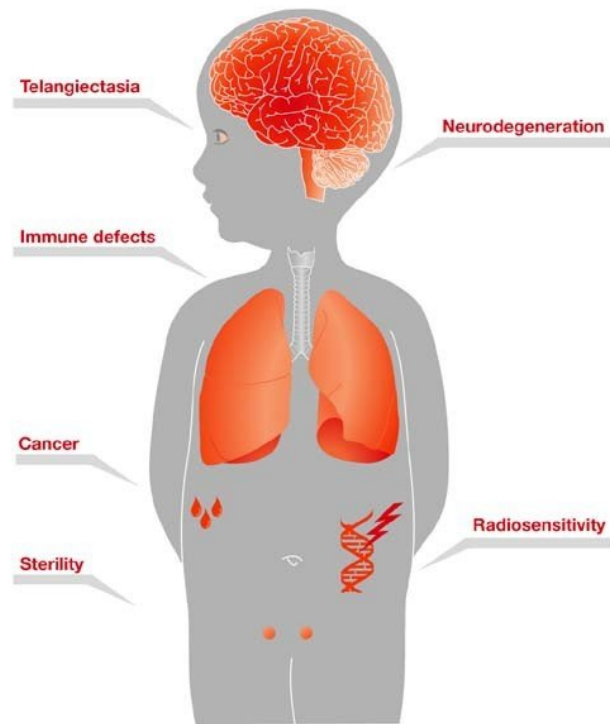
otros tipos de cánceres en los adultos, además del cáncer de mama.

De manera específica se ha establecido una relación entre la presencia de ciertas variantes de *ATM* y la alta incidencia de cánceres linfoides (Hall MJ. *et al.*, 2021), como el linfoma de células del manto y la leucemia prolinfocítica T, cáncer gástrico, cáncer de páncreas, cáncer gastroesofágico, cáncer colorrectal, cáncer de ovario, melanoma, cáncer de próstata, cáncer de tiroides cáncer, y cánceres de cabeza y cuello. Es importante recalcar que la estimación de los diversos riesgos de cáncer asociados a las variantes de *ATM* se ha visto limitada principalmente por los pequeños tamaños muestrales de participantes en estos estudios. Por ello, las estimaciones reportadas con los diferentes tipos de cáncer pueden variar entre los diferentes trabajos (Hall MJ. *et al.*, 2021).

### **1.1.3 Ataxia-telangiectasia (ATM): gen y proteína**

Un aspecto importante de cara a utilizar las variantes genéticas de *ATM* en procesos clínicos de diagnóstico y tratamiento es la comprensión de su impacto. Sin embargo, esto no es nada fácil, ya que la relación entre genotipo y fenotipo para este gen no es nada sencilla. Consideremos el caso de la Ataxia Telangiectasia AT, una enfermedad causada por variantes patogénicas de *ATM*. Debido a que el gen *ATM* está asociado con un patrón de herencia autosómico

recesivo, solo las personas con dos copias de variantes patogénicas en el gen *ATM* que muestran niveles bajos o indetectables de proteína ATM, presentan la enfermedad. Esta corresponde a un complejo trastorno neurodegenerativo con síntomas como sensibilidad a la radiación ionizante, e inmunodeficiencia, entre otras (Fig.1.5). Así como condiciones de envejecimiento prematuro y disgenesia gonadal (Stucci LS. *et al.*, 2021).



**Figura 1.5. Diagrama resumido de los Síntomas de AT.**

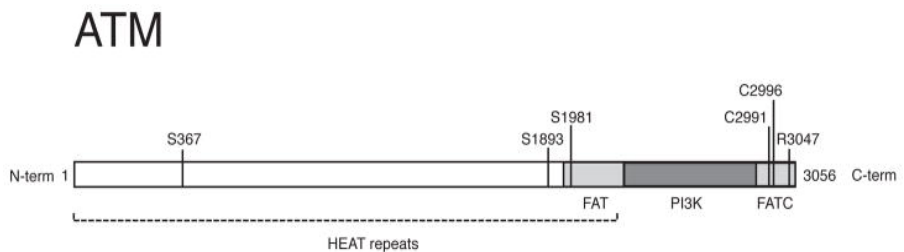
Tomado de:  
<https://globalgenes.org/2013/03/14/9-year-old-with-ataxia-telangiectasia-needs-our-help-now/>

Siendo un síndrome tan complejo, todavía no se comprende en su totalidad las razones por las cuales se presenta este fenotipo tan destructor. Sin embargo, también es cierto que nuestro conocimiento cada vez mayor de la función molecular de la proteína ATM empieza a desvelar esta compleja relación genotipo-fenotipo. Para facilitar su comprensión al lector, en lo que sigue recogemos algunos aspectos fundamentales de la función molecular de *ATM*.

El gen *ATM* para los humanos se encuentra en el cromosoma 11q23, presenta 69 exones, con 65 de ellos codificantes para una proteína de 3056 aminoácidos, con un tamaño aproximado de 370 kD. La proteína ATM es miembro de la familia de proteínas quinasas relacionadas con la fosfatidilinositol 3-quinasa (PIKK) que se encuentran involucradas en las respuestas al daño del ADN (Bernstein JL. *et al.*, 2017).

Esta función se distribuye, a lo largo de la secuencia, en un número limitado de dominios conservados. Si vamos del extremo N-terminal al C-terminal de la proteína, el primer dominio que encontramos es el dominio TAN, involucrado en el mantenimiento de la longitud de los telómeros. Este dominio es seguido por una variedad de motivos HEAT (Huntingtin, factor de elongación 1A, proteína fosfatasa 2A subunidad A, TOR) repetidos que pueden facilitar la formación de grandes complejos de proteínas. Finalmente, acercándonos al C-terminal encontramos varios dominios importantes para la actividad de la proteína: del aminoácido ~1960 al 2566 encontramos el dominio FAT, que limita con el dominio de serina/treonina quinasa C-terminal, que abarca desde el aminoácido ~2712 al 2962 y media

gran parte de la función de señalización de ATM. Por último, hallamos el dominio FATC ubicado desde el aminoácido ~2963 al 3056. En la Figura 6 se encuentra representada la distribución de los dominios y la ubicación de los sitios de autofosforilación en ATM (S367, S1893, S1981 y S2996) (Bernstein *et al.*, 2017).

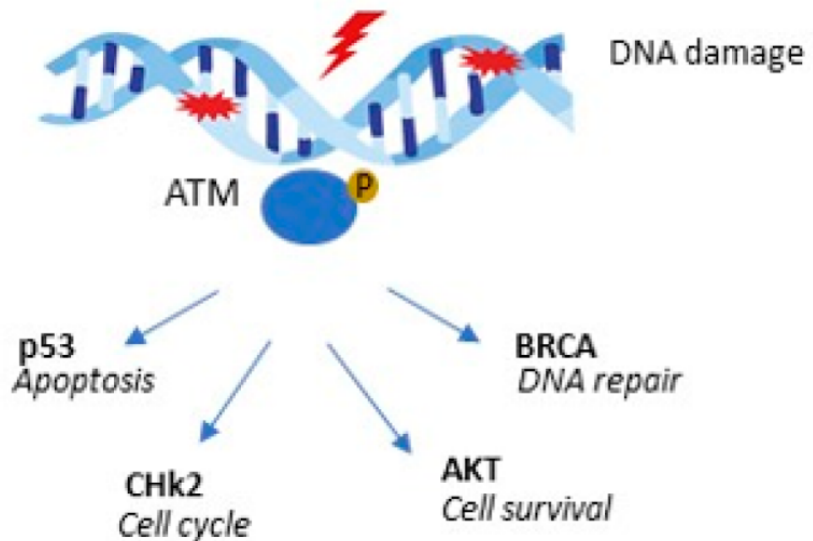


**Figura 1.6. Estructura de dominios de ATM, y lugares de autofosforilación.** Tomado de DOI:10.1016/j.tibs.2011.10.002

Cuando la célula se encuentra en buen estado, la proteína ATM se mantiene inactiva, formando dímeros u oligómeros en los que el dominio quinasa está unido a una región que rodea a la serina ubicada en el aminoácido 1981 y que está contenida dentro del dominio FAT. Cuando ocurren daños en el ADN, cambios en la estructura de la cromatina o daño celular, se induce la autofosforilación intermolecular de la serina 1981, lo que a su vez provoca la disociación del dímero dando lugar a los monómeros activos de esta proteína que inician su actividad la quinasa celular

(Bakkenist *et al.*, 2003).

En su forma activa, la proteína ATM fosforila varios de sus blancos, como *P53*, *CHEK2* y *BRCA1*, todos involucrados en procesos celulares vitales como el mantenimiento del ciclo celular, la apoptosis, el estrés oxidativo y el mantenimiento de los telómeros. *ATM* media así decisiones celulares importantes como son la activación de la maquinaria de control del ciclo celular, la reparación del ADN o el inicio de la apoptosis, por tal razón no es extraño que las mutaciones en *ATM* se relacionen con tantos tipos de cáncer (Figura 1.7) (Bernstein *et al.*, 2017).



**Figura 1.7. Representación de la actividad de ATM.** Una vez activada *ATM* sirve como transductor, fosforila y activa otras proteínas quinasas lo que resulta en la detención del ciclo celular, supervivencia celular, reparación del ADN o Apoptosis. Adaptado de: doi: 10.3390/genes12050727

Como vemos, la cantidad de información sobre la función de *ATM* nos permitirá comprender, de forma paulatina y cuantitativa, el mecanismo mediante el cual las variantes de *ATM* pueden contribuir a la aparición de los cánceres asociados a este gen. Sin embargo, hasta que llegemos a este punto, si queremos utilizar esta información en un entorno clínico, necesitamos modelos que nos respondan, aunque sea de forma aproximada, a la pregunta de si una variante de *ATM* es patogénica o benigna. Los métodos *in silico* o bioinformáticos de predicción de patogenicidad nos proporcionan esa respuesta, mediante el uso de métodos de machine learning (aprendizaje automático). En las próximas secciones se presentan estos métodos, sus características principales, su capacidad predictiva, y su marco de uso.

## **1.2 Uso de la información de las variantes en el entorno clínico**

En la sección anterior hemos visto que la posibilidad de que una paciente sea atendida con un tratamiento asertivo para cáncer de mama, aumenta gracias a la comprensión de la naturaleza de las variantes en los genes asociados al desarrollo de esta enfermedad. Por consiguiente, es importante poder clasificar las variantes de

acuerdo con el riesgo de padecer cáncer que confieren a su portadora.

Esta necesidad se ha visto incrementada por los avances en tecnologías genéticas como la secuenciación de nueva generación (NGS), cuya capacidad de detectar variantes en cantidades masivas, supera las capacidades de la investigación primaria y específica para cada una de estas. Siendo así, para muchas de las variantes halladas no se ha podido establecer todavía su naturaleza, son las que conocemos como VUS (variantes de significado desconocido) (Federici G. *et al.*, 2020).

### **1.2.1 Guías ACMG/AMP Richards. *et al.* 2015**

Ante la situación descrita anteriormente, el Colegio Americano de Genética Médica y Genómica (ACMG), en colaboración con los expertos de las asociaciones patológicas CAP (Colegio de Patólogos Estadounidenses) y moleculares, la AMP (Asociación de Patología Molecular) de Estados Unidos, publicó en el año 2015 un artículo donde se establecen las guías ACMG/AMP estandarizadas para la clasificación de las variantes genéticas (Richards S. *et al.*, 2015).

En estas guías de clasificación para variantes de enfermedades mendelianas, se categorizan las variantes en las siguientes cinco

clases: "patogénica", "probablemente patogénica", "clasificación incierta", "probablemente benigna" y "benigna". Para asignar una variante concreta a una de estas categorías, se describe un protocolo basado en la combinación de las siguientes fuentes de información: datos de población, datos computacionales, datos funcionales, datos de segregación, data de novo y otros datos (Figura 1.8 -1.9).

Destacaremos, entre los aspectos relacionados con el impacto molecular de las variantes, que los autores recomiendan bases de datos donde se puede encontrar información sobre la patogenicidad o benignidad de las variantes, aunque hacen la salvedad que muchas de ellas pueden estar clasificadas de una manera errónea. Asocian el origen de estos errores principalmente con los estudios primarios, ya que la información puede tener fallos en el momento de ser recolectada y/o durante el análisis de los datos. Por ello sugieren la aplicación de ciertos estándares antes de aceptar la información como verídica. Así mismo, para completar esta información, nombran algunos predictores *in silico* como fuentes de evidencia computacional, aunque sin descartar el empleo de otros predictores, ni limitar el número de predictores a usar.

Cada criterio propuesto en estas guías contribuye a la clasificación de las variantes, según el valor que se le otorga a la evidencia que aporta y al origen de la misma. Estos criterios se combinan siguiendo unas



reglas concretas, dando lugar a una clasificación de la variante en una de las cinco categorías mencionadas anteriormente.

En estas guías también se sugiere cómo usar esta información en el entorno clínico, ya que recomiendan las decisiones que se pueden tomar luego de la clasificación de una variante. Aunque siempre indican que se deben hacer esfuerzos para evitar usar esta clasificación como única evidencia de la existencia de enfermedad, más bien como parte de información de apoyo a la evidencia clínica.

<b>Population Data</b>	MAF is too high for disorder <i>BA1/BS1</i> OR observation in controls inconsistent with disease penetrance <i>BS2</i>	
<b>Computational And Predictive Data</b>		Multiple lines of computational evidence suggest no impact on gene /gene product <i>BP4</i> Missense in gene where only truncating cause disease <i>BP1</i> Silent variant with non predicted splice impact <i>BP7</i>
<b>Functional Data</b>	Well-established functional studies show no deleterious effect <i>BS3</i>	
<b>Segregation Data</b>	Non-segregation with disease <i>BS4</i>	
<b>De novo Data</b>		
<b>Allelic Data</b>		Observed in <i>trans</i> with a dominant variant <i>BP2</i> Observed in <i>cis</i> with a pathogenic variant <i>BP2</i>
<b>Other Database</b>		Reputable source w/out shared data = benign <i>BP6</i>
<b>Other Data</b>		Found in case with an alternate cause <i>BP5</i>

Figura 1.8. Criterios y sistema de clasificación de variantes benignas, según las guías ACMG/AMP.

Pathogenic →			
Supporting	Moderate	Strong	Very Strong
	Absent in population databases <i>PM2</i>	Prevalence in affecteds statistically increased over controls <i>PS4</i>	
Multiple lines of computational evidence support a deleterious effect on the gene /gene product <i>PP3</i>	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before <i>PM5</i>  Protein length changing variant <i>PM4</i>	Same amino acid change as an established pathogenic variant <i>PS1</i>	Predicted null variant in a gene where LOF is a known mechanism of disease <i>PVS1</i>
Missense in gene with low rate of benign missense variants and path. missenses common <i>PP2</i>	Mutational hot spot or well-studied functional domain without benign variation <i>PM1</i>	Well-established functional studies show a deleterious effect <i>PS3</i>	
Co-segregation with disease in multiple affected family members <i>PP1</i>	Increased segregation data →		
	<i>De novo</i> (without paternity & maternity confirmed) <i>PM6</i>	<i>De novo</i> (paternity & maternity confirmed) <i>PS2</i>	
	For recessive disorders, detected in <i>trans</i> with a pathogenic variant <i>PM3</i>		
Reputable source = pathogenic <i>PP5</i>			
Patient's phenotype or FH highly specific for gene <i>PP4</i>			

**Figura 1.9. Criterios y sistema de clasificación de variantes patogénicas, según las guías ACMG/AMP.** Abreviaturas: LOF, pérdida de función; PM, patogenicidad moderada; PP, apoya patogenicidad; PS, patogenicidad fuerte; PVS, patogenicidad muy fuerte.

## 1.2.2 Adaptación de las guías generales a genes concretos

A pesar que las guías ACMG/AMP fueron creadas con el objetivo de ser aplicadas a las variantes de los diferentes genes, dadas las peculiaridades de cada gen, ha sido necesario adaptarlas a enfermedades y genes específicos.

Las guías se han ajustado según los grupos de expertos en el tema, adaptándose a genes como: *CDH1*, *TP53*, *MYH7*, *HNFI1A*, *ATM* entre otros. En la página de paneles de Expertos de ClinGen se evidencian 53 entradas de adaptación de las guías ACMG/AMP, algunas de las cuales están todavía en proceso de construcción (ClinGen <https://clinicalgenome.org/affiliation/all>).

Para la adaptación se modifica principalmente la forma de aceptación de una variante en las categorías, siendo más restrictiva o más laxa según la información disponible sobre el gen, y las cualidades de herencia del mismo. También se especifica de forma precisa el predictor *in silico* a usar para la anotación de las variantes en cada enfermedad (ClinGen <https://clinicalgenome.org/affiliation/all>).

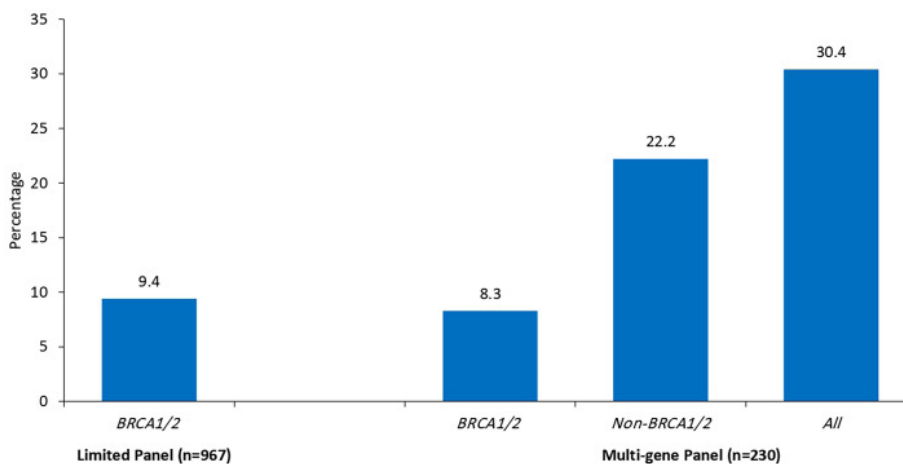
Para el caso particular de *ATM* fue necesario hacer ajustes a los criterios usados en las guías ACMG/AMP, ya que la penetrancia de este gen es moderada/baja, mientras las guías están pensadas para clasificar variantes de genes con penetrancia completa, como en los trastornos mendelianos clásicos (Feliubadaló L. *et al.*, 2021).

A la fecha se han llevado a cabo dos adaptaciones para *ATM*. La más reciente ha sido presentada por el comité de expertos de ClinGen en enero del 2022. En ella se presentan las pautas de interpretación para *ATM* según expertos en cáncer hereditario de mama, ovario y páncreas. Entre sus ajustes se encuentra el uso de REVEL como el predictor *in silico* recomendado para la clasificación de las variantes. Se redefine, además, la forma de uso de sus predicciones, estableciendo criterios más restrictivos que los usados normalmente con REVEL para la calificación de las variantes (ClinGen, Guías *ATM*., 2022).

Un poco antes, Feliubadaló, *et al.* en el 2021, publicaron una adaptación de las guías aplicadas a las variantes de *ATM* en el contexto de cáncer de mama. En esta adaptación se ajustaron diferentes criterios. Por ejemplo, la frecuencia de variantes poblacionales. Este cambio mejoró la clasificación de algunas variantes que pasaron de ser consideradas como VUS, según los criterios de las guías originales ACMG/AMP, a ser clasificadas como 'probablemente benignas', según los nuevos umbrales adoptados. En esta adaptación también se incluyó una comparación entre diferentes herramientas *in silico* para la predicción de las variantes missense en *ATM*, eligiendo la combinación de predictores que daba el mejor rendimiento. Este aspecto forma parte del trabajo de esta tesis y se desarrollará en el capítulo.

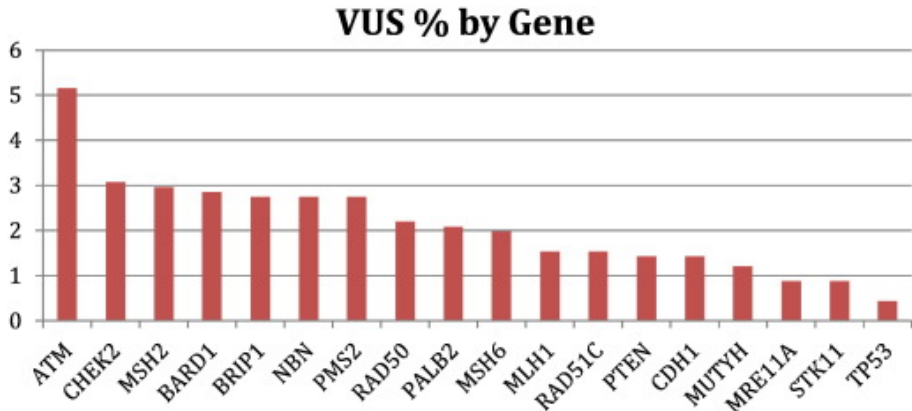
### 1.3 Clasificación e interpretación *in silico* de variantes

Las variantes de significado incierto o VUS (Variants of Uncertain Significance), representan alrededor del 40% del total de variantes registradas en las bases de datos genéticos (Federici G. *et al.*,2020). En el contexto clínico, esta situación genera incertidumbre y problemas para seleccionar el tratamiento adecuado de la mayoría de las pacientes (Abdel-Razeq H. *et al.*,2022). Esto no sería tan grave si el número de estas variantes fuese más reducido. Pero, lamentablemente, se ha visto que la mayoría de las variantes encontradas en pacientes con cáncer de mama son VUS (Figura 1.10) (Smith. *et al.*). Por ejemplo, en la investigación de Abdel-Razeq. *et al.* (Abdel-Razeq H. *et al.*, 2022) se observa, en una cohorte de 1197 mujeres de origen árabe, que aproximadamente un 30% de las variantes halladas estaban clasificadas como VUS y provenían principalmente de los genes *PALB2*, *CHEK2*, *ATM*, *BARD1*, *NBN* y *NF1*.



**Figura 1.10. Porcentaje de variantes VUS por gen encontrado en panel multigénico para Cáncer de mama.** Tomado de: doi:10.3389/fonc.2022.673094

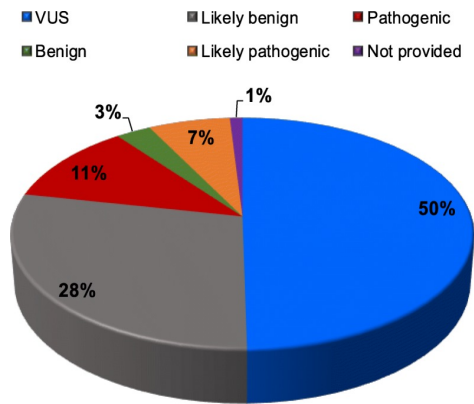
Hay que señalar que no todos los genes presentan el mismo número de VUS. Por ejemplo, Minion. *et al.* & Smith. *et al.* (Minion. *et al.* & Smith. *et al.* 2021), en sus trabajos sobre las variantes VUS obtenidas en un panel de cáncer hereditario de 19 genes asociados a riesgo de cáncer de mama y ovario hereditario, hallan que el gen que mayor número de VUS presenta es *ATM*, recogiendo un 5 % de todas las variantes encontradas (Figura 1.11).



**Figura 1.11. Porcentaje de VUS halladas por gen, en análisis multigénico para cáncer de mama.** Tomado de: doi: 10.1016

Cuando se analiza la información sobre las variantes clasificadas en ClinVar, se encuentra que para el 2020 al menos el 50 % de las variantes de *ATM*, lo que equivale a 2000 variantes, son VUS, intensificando el problema que se presenta con *ATM* en el entorno diagnóstico (Figura 1.12) (Federici G. *et al.*, 2020).





**Figura 1.12.** Gráfico circular que representa los porcentajes de variantes genéticas de *ATM* agrupadas en clases clínicas según ClinVar. Tomada del: <https://doi.org/10.1016/j.ygyno.2015.01.537>

### 1.3.1 Construcción de predictores

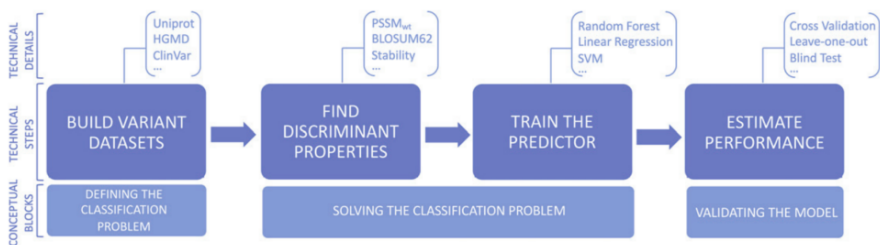
De la necesidad de clasificar la gran cantidad de variantes VUS que se encuentran, han surgido soluciones *in silico* como son los predictores de patogenicidad. Esta alternativa, cuestionada inicialmente, está siendo paulatinamente aceptada por la comunidad biomédica/clínica porque tal como lo hemos visto en la sección anterior, las guías ACMG/AMP contemplan en uno de sus criterios la predicción *in silico* de la variante.

Los predictores de patogenicidad se basan en algoritmos de aprendizaje automático o machine learning que tienen la capacidad de aprender sobre un problema determinado a partir de un conjunto

de ejemplos proporcionados por el usuario. (Özkan *et al.* 2021)

En el caso particular de las variantes missense es necesario que el predictor aprenda de un conjunto de variantes de naturaleza patogénica/benigna conocida. Un aspecto que hay que destacar es la importancia del número de ejemplos utilizados y su calidad. De hecho, cuanto mayor sea esta mejor aprenderá el algoritmo a discernir entre los dos tipos de variantes (benignas/patogénicas) y, por tanto, será más exacto.

Existen cuatro pasos estándar para la construcción de un predictor *in silico*, que resumimos en la Figura 1.13 (Özkan *et al.*, 2021) y explicamos a continuación.



**Figura 1.13. Protocolo para la construcción de predictores de patogenicidad.** Cada una de las cuatro cajas de azul intenso representa uno de los grandes pasos en la construcción de un predictor. Encima de cada caja se muestran los elementos específicos para el problema de predicción de patogenicidad. Tomado de (Özkan S. *et al.*, 2021)

### 1.3.1.1 Paso 1: Conjunto de datos

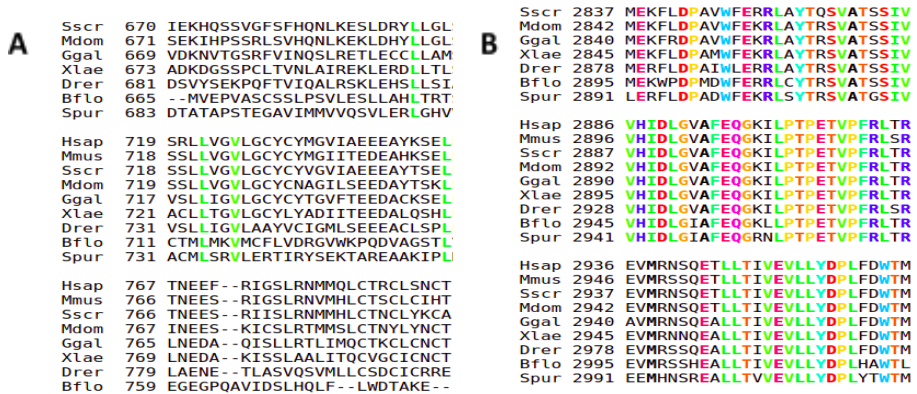
Para crear el conjunto de un predictor específico de proteína, e.g., de *ATM*, el primer paso es obtener un conjunto de datos de entrenamiento, para la proteína objeto de estudio. Si estamos diseñando un predictor general (válido para cualquier proteína) es necesario tener variantes de diferentes proteínas. En este trabajo nos centraremos en el caso de las variantes missense, aunque puede extenderse a otras de variantes.

Las variantes patogénicas pueden ser obtenidas de bases de datos como UniProt/SwissProt (Bateman. *et al.*, 2017a), HGMD (Stenson. *et al.*, 2012) o ClinVar (Landrum. *et al.*, 2016), bases de datos que son recomendadas por su actualización constante. Aunque para tener una mayor certeza de la clasificación como patogénica es recomendable revisar el artículo primario y su metodología.

Para las variantes benignas hay menos trabajos disponibles en la literatura. Por ello, diferentes grupos han desarrollado una alternativa a las búsquedas bibliográficas (Ferrer-Costa, Orozco y De La Cruz, 2004; Sunyaev. *et al.*, 2001), que es el uso de alineamientos múltiples de proteínas (MSA).

Este modelo toma la secuencia humana como referencia y utiliza las desviaciones de secuencia observadas en otras especies (con identidad de secuencia >95% respecto a la proteína humana) como variantes benignas (Figura 1.14). Utilizado habitualmente en el

desarrollo de predictores de patogenicidad (Adzhubei. *et al.* 2010; Riera. *et al.* 2014) este modelo ha mostrado resultados competitivos en la construcción de predictores tanto generales como específicos (Riera. *et al.* 2016; Padilla. *et al.* 2019).



**Figura 1.14. Ejemplo de MSA.** En el alineamiento se resaltan mediante diferentes colores aquellos aminoácidos que se conservan entre las especies. A. Zona poco conservada B. Zona altamente conservada. Por ejemplo la variante R720S, es benigna.

### 1.3.1.2 Paso 2: Las propiedades discriminantes

Las propiedades discriminantes son aquellas propiedades de las variantes que permiten relacionarla con su fenotipo clínico (Özkan *et al.*, 2021). Estas características, que pueden ser derivadas de la biología molecular, la biofísica, y la bioquímica, serán la información con la cual el algoritmo aprende a discriminar entre variantes

patogénicas y benignas.

Características biofísicas. Son medidas, en términos de energía libre, del efecto de las variantes sobre la estabilidad y capacidad de interacción de las proteínas.

Características moleculares. Estas propiedades son locales, a diferencia de las anteriores, que son globales. Hacen referencia a cambios en hidrofobicidad, volumen, y carga, entre otros, producto del cambio de un aminoácido. Por lo general, los valores grandes indican un impacto molecular importante, mientras que los valores pequeños indican cambios mejor tolerados por la proteína. Estas propiedades se resumen en matrices de sustitución como Blosum62 (Henikoff & Henikoff, 1992), cuyos valores corresponden a cambios molecularmente 'agresivos' cuando son negativos y más neutros cuando son positivos.

Características basadas en la conservación de secuencias. Estas propiedades miden el grado de conservación de las secuencias en el alineamiento múltiple de nuestra proteína y están relacionadas con la relevancia estructural/funcional de los aminoácidos. Aunque hay diferentes opciones, las más utilizadas son la entropía de Shannon y el PSSM (Position Specific Score Matrix). La entropía de Shannon (Cover & Thomas, 2006) refleja la variabilidad que tiene un aminoácido en una columna específica del MSA. Es igual a  $-\sum_i p_i \cdot \log_2(p_i)$ , donde  $p_i$  es la proporción del aminoácido de tipo  $i$  (e.g., alanina, valina, etc.) en la columna del MSA donde ocurre la variante

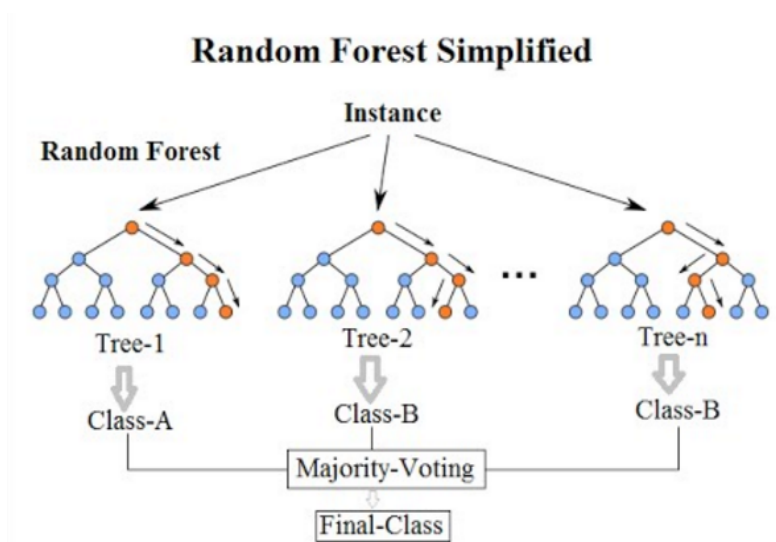
y abarca todos los aminoácidos naturales. Valores bajos de entropía son característicos de aminoácidos altamente conservados entre especies y sugieren poca tolerancia al cambio. Por el contrario, valores altos de entropía indican una mayor tolerancia al cambio. El PSSM mide la frecuencia del aminoácido nativo en la ubicación variante, normalizada por la frecuencia de dicho aminoácido en todo el MSA. Es igual a  $\log_2(f_{\text{nati}}/f_{\text{nATMSA}})$ , donde  $f_{\text{nati}}$  es la frecuencia del aminoácido nativo en el locus de la variante y  $f_{\text{MSA}}$  es la frecuencia del mismo aminoácido en todo el MSA. Valores altos del PSSM indican cambios disruptivos, mientras que los valores bajos indican cambios más tolerables.

Características en los Metapredictores. Los metapredictores son un tipo de predictores de patogenicidad de segunda generación que utilizan, como propiedades predictivas, los resultados de otros predictores. Por ejemplo, un metapredictor puede utilizar como input los resultados sobre una variante de predictores primarios como SIFT, CADD, y Align-GVGD. En general, se observa que los metapredictores mejoran la capacidad predictiva respecto a los predictores primarios, por lo cual se han convertido en una herramienta muy útil.

### **1.3.1.3 Paso 3: Entrenamiento de un algoritmo de machine learning**

Conceptualmente, el entrenamiento de un predictor es un paso muy sencillo. Consiste en escoger uno de los algoritmos de machine learning disponibles, y presentarle un conjunto de ejemplos de variantes patogénicas y benignas, etiquetadas con las características discriminantes elegidas, para que aprenda a distinguir las. A nivel técnico, este proceso es matemáticamente sofisticado (Bishop, 2006) y su descripción queda fuera de los límites de esta tesis.

Para la selección del tipo de algoritmo se debe considerar que su capacidad predictiva depende del problema considerado. Por ejemplo, las Artificial Neural Networks (ANN) son especialmente recomendadas en el análisis de imágenes (Currie *et al.*, 2019). Ello es debido a que, matemáticamente, cada algoritmo está estructurado de una manera diferente. Por ejemplo, Random Forest es un algoritmo basado en la combinación aleatoria de árboles de decisión (Figura 1.15), completamente diferente de la estructura de las ANN. Inyectar el tipo correcto de aleatoriedad los convierte en clasificadores precisos, con una importante capacidad para generar predicciones no sesgadas (Breiman, 2001).



**Figura 1.15. Representación de la aleatoriedad en los árboles de decisiones usados en Random Forest.** Los círculos azules representan todas las posibles ramas en las que un dato inicial puede ser incluido. En rojo, como ejemplo, se muestra un caso en particular como es distribuido en las diferentes ramas. La clasificación de este como A/B por cada rama será integrada por el sistema para concluir sobre la naturaleza del ejemplo. Tomado de: [https://en.wikipedia.org/wiki/Random\\_forest#/media/File:Random\\_forest\\_diagram\\_complete.png](https://en.wikipedia.org/wiki/Random_forest#/media/File:Random_forest_diagram_complete.png)

De forma general es recomendable hacer pruebas con diferentes algoritmos, para tener un panorama de los resultados, recomendándole además elegir el algoritmo con mayor interpretabilidad y simplicidad, para limitar los problemas de sobreajuste (Rudin, 2019).

Una vez escogido el algoritmo, se procede a su entrenamiento,



siguiendo los pasos recomendados en el manual asociado. El predictor resultante ya se podrá aplicar a cualquier variante desconocida y generará valor numérico que, mediante un umbral de corte, se transformará en una predicción binaria.

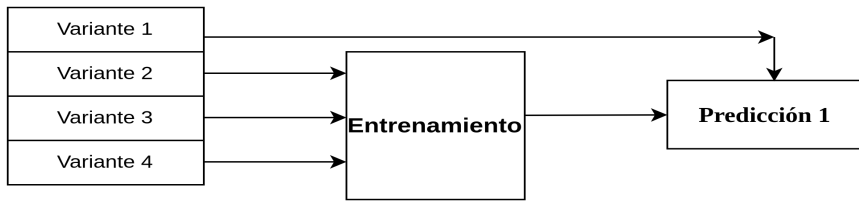
#### **1.3.1.4 Paso 4: Estimación y validación del rendimiento del predictor**

En este paso se estima la capacidad de acierto, o desempeño, de nuestro predictor en casos desconocidos; también se conoce como la capacidad de generalización. Para realizar dicha estimación se pueden usar diferentes estrategias donde siempre el objetivo es someter al algoritmo entrenado a nuevos casos que no hayan sido parte de su entrenamiento.

Las estrategias para estimar la capacidad de acierto de un predictor dependen en gran medida de la cantidad de datos disponibles. Por ejemplo, si tenemos un conjunto de datos lo suficientemente grande podemos separarlo en dos grupos de tamaño diferente: el mayor para entrenar al algoritmo y el segundo para validarlo.

Sin embargo, hay casos donde el conjunto inicial de datos es pequeño. En este caso existen otras estrategias para realizar su análisis de desempeño, como es la validación cruzada ('Leave-one-out cross-validation', LOOCV). Esta estrategia se ilustra en la Figura 1.15.

En ella, del total de variantes que se tiene inicialmente, se saca una de ellas. El entrenamiento se realiza con el conjunto resultante. La variante que ha quedado afuera pasa a ser predicha luego del entrenamiento. El proceso anterior se itera tantas veces como variantes se tengan, de manera que al final se obtiene la predicción de todas las variantes.



Iteraciones	Variante 1	Variante 2	Variante 3	Variante 3
Predicción 1	Prueba	Entrenamiento	Entrenamiento	Entrenamiento
Predicción 2	Entrenamiento	Prueba	Entrenamiento	Entrenamiento
Predicción 3	Entrenamiento	Entrenamiento	Prueba	Entrenamiento
Predicción 4	Entrenamiento	Entrenamiento	Entrenamiento	Prueba

**Figura 1.15. Diseño de validación cruzada.** En el esquema superior se muestra como una variante sale del conjunto de datos de entrenamiento y pasa hasta el punto de la predicción (transformándose en una variante de Prueba). En la parte inferior, en rojo están coloreadas las variantes que van saliendo en cada repetición para pasar a variante de prueba.

La estimación final de la capacidad predictiva del algoritmo se obtiene a partir de las predicciones generadas con las variantes de prueba. Hay diferentes métricas que permiten medir el rendimiento de nuestro predictor, todas ellas basadas en la comparación de las predicciones con las observaciones (Vihinen, 2012). De forma general, todas las medidas de rendimiento se basan en la matriz de confusión (Figura 1.16). En esta matriz, el número de variantes bien predichas quedará consignado en los cuadros de los 'TRUE POSITIVES' (TP) y 'TRUE NEGATIVES' (TN). Mientras que el número de errores que ha cometido el predictor aparecerá en las casillas de 'FALSE POSITIVES' (FP) y 'FALSE NEGATIVES' (FN).

		Observados	
		Patogénico	Benigno
Predichos	Patogénico	TP	FN
	Benigno	FP	TN

Figura 1.16. Matriz de confusión

A partir de esta matriz se obtienen diferentes métricas. Las más utilizadas habitualmente, y que vamos a tener en cuenta en esta tesis, son:

--Sensibilidad: mide la proporción de casos positivos bien predichos respecto al total de casos positivos disponibles. Se expresa como:

$$\text{Sensibilidad} = \frac{TP}{TP+FN}$$

--Especificidad: mide la proporción de casos negativos bien predichos con respecto al total de casos negativos disponibles. Se expresa como:

$$\text{Especificidad} = \frac{TN}{TN+FP}$$

El análisis combinado de estas dos características ayuda a tener una idea sobre el posible sesgo que presenta un predictor en sus resultados (Ernst. *et al.*, 2018)

--Positive predictive value (PPV): mide la proporción de casos positivos bien predichos con respecto al total de los casos positivos predichos por el algoritmo. Se expresa como:

$$PPV = \frac{TP}{TP+FP}$$

--Negative predictive value (NPV): mide la proporción de casos

negativos bien predichos con respecto al total de los casos negativos predichos por el algoritmo. Se expresa como:

$$NPV = \frac{TN}{TN+FN}$$

--Accuracy: se refiere a la fracción total de predicciones acertadas con respecto al total de casos analizados. Da una visión general de la capacidad de acierto del predictor, sin embargo, cuando una clase es más frecuente que la otra, la accuracy puede ser engañosa (Baldi *et al.*, 2000). Se expresa como:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

--Matthews Correlation Coefficient (MCC): Es un parámetro comprendido entre -1 y 1, donde 1 refleja una concordancia total entre los datos observados y los predichos, mientras -1 es un desacuerdo total; 0 corresponde a un predictor aleatorio. MCC se considera más informativo que las medidas anteriores (Chicco, 2017). Se expresa como:

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP+FP).(TP+FN).(TN+FP).(TN+FN)}}$$

### **1.3.2 Necesidad de interpretación en predictores *in silico***

Como hemos visto en las secciones anteriores, los predictores de patogenicidad se basan en el uso de algoritmos de machine learning o, lo que es lo mismo, algoritmos de inteligencia artificial (IA). Y por ello, aunque recogen todas las virtudes de esta rama del saber, también presentan sus problemas. Entre ellos está el de la baja interpretabilidad, un problema asociado a las técnicas más efectivas de IA y que describimos a continuación.

Un aspecto destacado de la inteligencia artificial es que los avances recientes en este campo a menudo se han logrado aumentando la complejidad de los modelos predictivos, convirtiéndolos en lo que se conoce como sistemas "caja negra".

Esta complejidad ha generado una preocupación por la falta de interpretabilidad en los modelos predictivos que puede generar desconfianza en ellos, especialmente en el área de la salud donde se aplican en situaciones clínicas cuyo desenlace afecta a la vida y salud del paciente (Petch *et al.*, 2022). De hecho, cada vez está más aceptada la idea de que no deben de tomarse decisiones sobre situaciones críticas, como la predicción de una variante para el tratamiento contra el cáncer de mama, utilizando modelos de caja negra. Estos pueden demostrar una alta capacidad predictiva, pero no proporcionan ningún tipo de explicación para su predicción, lo que impide la identificación de posibles errores mediante análisis manual.

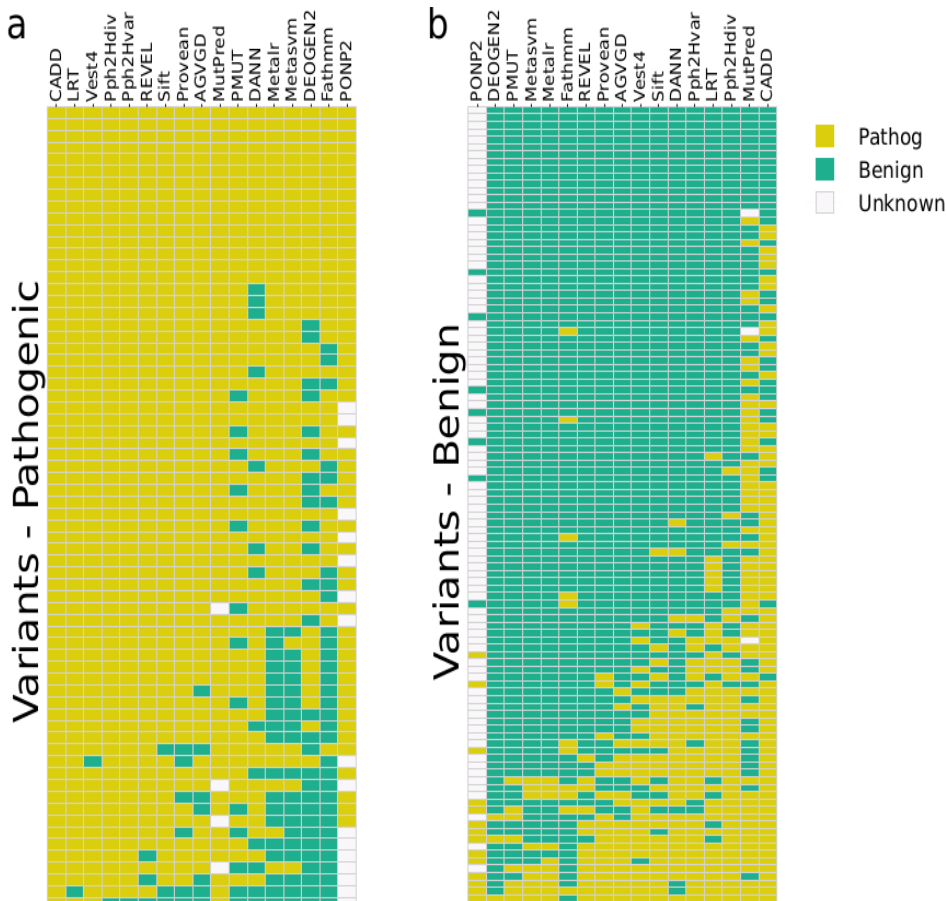
Ello es totalmente insuficiente e incluso éticamente inaceptable (AI-HLEG-group, 2019). La explicación es esencial para comprender y, por lo tanto, confiar en estos modelos y respaldar sus predicciones (Pintelas E. *et al.*, 2020).

Esta necesidad, tanto en el campo de la medicina como en otras áreas, ha dado lugar en los últimos años a una nueva área de la inteligencia artificial en la que se busca el desarrollo de nuevos métodos para acercar los modelos de aprendizaje automático al usuario, facilitando la interpretación de los resultados más complejos (Pintelas E. *et al.*, 2020).

Esta tendencia ha sido una de las motivaciones que se hallan detrás de esta tesis. En *ATM*, nuestro campo clínico de interés, los predictores desarrollados hasta ahora no son perfectos y sus rendimientos oscilan entre un 60-80% de acierto (Riera. *et al.*, 2016). Esta limitación se ilustra en la Figura 1.17, en la que se observa cómo algunos clasificadores son desbalanceados y por tanto predicen mejor solo una de las clases de variantes, ya sean las patogénicas o las benignas. Aún más importante, se debe resaltar el hecho que ninguna de las herramientas analizadas predice perfectamente todas las variantes mostradas, generando así una mayor sospecha de error al momento de usar estos predictores. Por lo tanto, resulta acuciante la necesidad de disponer de medios de análisis independientes que permitan al usuario clínico juzgar la fiabilidad de las predicciones *in silico*.

En línea con lo expuesto anteriormente, en el trabajo de esta tesis se abordan dos aspectos principales en la anotación *in silico* de las variantes de *ATM*: (i) la obtención de predictores competitivos para estas variantes y (ii) el desarrollo de herramientas que faciliten la interpretación de las predicciones computacionales. Estos aspectos conforman los objetivos de este trabajo que se enuncian con mayor detalle a continuación.





**Figura 1.17. Resultados de diecisiete predictores de patogenicidad aplicados a las variantes de ATM en nuestro conjunto de datos.** Cada columna corresponde a un predictor y cada fila a una variante. Las celdas individuales se colorean según si la variante se predice como patogénica (amarillo), benigna (verde) o de naturaleza desconocida (VUS) (blanco). (a) Predicción de variantes patogénicas (b) Predicción de variantes benignas.

## 2 OBJETIVOS

# OBJETIVOS

El objetivo central de esta tesis es el desarrollo de nuevas herramientas *in silico* altamente interpretables que sean específicas de la proteína *ATM* y que puedan ser usadas (i) en la clínica, apoyando al diagnóstico de cáncer de mama, y (ii) en la investigación biomédica, para priorizar análisis funcionales de variantes. Para lograr este objetivo, la presente tesis aborda los siguientes **objetivos específicos**:

1. Realizar un análisis comparativo de las herramientas *in silico* estándar para la predicción de variantes de *ATM*. Seleccionar los predictores individuales o combinación de ellos que den los mejores rendimientos de cara a su uso en la adaptación de las guías ACMG/AMP.
2. Generar predictores específicos para *ATM*, elaborando herramientas con alta capacidad predictiva e interpretabilidad. Para ello, se analizarán tres familias de predictores basados en tres formas diferentes de caracterizar de las variantes, mediante (i) características biofísicas, o (ii) características evolutivas y moleculares; o (iii) con características de metapredictor.
3. Diseño y generación de nuevas herramientas gráficas para el análisis de las predicciones de patogenicidad.
4. Aplicar las herramientas generadas a dos problemas de interés en

diagnóstico molecular: comparación de las dos versiones disponibles de las guías ACMG/AMP adaptadas a *ATM* y priorización de variantes VUS en *ATM* para su estudio posterior.

## 3 MATERIALES Y MÉTODOS

En este capítulo presentamos los aspectos metodológicos comunes a los dos capítulos de resultados (capítulos 4 y 5). Los aspectos metodológicos específicos se presentan en el capítulo correspondiente.

### **3.1 Conjunto de variantes para la construcción de predictores de patogenicidad**

El conjunto de variantes descrito en esta sección ha sido utilizado para generar los resultados de los capítulos 4 y 5 de esta tesis.

Las variantes patogénicas se obtuvieron de la literatura en el campo (Feliubadaló. *et al.*2021)

Las variantes benignas se obtuvieron mediante un modelo desarrollado en el grupo de Clinical and Translational Bioinformatics, ampliamente contrastado en problemas de predicción de patogenicidad (Riera. *et al.*, 2016; Padilla. *et al.*, 2019), y descrito en la sección 1.3.1.1 de la Introducción. Un elemento clave de este modelo es el alineamiento múltiple de secuencias (MSA) que describimos a continuación para el caso de *ATM*.

### 3.1.1 MSA

Para obtener las variantes benignas, tomamos el MSA disponible en el servidor de Align-GVGD (Tavtigian *et al.*, 2008). A partir de este alineamiento de proteínas se obtuvieron 322 variantes benignas. En la Figura 3.1 se observa cómo el entorno local del MSA para dos ejemplos de variantes benignas y en la Figura 3.2 para dos variantes patogénicas.

Cerca al N-terminal el MSA presenta menos zonas conservadas, razón por la cual la mayoría de variantes benignas se encuentran en esta mitad de la proteína. Por el contrario, cerca al C-terminal encontramos buena parte de la secuencia conservada (Figura 3.1).

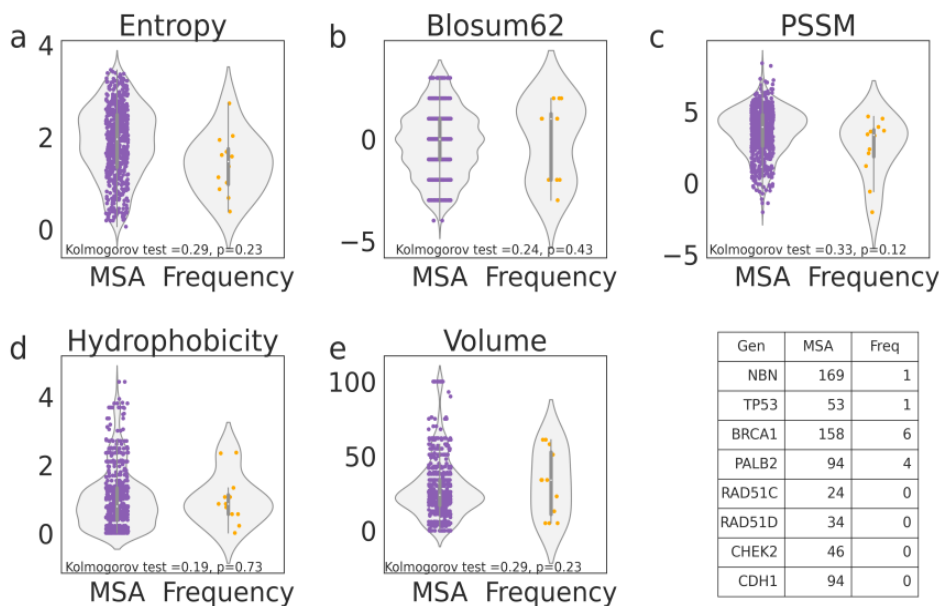




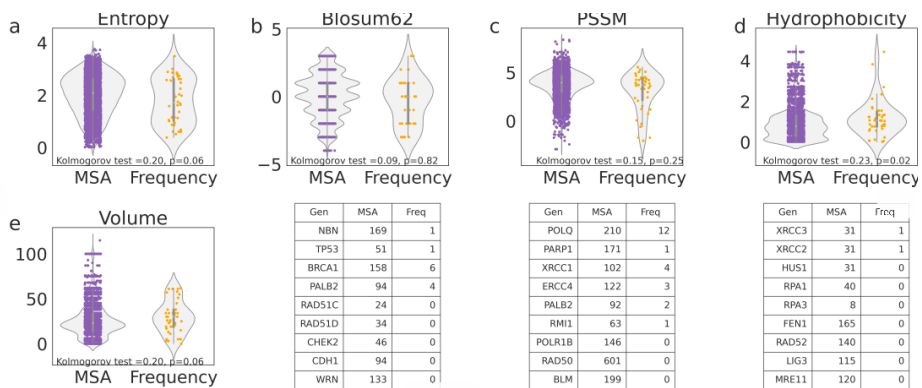
### **3.1.2 Comparación con el modelo de frecuencias de variantes benignas**

El modelo basado en MSA, produjo un mayor número de variantes que el modelo basado en la frecuencia de alelos, construido a partir de la base de datos 1000 Genomes (Auton. *et al.* 2015). Por eso lo elegimos para todo el trabajo.

Para cerciorarnos de que no hubiese diferencia fundamental entre ambos modelos, los comparamos en dos conjuntos diferentes de genes. El primero está constituido por diez genes utilizados en paneles de genes para diagnóstico de cáncer hereditario de mama y de ovario (Schroeder. *et al.* 2015) (Figura 3.3). El segundo conjunto de genes estaba constituido por veinte genes relacionados con el mecanismo de reparación de ruptura de doble cadena. Este listado de genes se obtuvo a partir de UniProt, utilizando la consulta “GO:0000724 y HDR y homo y sapiens” (Figura 3.4). Como podemos ver en las figuras 3.3 y 3.4, los modelos basados en homología y de frecuencia alélica muestran un comportamiento similar para las diferentes características discriminantes utilizadas en este trabajo. Ello confirma que el modelo basado en el uso de MSA dará resultados comparables al uso de un modelo basado en la frecuencia de las variantes.



**Figura 3.3. Comparación de la distribución de variantes benignas obtenidas por MSA y frecuencias alélicas.** En morado se representan las variantes obtenidas por MSA, y en amarillo las obtenidas por frecuencia, para las siguientes características: a) Entropía, b) Blosum62, c) PSSM, d) Hidrofobicidad, y e) Volumen. La tabla contiene el grupo de genes utilizados para la comparación.



**Figura 3.4. Comparación de la distribución de las variantes benignas obtenidas por MSA y Frecuencias alélicas.** En morado se representan las variantes obtenidas por MSA, y en amarillo las obtenidas por frecuencia, para las siguientes características: a) Entropía, b) Blosum62, c) PSSM, d) Hidrofobicidad, y e) Volumen. Las tablas contienen el grupo de genes utilizados para la comparación.

### 3.1.3 Variantes de entrenamiento

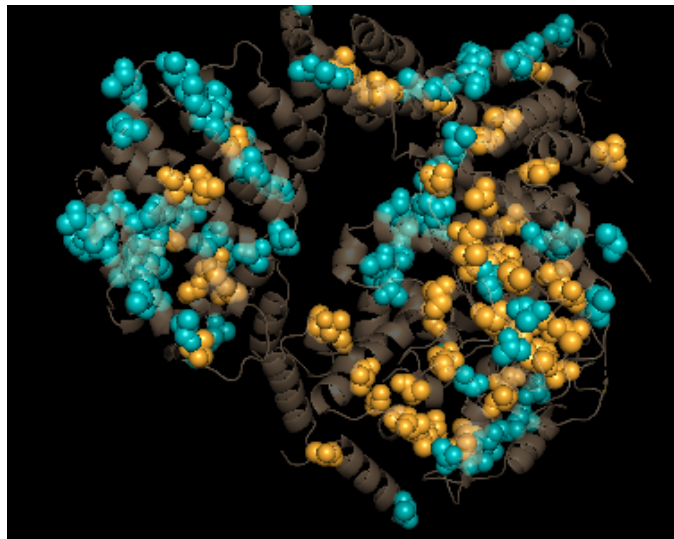
El conjunto de datos final estuvo constituido por un total de 425 variantes, 322 variantes benignas y 103 patogénicas. Estas variantes están distribuidas de la siguiente manera en la secuencia de la proteína *ATM*:

**Mitad N-terminal:** comprende del aminoácido 1 al 1959, contno aceptable iene 248 variantes en total, de ellas 213 benignas y 35 patogénicas.

**Mitad C-terminal:** comprende las posiciones 1960 al 3056, contiene

177 variantes en total, de ellas 109 benignas y 68 patogénicas.

En la Figura 3.5 representamos la ubicación de las variantes halladas en la mitad C-terminal. En esta mitad se encuentra más equilibrado el número de variantes patogénicas/benignas, y la representación basada en la estructura tridimensional de *ATM* (código PDB: 5NP0) muestra su disposición relativa. Observamos que ambos tipos de variantes se distribuyen de manera similar, y en muchos casos las variantes patogénicas/benignas ocupan ubicaciones vecinas.



**Figura 3.5. Distribución de las variantes benignas/patogénicas de la mitad C-terminal en la estructura de *ATM*.** Las esferas corresponden a las variantes de entrenamiento; en amarillo representamos las variantes patogénicas y en verde las benignas.

## **4 DESARROLLO DE UNA METODOLOGÍA *in silico* PARA LA ADAPTACIÓN DE LAS GUÍAS ACMG/AMP AL GEN *ATM***

## 4.1 Introducción

Las guías ACMG/AMP constituyen un valioso documento que da forma a un protocolo para la correcta clasificación de la naturaleza de las variantes. A fin de lograr este objetivo se ponderan e integran las puntuaciones que obtiene una variante de acuerdo con ocho criterios diferentes (Figura 1.18–1.19) (Richards S. *et al.*, 2015). Estas guías fueron concebidas para clasificar las variantes independientemente del gen en el que aparecen. Sin embargo, con el paso del tiempo, ha sido necesario adaptarlas a genes específicos e incluso para enfermedades particulares, ya que cada gen presenta características intrínsecas singulares así como cantidades y tipos de información diferentes (Feliubadaló. *et al.*, 2021).

En este capítulo presentamos el trabajo realizado en nuestro grupo dentro del proyecto liderado por el ‘Consortio español para la interpretación de las variantes genéticas en el gen *ATM*’ encaminado a adaptar las recomendaciones ACMG/AMP a dicho gen. Más concretamente, nuestro trabajo se centró en identificar y validar una combinación de herramientas *in silico* con la mayor capacidad predictiva posible para las variantes missense del gen *ATM*.

La necesidad de seleccionar un método *in silico* o combinación de ellos para un gen concreto surge de la observación de acuerdo con la cual los predictores normalmente utilizados tienen capacidades predictivas muy diferentes dependiendo del gen considerado (Riera *et*

*al.*, 2016). Ello se debe a factores técnicos complejos cuyo impacto en la capacidad predictiva en diferentes genes no puede establecerse a priori. Estos factores incluyen la forma como fue construido el algoritmo, el conjunto de datos utilizados, etc.

En este contexto, nuestro objetivo fue: (i) determinar la mejor combinación de predictores para la clasificación de las variantes missense en la proteína *ATM*, y (ii) evaluar la capacidad predictiva con variantes patogénicas que han sido obtenidas de la literatura revisando uno a uno los artículos primarios (Feliubadaló, *et al.*, 2021). Para lograr este objetivo consideramos dos estrategias. En la primera nos inspiramos en la metodología de Hart *et al.* (Hart *et al.*, 2019), diseñada originalmente para predicción de variantes patogénicas en los genes *BRCA1* y *BRCA2*, y que aquí hemos modificado para adecuarla a *ATM*. En la segunda estrategia utilizada seguimos la metodología propuesta por las guías ACMG/AMP respecto a la combinación de predictores, en una aproximación similar a la empleada por Fortuno *et al.* en *TP53* (Fortuno *et al.*, 2021).

## 4.2 Materiales y método específicos

### 4.2.1 Metodología adaptada de Hart *et al.*

La metodología Hart *et al.* desemboca en la creación de dos metapredictores: NVM y RF. En las secciones siguientes describiremos los puntos fundamentales de esta metodología. Posteriormente, en las secciones 4.3.1, mostraremos cómo la hemos adaptado a la construcción de nuestros predictores para *ATM*.

#### 4.2.1.1 Predictores primarios

Para la creación de los metapredictores de Hart *et al.* se utilizan los resultados predictivos de 30 herramientas *in silico* disponibles en la base de datos dbNSFP versión 4.0 (Liu X *et al.*, 2016). Dichos predictores son: SIFT, Polyphen-2-Hdiv, Polyphen-2-Hvar, MutationAssessor, FATHMM, VEST4, CADD, DEOGEN2, PROVEAN, Metalr, Metasvm, CADD, Damm, MetaLr, GerpNr, LRT, Phastcons30wayMammalian, Phastcons100wayVertebrate, Phylop30wayMammalian, Phylop100wayMammalian, Siphy29waylogodds. Adicional, dbNSFP genera una nueva clasificación sobre los valores de predicción para algunos predictores, que consiste



en una puntuación de clasificación que varía de 0 a 1, una puntuación de 0.9 significa que es más probable que sea perjudicial que el 90 % de todas las missense potenciales predichas por ese método, esto los autores lo han denominado RankScore, de estos, usamos: SIFT\_RankScore, Polyphen-2-Hdiv\_RankScore, Polyphen-2-Hvar\_RankScore, Metalr\_RankScore, Metasvm\_RankScore, MetaLr\_RankScore, GerpNr\_RankScore, Mutationassessor\_RankScore. Y Align-GVGD, tomada de la web <http://agvgd.hci.utah.edu/about.php> (Tavtigian *et al.*, 2008).

#### **4.2.1.2 Optimización de los puntos de corte**

Uno de los elementos claves que Hart emplea en la elaboración de los metapredicadores, es la modificación previa de los puntos de corte o umbrales de decisión de los predictores primarios. Para una variante determinada, cada uno de estos predictores emite un valor numérico que posteriormente transformaremos en una predicción binaria (patogénica/benigna), mediante el uso de un umbral de decisión. Por ejemplo, si el valor numérico es mayor que 0.75 diremos que la variante es patogénica (o benigna, ya que la escala varía con el método). Para optimizar los puntos de corte para cada predictor empleamos una muestra aleatoria del 40 % de nuestra base de datos de variantes. La optimización se realizó usando el paquete “optimal.cutpoint” en R (versión 3.6.3 <http://www.R-project.org>). Los

umbrales obtenidos buscan maximizar la sensibilidad y la especificidad de la predicción de nuestro conjunto de datos.

#### **4.2.1.3 Metapredicador Naïve Voting Method (NVM )**

El entrenamiento del metapredicador NVM se basó en la obtención del MCC (ver capítulo Introducción sección 13.1.4) máximo. Primero, se obtienen los MCC de todos los predictores primarios y se ordenan de mayor a menor valor de MCC, obteniendo: Predictor P1, P2, P3, P4, ..., P30. Posteriormente, combinamos los predictores de la siguiente manera:

combinación 1: P1+P2

combinación 2: P1+P2+P3

combinación 3: P1+P2+P3+P4

...

combinación 29 : P1+P2+P3+P4+...+P30

Para cada combinación se obtienen todas las métricas de MCC, sensibilidad, especificidad y ACC, y finalmente elegimos la combinación con mayor MCC como el nuevo predictor NVM. En este metapredicador excluimos los predictores primarios de RankScore.

NVM (NVM-validation) fue validado con un grupo de variantes

independientes que no formaron parte del proceso de selección de la mejor combinación. El método con mayor MCC obtenido durante la selección de la mejor combinación lo llamamos NVM-Training.

#### **4.2.1.4 Metapredictor Random Forest (RF)**

Para la construcción del modelo RF se utilizaron solo las predicciones de los predictores RankScore. Se usó el paquete randomForest de R con 500 árboles y demás parámetros por defecto.

#### **4.2.1.5 Variantes**

Para el desarrollo de los metapredictores según la metodología Hart usamos el total de las variantes descritas en la sección de variantes en Materiales y Métodos generales.

### **4.2.2 Metodología final aplicada al caso de ATM**

Tal como veremos en la sección de resultados, la aplicación de los métodos de Hart *et al.* a ATM no resultó totalmente convincente y decidimos probar una estrategia alternativa. En esta nueva metodología utilizamos pares de predictores primarios: en caso de coincidir damos el resultado por bueno; en caso contrario lo

descartamos. Así, la aplicación de nuestro criterio dará tres resultados: patogénica, benigna, y no predicción. Posteriormente, evaluamos qué combinación de predictores es mejor.

#### **4.2.2.1 Predictores primarios analizados**

En la búsqueda de los mejores predictores disponibles para *ATM* analizamos 13 predictores, SIFT, Polyphen-2-Hdiv, Polyphen-2-Hvar, Mutationassessor, FATHMM, VEST4, CADD, DEOGEN2, PROVEAN, Metalr, Metasvm, y REVEL. Sus predicciones fueron obtenidas de la base de datos dbNSFP versión 4.0 (Liu X *et al.*, 2016). También utilizamos Align-GVGD, cuyas predicciones se tomaron de la web <http://agvgd.hci.utah.edu/about.php> (Tavtigian *et al.*, 2008). Escogimos únicamente trece predictores debido a que nuestro número de variantes no es muy elevado e impide la construcción de modelos complejos (con muchos parámetros).

#### **4.2.2.2 Variantes**

A efectos de entrenamiento y validación, empleamos las variantes descritas en Materiales y Métodos, tratando por separado las pertenecientes a las mitades C-terminal y N-terminal. Ello dio lugar a

un predictor para cada mitad de la proteína.

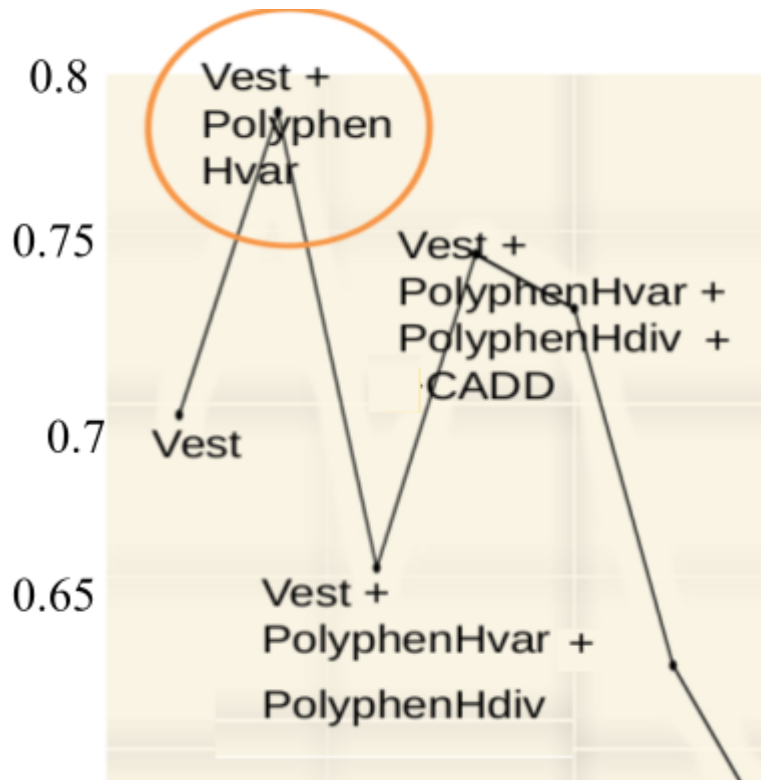
## **4.3 Resultados**

### **4.3.1 Adaptación Hart**

Con la metodología de Hart creamos dos metapredictores NVM y RF, construyéndose a partir de los predictores primarios, previa optimización de los puntos de corte.

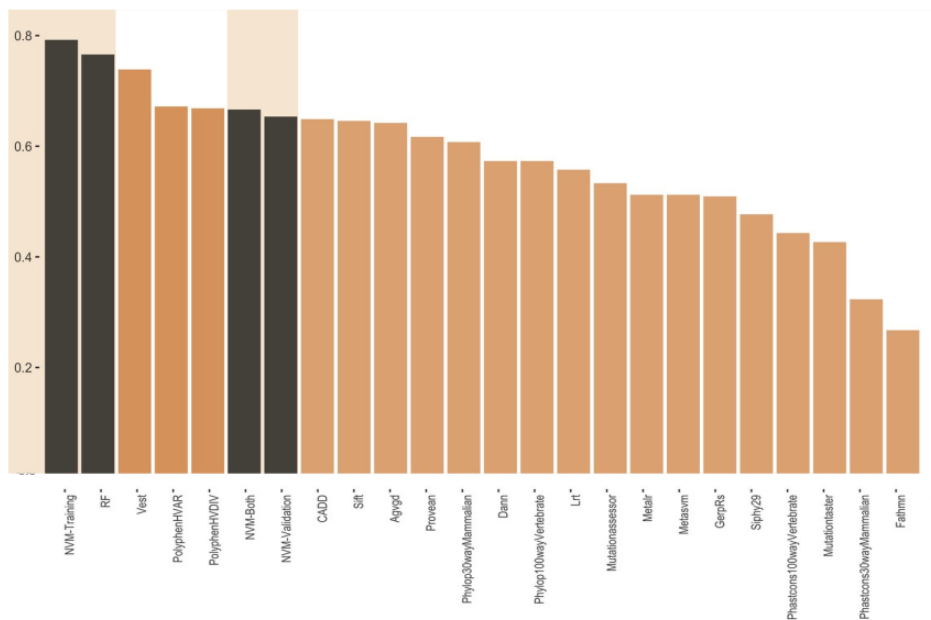
#### **4.3.1.1 Metapredictores NVM Y RF**

Para el entrenamiento de NVM analizamos los valores de MCC obtenidos desde el predictor primario con mayor MCC que fue el predictor VEST4, y de las posteriores combinaciones incrementales de predictores (Figura 4.1). En nuestros resultados, las mejores combinaciones fueron aquellas que tenían un menor número de predictores. Entre todas ellas, la combinación VEST4+PolyphenHvar fue quien presentó el mejor resultado, con un MCC de 0.78 (evaluado con las variantes de entrenamiento), mientras que en las combinaciones con más de cuatro predictores se obtuvieron valores de MCC inferiores a 0.6.



**Figura 4.1** *Proceso de construcción de NVM y su relación con valores de MCC. Cada punto representa la adición de un nuevo predictor. En esta gráfica vemos que las cuatro primeras adiciones dan los mejores resultados. Posteriormente, se observa una caída importante en los valores de MCC.*

Con base a los resultados obtenidos en la combinación de predictores, para NVM finalmente escogimos la combinación de dos predictores VEST4+PolyphenHvar. Esta combinación presenta un MCC de 0.66 en el conjunto de validación y de 0.78 en el de entrenamiento (NVM-Training) (Figura 4.2).



**Figura 4.2 Comparativa de los valores MCC de NVM y RF con respecto a otros predictores tras la optimización del punto de corte.** En marrón oscuro se representan los metapredictores desarrollados. El MCC de NVM-Training se pone únicamente como referencia, puesto que incluye las variantes con las cuales fue realizada la optimización del punto de corte y de entrenamiento. Por el contrario, NVM-Validation proporciona una estimación menos sesgada del rendimiento de NVM, ya que el MCC no incluye variantes usadas en el proceso de entrenamiento ni de optimización. NVM-Both incluye todas las variantes. En naranja se representan otros predictores como punto de comparación.

Como es natural, durante el entrenamiento de los metapredictores obtuvimos valores MCC altos: 0.75 en el RF y 0.78 en NVM-Training (Figura 4.2). Ello sugiere que las variables empleadas pueden tener un valor predictivo. En NVM-validation vemos (Figura 4.2), tal como se

esperaba, que al evaluar el rendimiento del predictor en un conjunto independiente de variantes los valores MCC bajan. Este también está presente en NVM-Both, donde usamos todas las variantes.

Hay que señalar que los resultados de validación de NVM son inferiores al rendimiento de predictores primarios como VEST4 y PolyPhen utilizados en su construcción. Este es un aspecto importante, ya que en principio desarrollamos NVM con la idea de obtener una capacidad predictiva mejor a la de los métodos primarios. Este resultado es un indicador de una situación de sobreajuste que se da cuando el número de parámetros del predictor es muy elevado en relación a la muestra que disponemos para entrenarlo.

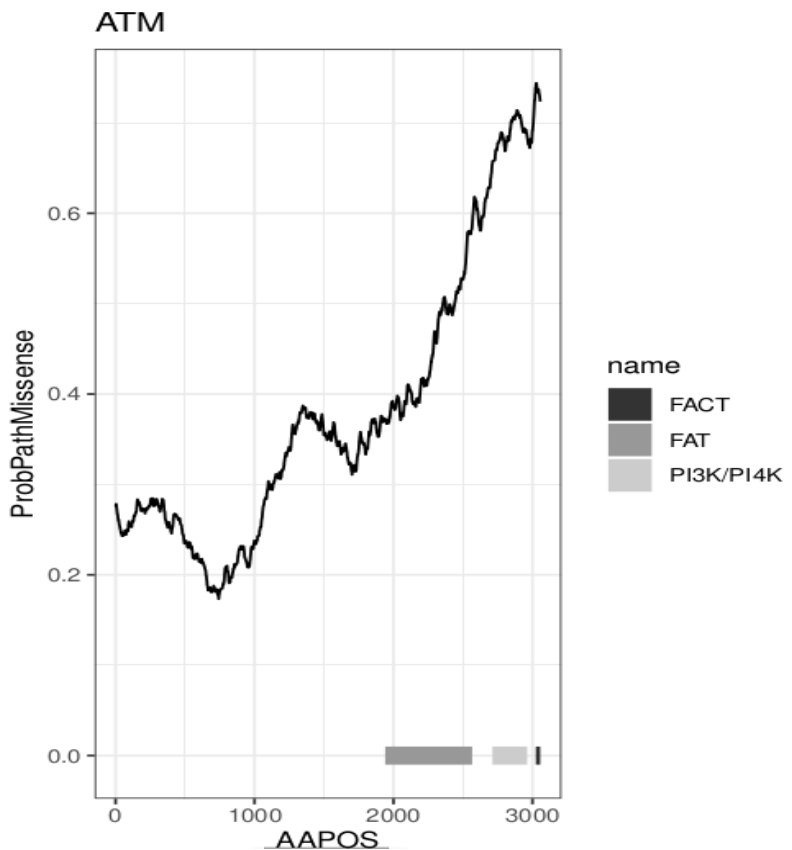
#### **4.3.1.2 Predicción sistemática de variante missense con NVM**

Para explorar el origen del posible sobreajuste, estudiamos si las predicciones de NVM mostraban algún sesgo especial, asociado a la localización en la secuencia. Para ello evaluamos con NVM todas las variantes missense posibles en *ATM* (19 por cada posición) y analizamos la distribución resultante de estas predicciones a lo largo de la secuencia.

Se observó un enriquecimiento marcado para las variantes patogénicas predichas por NVM en los dominios funcionales



conocidos de *ATM* (Figura 4.3). Concretamente, el análisis del dominio FACT mostró que entre el 30 y el 40 % de todas las variantes predichas generaban un daño en la proteína. De manera más agresiva, del 50 al 60 % de las missense en el dominio PI3K fueron predichas como patogénicas y más del 65 % de aquellas que ocurrían en el dominio FACT. En general, en la región C-terminal (aproximadamente residuo 1960–3056) se pronosticó el mayor porcentaje de variantes patogénicas. Curiosamente, hay un pequeño pico antes de los dominios, alrededor de las variantes 1800-1900 que llega casi a un 40% de variantes predichas como perjudiciales (Figura 4.3). Este pico podría ser debido a alguna característica funcional/estructural específica de esta región, que adopta, de acuerdo con las anotaciones de InterPro, un ‘armadillo-type fold’ que iría del residuo 256 al 2496.



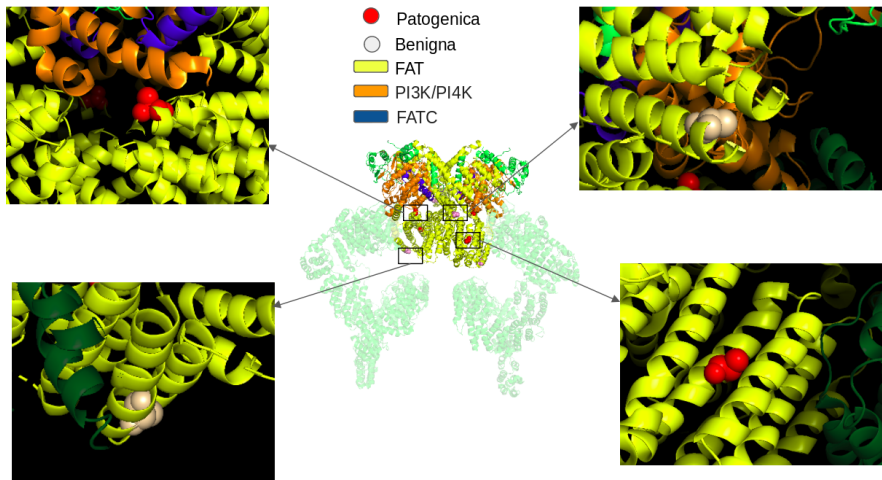
**Figura 4.3** *Fracción de variantes patogénicas predichas por NVM a lo largo de la proteína ATM.* En la parte inferior del gráfico se representa, mediante cajas, las posiciones de los dominios funcionales de ATM en la escala de grises FAT (Tonalidad intermedia), PI3K/PI4K (Tonalidad clara), y FACT (en Negro).

En general vemos que las predicciones de patogenicidad de NVM son más frecuentes en las regiones funcionales más relevantes de la proteína. Aunque este efecto podría ser debido a que las variantes en

las regiones funcionales son potencialmente más dañinas, también coincide con un posible sesgo en el entrenamiento, debido a una distribución desigual de las variantes patogénicas a lo largo de la secuencia. Por lo tanto, este análisis no nos permite descartar la posibilidad de sobreajuste en NVM.

#### **4.3.1.3 Visualización estructural**

Para ver si la distribución estructural de las variantes podía explicar el comportamiento predictivo de NVM, representamos en la estructura tridimensional de *ATM* las variantes benignas y patogénicas de nuestro conjunto de datos. El análisis de los datos resultantes muestra que este no ha sido el caso, no observándose diferencias patentes entre ambas poblaciones. Por ejemplo, en la Figura 4.4 vemos varios casos de variantes benignas y patogénicas que muestran que ambos tipos de variantes conviven, generalmente, en ubicaciones espaciales muy similares y cercanas en el espacio.

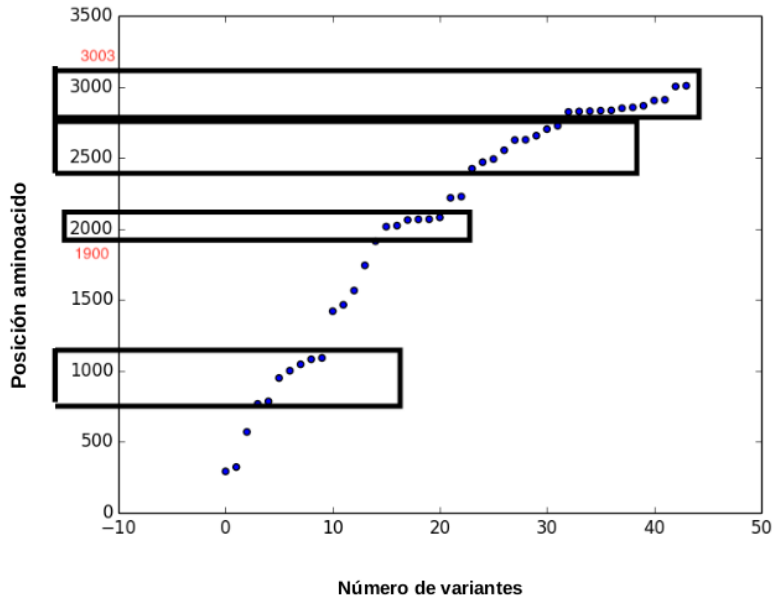


**Figura 4.4 Visualización estructural del dímero de ATM.** En amarillo resaltamos el dominio FAT, en naranja PI3K/PI4K, y en azul FACT, Los dos casos de variantes patogénicas se muestran en rojo y las dos variantes benignas en blanco. El resto de la proteína se representa en verde.

#### 4.3.1.4 Posible existencia de un sesgo de dominios en NVM

Si resumimos los resultados anteriores tenemos lo siguiente. Por una parte, la Figura 4.2 indica la posibilidad de un sobreajuste en NVM. Los resultados de la Figura 4.3 no permiten descartar esta posibilidad, ya que son consistentes con un posible sesgo asociado a la distribución de dominios funcionales a lo largo de la secuencia. Este

sesgo sería consistente con la distribución irregular de las variantes patogénicas y benignas a lo largo de la secuencia (Figura 4.5). Finalmente, la Figura 4.4 indica que la diferencia observada en la distribución a lo largo de la secuencia no existe a nivel estructural. En su conjunto, estas observaciones son consistentes con el hecho de que la capacidad predictiva del método fuese parcialmente debida a diferencias en las propiedades evolutivas entre regiones de la proteína, e.g., fragmento N-terminal vs. fragmento C-terminal. Es decir, que un componente de la predicción de NVM fuese la localización de la variante en uno de estos dos fragmentos. Desde el punto de vista de la aplicabilidad de NVM, esto supone un problema, ya que puede generar predicciones erróneas, con un sesgo que favorece las predicciones patogénicas para aquellas variantes que se localizan en el fragmento C-terminal, por ejemplo.



**Figura 4.5** *Representación del número de variantes de entrenamiento patogénicas respecto a la posición en la proteína ATM. Con recuadros negros resaltamos posiciones en las cuales se encuentran la mayoría de variantes patogénicas.*

### 4.3.2 Combinación de predictores

Los resultados anteriores nos llevaron a replantear completamente la metodología predictiva. En lo que respecta al modelo computacional, decidimos primar principalmente la simplicidad. Adicionalmente, tuvimos en cuenta la distribución de dominios funcionales, para evitar posibles sesgos.

Para construir el predictor tomamos como referencia el trabajo de Fortuno *et al.* en el caso análogo de *TP53*. Estos autores estudian

sistemáticamente diferentes combinaciones de predictores, siguiendo las reglas de combinación de las guías ACMG/AMP porque no requieren ningún ajuste de parámetros.

Para evitar el sesgo por distribución de variantes en la secuencia, optamos por trabajar independientemente con las mitades N-terminal (residuos 1 al 1959) y el C-terminal (residuos 1960 al 3056) de la proteína. Esta división se basó en consideraciones funcionales: la mitad N-terminal comprende los dominios HEAT y TAN, mientras que la mitad C-terminal comprende los dominios FACT, FAC y PI3K/PI4K (Bernstein JL. *et al.*, 2017). Esta nueva aproximación y sus resultados, que presentamos a continuación, constituyen nuestra contribución a la adaptación al gen *ATM* de las recomendaciones ACMG/AMP realizada por el ‘Consortio español para la interpretación de las variantes genéticas en el gen *ATM*’.

A continuación presentamos por separado los predictores obtenidos para las partes N- y C-terminal de la proteína.

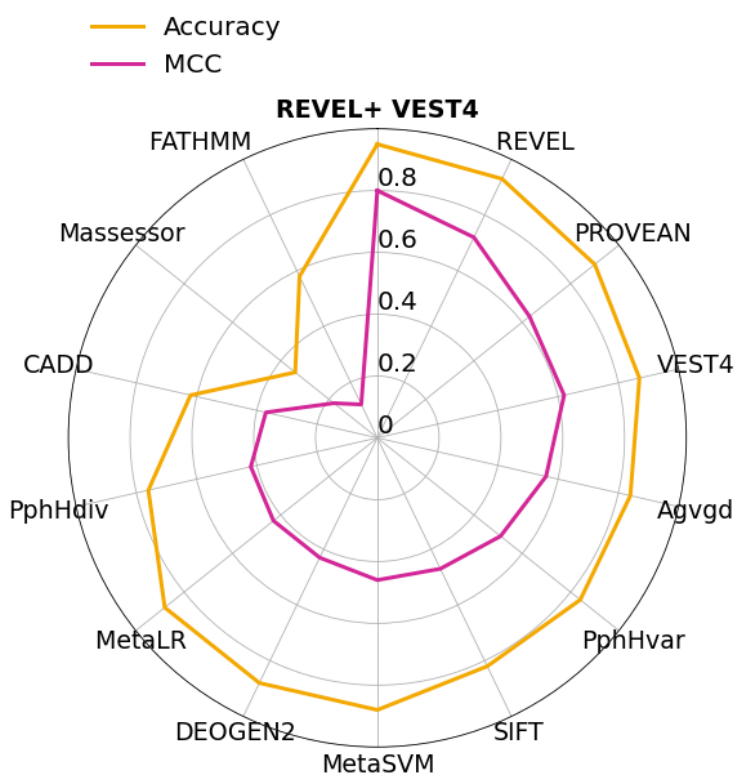
#### **4.3.2.1 N-terminal**

En la mitad N-terminal los predictores individuales con mejor capacidad predictiva son REVEL, PROVENA y VEST4, con valores respectivos del MCC de 0.72, 0.63 y 0.62 (Figura 4.6). Los valores de sensibilidad de REVEL y PROVENA se encuentran por debajo del 0.8, pero destaca su especificidad elevada de 0.97 y 0.94. VEST presenta

estas métricas más balanceadas y cercanas al 0.8 (Tabla 4.1). Los predictores que obtuvieron las peores capacidades predictivas fueron Fathmm y Mutationassessor, con valores del MCC de 0.12 y 0.18 respectivamente (Figura 4.6). Estos predictores tienen un desequilibrio predictivo muy importante, ya que Fathmm muestra una especificidad de 0.98, Mutationassessor tiene una sensibilidad muy elevada cercana a 1 (Tabla 4.1). Ello los hace inadecuados para su uso en diagnóstico.

Siguiendo el trabajo de Fortuno *et al.*, exploramos todas las combinaciones posibles de pares de predictores. Hallamos que REVEL+VEST daba los mejores resultados, con un MCC de 0.8 y una sensibilidad y especificidad de 0.83 y 0.97 respectivamente, un rendimiento superior al de cualquier predictor individual (Figura 4.6).





**Figura 4.6 Radar comparativo de las métricas MCC Y ACC para los predictores analizados y la combinación con mayor capacidad predictiva en la mitad N-Terminal.** Los predictores se ordenan de mayor a menor valor de MCC siguiendo el sentido anti-horario.

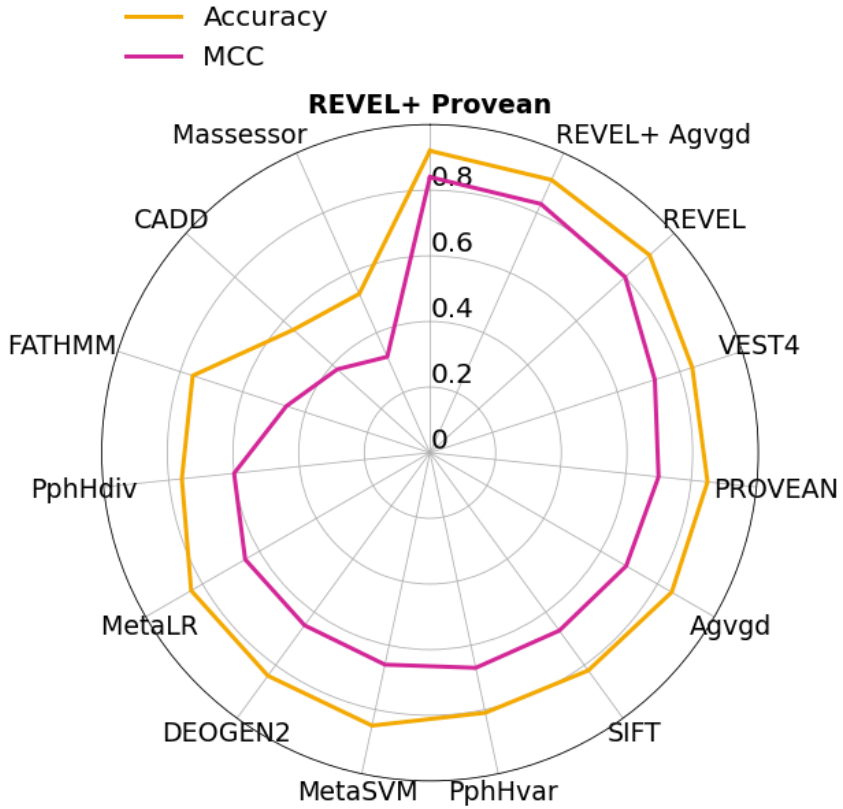
<b>Predictor</b>	<b>mcc</b>	<b>acc</b>	<b>sens</b>	<b>spec</b>
REVEL	0.72	0.93	0.69	0.97
PROVEAN	0.63	0.9	0.69	0.94
VEST4	0.62	0.87	0.86	0.88
Align-GVGD	0.56	0.88	0.67	0.92
Polyphen2Hvar	0.51	0.84	0.74	0.86
SIFT	0.47	0.82	0.74	0.83
Metasvm	0.46	0.88	0.33	0.98
DEOGEN2_	0.43	0.88	0.21	1
Metalr	0.43	0.88	0.33	0.97
Polyphen2Hdiv	0.42	0.76	0.79	0.76
CADD	0.37	0.62	0.95	0.56
Mutationassessor	0.18	0.34	0.98	0.23
Fathmm	0.12	0.85	0.05	0.99
REVEL+VEST4	0.8	0.952	0.83	0.97

**Tabla. 4.1. N-Terminal (1-1959). Resumen de las métricas MCC,ACC, Sensibilidad y Especificidad evaluadas.**

#### 4.3.2.2 C-terminal

Para el C-terminal encontramos que los predictores individuales que mejor predecían variantes de *ATM* fueron REVEL, VEST4 y PROVEAN, con MCC de 0.8, 0.72 y 0.7 respectivamente (Fig. 4.7). Los valores de sensibilidad y especificidad se encontraban por encima o muy cercanos al 0.8 (Tabla 4.2), lo cual indica un buen balance en las predicciones de las poblaciones benignas y patogénicas. Mientras tanto, los peores predictores fueron Mutationassessor y CADD, con unos MCC de 0.32 y 0.38 respectivamente (Figura 4.7), presentando además una especificidad muy pobre (Tabla 4.2).

En cuanto a la combinación de predictores, obtuvimos los mejores resultados para el C-terminal con dos combinaciones: REVEL+PROVEAN y REVEL+Align-GVGD. Los MCCs obtenidos fueron 0.84 y 0.83 respectivamente, valores que superan los encontrados en los predictores individuales (Figura 4.7).



**Figura 4.7 Radar comparativo de las métricas MCC Y ACC para los predictores analizados y la combinación con mayor capacidad predictiva en la mitad C-Terminal.** Los predictores se ordenan de mayor a menor valor de MCC siguiendo el sentido anti-horario.

<b>Predictor</b>	<b>mcc</b>	<b>acc</b>	<b>sens</b>	<b>spec</b>
REVEL	0.8	0.9	0.97	0.86
VEST4	0.72	0.84	0.98	0.76
PROVEAN	0.7	0.85	0.9	0.83
Align-GVGD	0.69	0.85	0.9	0.82
SIFT	0.67	0.82	0.95	0.75
Polyphen2Hvar	0.67	0.81	0.98	0.71
Metasvm	0.66	0.85	0.74	0.91
DEOGEN2_	0.65	0.84	0.62	0.96
Metalr	0.65	0.84	0.74	0.9
Polyphen2Hdiv	0.6	0.76	0.98	0.64
Fathmm	0.46	0.76	0.52	0.89
CADD	0.38	0.56	1	0.32
Mutationassessor	0.32	0.53	0.98	0.27
REVEL+PROVEAN	0.844	0.922	0.982	0.888
REVEL+Align-GVGD	0.833	0.916	0.982	0.879

**Tabla. 4.1. C-Terminal (1960-3056). Resumen de las métricas MCC,ACC, Sensibilidad y Especificidad.**

## 4.4 Discusión

### 4.4.1 Metodología Hart

La metodología Hart (Hart *et al.*, 2019), la cual implementamos inicialmente, nos permitió conocer los datos y generar metapredictores NVM y RF cuyos rendimientos fueron aceptables. Sin embargo, observamos que el proceso de reajuste de los umbrales o puntos de corte podría inducir a un sobreajuste no deseado. Este fenómeno podría verse favorecido por el elevado número predictores, que además generaba una complejidad formal que quisimos minimizar. Gracias a estos resultados iniciales, también encontramos otras posibles fuentes de error en la construcción de predictores para *ATM*, relacionadas con la distribución no homogénea de las variantes a lo largo de la secuencia. Este último resultado nos llevó a tratar las dos mitades de la proteína por separado.

### 4.4.2 Combinación de predictores

De forma general, para la adaptación de los criterios de predicción *in silico* a *ATM*, nos basamos en las recomendaciones de los autores de las guías ACMG/AMP (Richards S. *et al.*, 2015), evaluando la capacidad predictiva de los predictores diseñados con variantes patogénicas bien establecidas y descartando las predicciones cuando los predictores

seleccionados discrepaban.

En el desarrollo de la mejor estrategia de predicción *in silico* para *ATM* propusimos una metodología donde se considera por separado cada mitad de la proteína, ya que cada una de estas presenta dominios funcionales diferentes (Figura 1.6). Esta estrategia demostró ser un sistema más ajustado para la clasificación de variantes *ATM* porque evitaba un posible sesgo predictivo debido a diferencias en la distribución de las variantes a lo largo de la secuencia (Figura 4.5).

Para el criterio de predicción computacional, se estableció un sistema de combinación de dos predictores inspirado en el reciente trabajo de Fortuno *et al.* (Fortuno *et al.*, 2021). Para la mitad N-terminal se estableció la combinación de REVEL+VEST4, y para el C-terminal la REVEL+PROVEAN. Merece la pena señalar que en ambas combinaciones participa el predictor REVEL, ya que este predictor ha sido utilizado como único predictor en la adaptación de las guías ACMG/AMP a *ATM* realizada por el comité de expertos de ClinGen (HBOPC VCEP. ClinGen; 2022). Ello por una parte confirma la validez de nuestro trabajo y, por otra, abre la posibilidad a comparar las diferentes estrategias para la generación de evidencia computacional. Este último aspecto se tratará en el próximo capítulo en el que realizamos un análisis comparativo de ambas aproximaciones.

Finalmente, conviene comentar que la capacidad predictiva de la estrategia de predicción que hemos presentado, aunque es viable y

competitiva, puede estar ligeramente sobreestimada. Ello es debido a que no conocemos con exactitud hasta qué punto nuestras variantes de entrenamiento han sido usadas en la construcción de alguno de los predictores seleccionados (REVEL, PROVEAN, y Align-GVGD). Adicionalmente, el hecho de que los predictores seleccionados están basados en tecnología de tipo caja negra y no han sido pensados en la predicción exclusiva de *ATM* limita su interpretabilidad, un aspecto cada vez más valorado en el desarrollo de herramientas de machine learning (Rudin, 2019). Todas estas consideraciones nos llevaron a plantearnos el desarrollo de predictores más simples que a su vez fuesen más interpretables y específicos para la proteína *ATM*. En el próximo capítulo se describe el trabajo que hemos realizado en esta nueva dirección.



## **5 Hacia una mejor clasificación *in silico* de las variantes missense de *ATM* utilizando herramientas de análisis visual**

Nos gustaría señalar que el trabajo descrito en este artículo ha sido escrito y está en proceso de revisión en la revista *Journal of Molecular Diagnostics*.

## 5.1 Introducción

En la actualidad sabemos que las variantes patogénicas de la línea germinal en el gen mutado de ataxia-telangiectasia (*ATM*) aumentan el riesgo de varios cánceres hereditarios (Hu *et al.*, 2021; Kaur *et al.*, 2020; Hannan *et al.*, 2021). La identificación de los portadores de estas variantes mediante pruebas de secuenciación ofrece a los pacientes y familiares la posibilidad de un manejo clínico personalizado (González-Santiago *et al.*, 2019). Sin embargo, el alcance de este enfoque actualmente está limitado por la pequeña cantidad de variantes de *ATM* con impacto clínico conocido. En este trabajo, abordamos esta problemática, centrándonos en la clasificación *in silico* de variantes missense de *ATM* y cómo podemos mejorarla. En particular, abordamos dos carencias de los predictores de patogenicidad que limitan nuestra confianza en las clasificaciones computacionales.

La primera carencia que consideramos es la baja interpretabilidad de las predicciones de estas herramientas. Sabemos que estas predicciones no son totalmente precisas (Figura 1.17), por lo tanto,

para evitar la propagación de errores en la cadena diagnóstica, nuestra primera opción es la interpretación de las predicciones para establecer su fiabilidad. Con este fin, necesitamos comprender la relación entre las predicciones y las propiedades discriminantes utilizadas para generarlas (Rudin, 2019). Sin embargo, esto es técnicamente difícil debido a la complejidad de los modelos predictivos y a la cantidad de propiedades que utilizan. Alternativamente, podemos desarrollar predictores más simples o más interpretables. Basándonos en investigaciones previas del grupo (Riera *et al.*, 2016; Padilla *et al.*, 2021), hemos construido dos predictores minimalistas de patogenicidad específicos de *ATM*; son minimalistas en el sentido de que utilizan pocos descriptores de las variantes o, lo que es lo mismo, pocas propiedades discriminantes. Esa baja complejidad los vuelve automáticamente más interpretables. En esta línea de conseguir una alta interpretabilidad, otra opción que también exploramos es la de construir un predictor intrínsecamente más interpretable, utilizando estimaciones biofísicas del impacto molecular de las variantes (por ejemplo, el cambio de estabilidad en la proteína tras la mutación). Hasta ahora, el uso de esta opción estaba limitado por la baja calidad de la información estructural disponible para estos cálculos. Sin embargo, la reciente publicación de una estructura *ATM* de buena resolución (Warren *et al.*, 2022) ha abierto la puerta a cálculos de energía más precisos. Aquí se ha utilizado esta información para explorar si podemos desarrollar un predictor de patogenicidad competitivo para variantes missense de

*ATM*, basado en el uso de estimaciones del impacto de las variantes en la energía libre de la proteína.

La segunda carencia que abordamos en este trabajo es la falta de herramientas simples, más allá de los predictores de patogenicidad, que permitan a los profesionales evaluar las variantes y sus predicciones a la luz de lo que sabemos sobre las variantes de *ATM*. Para hacer frente a esta necesidad, hemos desarrollado una representación gráfica en la que las poblaciones de *ATM* missense patogénicas y benignas sirven como referencia para analizar las variantes desconocidas. Ilustramos la utilidad de esta herramienta en tres problemas relevantes en la clasificación clínica de variantes missense de *ATM*. Primero, describimos cómo nuestra representación puede respaldar la evaluación de la predicción. En segundo lugar, ilustramos su uso para analizar y priorizar variantes de significado desconocido (VUS). Y, en tercer lugar, mostramos cómo nuestra representación permite comparar las dos adaptaciones a *ATM* de las guías ACMG/AMP (Feliubadaló. *et al.*2021; HBOPC VCEP. ClinGen; 2022) en su tratamiento de la evidencia computacional.

Los predictores desarrollados en esta sección están dirigidos para la mitad C-terminal  $\geq 1960$  ya que la mayor parte de variantes patogénicas se concentran en esta mitad.

## 5.2 Materiales y métodos específicos

### 5.2.1 Características de entrenamiento

En las tres secciones siguientes, describimos las propiedades discriminantes empleadas para la construcción de cada predictor de patogenicidad presentado, así como las fuentes de las cuales se obtuvieron. Estas propiedades reflejan diferentes visiones del problema de clasificación de variantes.

#### 5.2.1.1 RF\_Biophys

Para el predictor biofísico, las variantes se caracterizaron con seis estimaciones del impacto de la variante sobre la energía de la proteína y sus interacciones: tres versiones del cambio de estabilidad ( $\Delta\Delta G$ ) del monómero de *ATM* y tres del impacto en la energía libre de unión del dímero ( $\Delta\Delta G_{\text{unión}}$ ).

Estas estimaciones se obtuvieron a partir de los siguientes modelos. Del grupo de M.Roman, utilizamos SNPmusic (Dehouck *et al.* 2011) (<http://babylone.3bio.ulb.ac.be/3bio/softs/>) y BeATMusic (Dehouck *et al.* 2013) (<http://babylone.ulb.ac.be/beATMusic/>), para  $\Delta\Delta G$  y

$\Delta\Delta G_{\text{unión}}$ , respectivamente. Del grupo de T.L.Blundell, usamos mCSM (<http://biosig.unimelb.edu.au/mcsm/stability>) y mCSMP ([http://biosig.unimelb.edu.au/mcsm/protein\\_protein](http://biosig.unimelb.edu.au/mcsm/protein_protein)), para  $\Delta\Delta G$  y  $\Delta\Delta G_{\text{unión}}$ , respectivamente. Del grupo de E.Alexov usamos SAAFEC (<http://compbio.clemson.edu/SAAFEC-SEQ/>) y SAAMBE ([http://compbio.clemson.edu/saambe\\_webserver/index3D.php](http://compbio.clemson.edu/saambe_webserver/index3D.php)), para  $\Delta\Delta G$  y  $\Delta\Delta G_{\text{unión}}$ , respectivamente.

Estas estimaciones de energía se obtuvieron para dos versiones distintas de la estructura de ATM, una de baja resolución (PDB: 5NPO) y otra de alta resolución (PDB: 7SIC). Los predictores resultantes fueron RF\_Bphyslr y RF\_Bphys, respectivamente.

#### 5.2.1.2 RF\_Bioinf

Para RF\_Bioinf, utilizamos una mezcla de propiedades basadas en el patrón de conservación de la familia *ATM* y en índices biofísicos simples. Más precisamente, utilizamos cinco propiedades elegidas por su buen rendimiento cuando se utilizan para construir predictores específicos de proteínas para BRCA1/2 (Padilla *et al.* 2019). Las tres primeras son tres descriptores sencillos de la mutación en términos de propiedades del cambio aminoacídico: el cambio de volumen y de hidrofobicidad (Fauchere y Pliska 1983) entre los aminoácidos nativo y mutante, y los elementos de la matriz Blosum62 (Henikoff y Henikoff 1992). Las otras dos propiedades se calculan a partir del

alineamiento de las secuencias ortólogas de la familia *ATM*, recuperado del servidor Align-GVGD (<http://agvgd.hci.utah.edu>) (Tavtigian *et al.* 2008), y son: la entropía de Shannon y los elementos de la matriz de puntuación (pssm), ambos descritos anteriormente.

### 5.2.1.3 RF\_Metap

Los metapredictores de patogenicidad, son clasificadores entrenados a partir de las predicciones de herramientas preexistentes y constituyen una prometedora aproximación al problema de predicción de patogenicidad (Özkan *et al.* 2021). Para representar dicha estrategia, obtuvimos un metapredictor, que denominamos RF\_Metap, y para el que caracterizamos las variantes con las predicciones de seis clasificadores conocidos: REVEL (Ioannidis *et al.* 2016), PROVEAN (Choi *et al.* 2012), VEST4 (Carter *et al.* 2013), Polyphen-2-Hdiv (Adzhubei *et al.* 2010), SIFT (Kumar *et al.* 2009), y Align-GVGD (Tavtigian *et al.* 2006)

## 5.2.2 Algoritmo Random Forest

Utilizamos el algoritmo Random Forest para construir nuestros tres predictores de patogenicidad, RF\_Bphys, RF\_Bioinf y RF\_Metap. Para ello usamos el paquete en R “randomForest”, entrenándolo con 500

árboles y los parámetros por defecto.

Dado que en nuestro conjunto de datos, las variantes benignas (109) eran más abundantes que las patógenicas (68), aplicamos el algoritmo de remuestreo SMOTE, con el paquete R “smotefamily”, para equilibrar el conjunto de datos.

Para estimar el rendimiento de estos predictores empleamos una versión particularmente rigurosa de la estrategia de validación cruzada LOOCV. En LOOCV estándar (Riera *et al.*, 2014), se recorre una a una el conjunto de variantes. A cada paso, se retira una variante del grupo de datos de entrenamiento y se utiliza el resto para entrenar el predictor que, una vez obtenido, se aplicará a la variante retirada. En nuestra versión del LOOCV, buscamos evitar una posible transferencia de información entre el conjunto de entrenamiento y la variable separada, debido a la existencia de variantes que comparten la misma ubicación de secuencia. Para ello, en lugar de omitir una variante a cada paso, omitimos simultáneamente todas las variantes que ocurren en la misma posición. Estas variantes luego se predicen con el RF entrenado.

Todo el proceso anterior se repitió 100 veces, para suavizar las fluctuaciones estadísticas resultantes del muestreo SMOTE. Como resultado, obtuvimos 100 predicciones por variante. La predicción final se obtuvo aplicando la regla de la mayoría: la variante fue anotada como patógena cuando tenía 50 o más predicciones de patogenicidad, de lo contrario se consideraba benigna.



### 5.2.3 Evaluación comparativa de la capacidad predictiva

A fin de comparar la capacidad predictiva de nuestros métodos respecto a la de otros métodos ya existentes, calificamos las variantes en nuestro conjunto de datos con diecisiete predictores de patogenicidad estándar. Estos predictores representan una muestra amplia de las herramientas disponibles/utilizadas en la actualidad. Las puntuaciones de catorce de estos predictores se recuperaron de la base de datos dbNSFP4 (Liu *et al.* 2016), versión 4.0: REVEL (Ioannidis *et al.* 2016), PROVEAN (Choi *et al.* 2012), VEST4 (Carter *et al.* 2013), Polyphen-2-Hvar y Polyphen-2-Hdiv (Adzhubei *et al.* 2010), SIFT (Kumar *et al.* 2009), MetaSVM y MetaLr (Dong *et al.* 2015), DEOGEN2 (Raimondi *et al.* 2017), CADD (Rentzsch *et al.* 2019), MutationAssessor (Reva *et al.* 2011), FATHMM (Shihab *et al.* 2013), LRT (Chun y Fay 2009) y GenoCanyon (Lu *et al.* 2015). Para las tres herramientas restantes, recuperamos las predicciones de los sitios web correspondientes: Align-GVGD (Tavtigian *et al.* 2006) ([http://agvgd.hci.utah.edu/agvgd\\_input.php](http://agvgd.hci.utah.edu/agvgd_input.php); opción de alineamiento de secuencia múltiple ATM: humano a erizo de mar), PON-P2 (Niroula *et al.* 2015) (<http://structure.bmc.lu.se/PON-P2/>), y PMut (López-Ferrando *et al.* 2017a) (<http://mmb.irbbarcelona.org/PMut/>).

## 5.2.4 Cálculo de Fiabilidad

Para ayudar a los usuarios a establecer una escala de confianza en nuestros predictores, diseñamos un índice de fiabilidad para las predicciones. Este cálculo implementa la idea según la cual la fiabilidad de una predicción depende de su ubicación respecto al umbral de decisión que utiliza el predictor para binarizar su resultado (Ferrer-Costa *et al.* 2004).

En este trabajo, como utilizamos combinaciones de un centenar de RF, no podemos definir un único output ya que cada RF tiene el suyo. Para superar este problema decidimos utilizar una aproximación empírica al uso de umbrales. Nuestro cálculo de la fiabilidad se divide en dos partes. La primera, es una clusterización que produce una partición del espacio de las propiedades predictivas del clasificador. Esta parte permitirá localizar nuestra predicción de interés en relación al pseudo-umbral de decisión. La parte de clusterización se realiza solo una vez, y sus resultados se pueden utilizar para calcular la fiabilidad de cualquier predicción. La segunda parte del cálculo corresponde a la asignación del valor numérico de la fiabilidad. Esta parte se basa en la clusterización pero, a diferencia de esta, es un paso específico para cada predicción. A continuación detallamos los aspectos técnicos de cada una de las partes.

**Clusterización.** Aquí se explica una versión general, válida para un

predictor de patogenicidad arbitrario, con  $N$  características de entrada y una predicción de la variante cuya fiabilidad queremos conocer. El procedimiento de clusterización es el siguiente. Primero, normalizamos cada propiedad predictiva de modo que sus valores se comprendan entre 0 y 1. La normalización se obtiene mediante un método min-max:  $(x-x_{\min})/(x_{\max}-x_{\min})$ , donde  $x$  es el valor por normalizar, y  $x_{\min}$  y  $x_{\max}$  son los valores mínimo y máximo, respectivamente, de la propiedad considerada. Segundo, definimos una red  $N$ -dimensional en la que cada eje corresponderá a una propiedad, y habrá once valores por eje: 0, 0.1, 0.2, ..., 0.9 y 1. Todas las combinaciones posibles de estos valores para las diferentes propiedades constituyen los nodos de la red. Tercero, obtenemos la predicción de patogenicidad para cada nodo. Cuarto, hacemos una clusterización de K-means de estos puntos utilizando la implementación Scikit-learn (Pedregosa *et al.* 2011) (`sklearn.cluster.KMeans`) de este algoritmo. Dado el gran número de puntos de la red, hacemos dos ejecuciones consecutivas de K-means. La primera proporciona un conjunto de grandes clústeres que son divididos en la segunda ejecución. Para cada uno de los clústeres finales, calculamos dos medidas,  $rp$  y  $rb$ , que corresponden a la fracción de puntos patogénicos y benignos en el clúster, respectivamente. Se definen como sigue:  $rp = np/(np+nb)$ , donde  $np$  y  $nb$  son el número de predicciones patógenicas y benignas para los puntos del clúster;  $rb$  se define de manera análoga. Hay que señalar que la ejecución de K-means requiere decidir a priori el número de

clústers. Dicho número lo obtuvimos equilibrando la necesidad de dividir finamente el espacio de las propiedades y la de obtener un número de puntos por clúster suficiente para que las estimaciones de  $r_p$  y  $r_b$  sean robustas. Además, exploramos visualmente la distribución de fiabilidades obtenidas de las variantes en nuestras gráficas, para confirmar que existiera congruencia entre estos valores y la distribución de las poblaciones de variantes patogénicas/benignas. Los números de clúster finales fueron 14 y 47 para agrupar los espacios de propiedades correspondientes a RF\_Bioinf y RF\_Metap, respectivamente.

**Asignación de fiabilidad.** Los resultados de la clusterización se pueden utilizar para asignar una fiabilidad R a la predicción de la variante. Dicha asignación se realiza de la siguiente manera. Primero, encontramos el clúster cuyo centro de masas es geoméricamente más cercano a la variante, en el espacio de propiedades discriminantes del predictor. R será entonces igual a la  $r_p$  o  $r_b$  del grupo, dependiendo de si la predicción de la variante era patógena o benigna, respectivamente.

También derivamos un índice de confianza para la adaptación de Feliubadaló *et al.* (Feliubadaló *et al.* 2021) de las guías ACMG/AMP a ATM. Para ello seguimos una variante del enfoque anterior. Primero, no se aplicó el paso de clusterización, debido a que el espacio de las propiedades es bidimensional y no hace falta reducir el número de

dimensiones. Segundo, la red se construyó utilizando un ancho de malla de 0.001, y se utilizaron 120 clusters (valor obtenido tras explorar diferentes valores). La fiabilidad R se calculó de la siguiente manera:  $rp=np/(np+nb+ndis)$ , donde np y nb son el número de predicciones patogénicas y benignas para los puntos en el clúster, y ndis es el número de casos para los cuales los dos predictores fueron discordantes (p. ej., las predicciones de REVEL y VEST4 serían patogénica y benigna, respectivamente, etc). rb se define de manera análoga.

### **5.2.5 Análisis de componentes principales (PCA)**

Se utilizó el PCA como herramienta para la reducción de dimensiones del espacio de las propiedades. Para ello, primero fusionamos las poblaciones de variantes entre patógeno/benigna en un solo conjunto de datos; segundo, rotulamos cada variante con las características de interés (las características utilizadas como input en RF\_Bioinf o RF\_Metap); y, tercero, se hizo un PCA usando la implementación de Scikit-learn (Pedregosa *et al.* 2011) (sklearn.decomposition, importación PCA). Los dos primeros componentes del output se utilizaron en las representaciones gráficas.

## 5.2.6 Representaciones Gráficas

En este trabajo presentamos representaciones gráficas de las distribuciones de variantes patogénicas y benignas en una y dos dimensiones. Estas representaciones se obtuvieron utilizando como materia prima el conjunto de datos de variantes patogénicas y benignas. Tanto para las gráficas uni- como para las bidimensionales utilizamos la función `kdeplot` que pertenece a la suite Seaborn (Waskom 2021). `kdeplot` emplea una estimación de la densidad del kernel para aproximar la función de densidad de probabilidad subyacente a nuestros datos.

## 5.3 Resultados

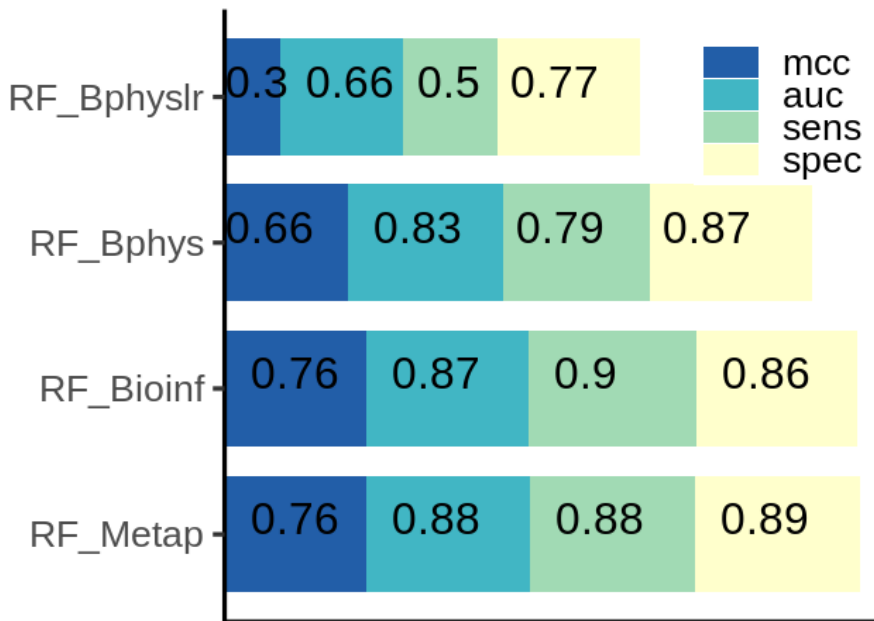
### 5.3.1 Predictores *in silico* para *ATM*

Primero describiremos la capacidad predictiva de los tres predictores específicos de *ATM*, *RF\_Bphys*, *RF\_Bioinf*, y *RF\_Metap*, obtenidos utilizando diferentes características de las variantes. Posteriormente, compararemos su rendimiento con el de otros métodos conocidos en el campo.

### 5.3.1.1 Capacidad predictiva de nuestros predictores RF

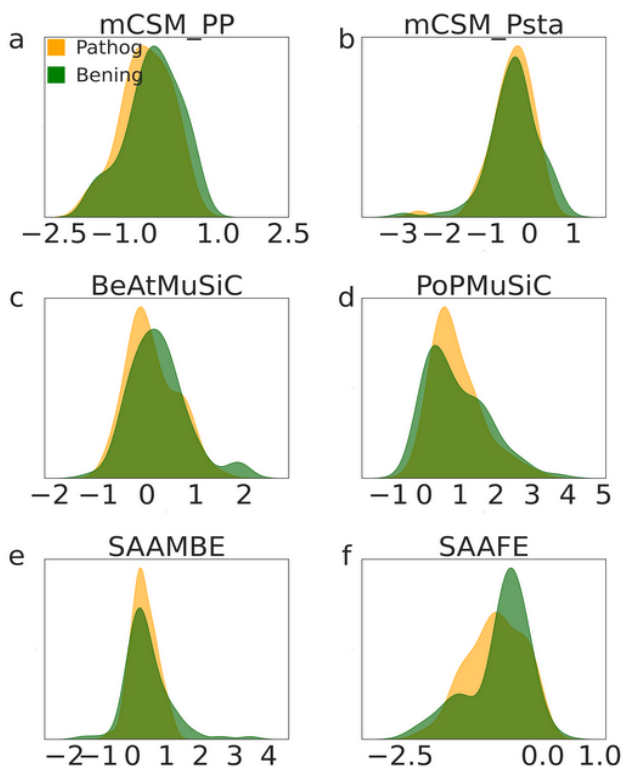
En la Figura 5.1 mostramos la capacidad predictiva de estas herramientas utilizando cuatro medidas (MCC, AUC, Sensibilidad y Especificidad). Vemos que los resultados para RF\_Bioinf y RF\_Metap son similares entre sí y mejores a los de RF\_Bphys y RF\_Bphyslr. El bajo poder predictivo de RF\_Bphyslr es atribuible a dos causas diferentes: la primera es la gran superposición observada para las distribuciones de propiedades biofísicas (Figura 5.2a-f), debida a la pobreza de las estimaciones generadas por los diferentes programas al utilizar la estructura de baja resolución; y (ii) la incoherencia entre estas estimaciones (Figura 5.3). Esta situación cambia si consideramos RF\_Bphys, obtenido con estimaciones de energía basadas en la estructura de ATM de alta resolución publicada a finales de 2021. Utilizando esta estructura calculamos de nuevo los valores de  $\Delta\Delta G$  y  $\Delta\Delta G_{\text{unión}}$  y reentrenamos RF\_Bphys. Esta nueva versión del predictor proporcionó unos resultados claramente mejores (Figura 5.1), mejorando la distribución de las propiedades (Figura 5.4a-f) y presentando mayor coherencia entre las estimaciones energéticas de los diferente algoritmos usados (Figura 5.5). Lo anterior indica que el uso de una estructura de buena resolución es particularmente relevante para estimar  $\Delta\Delta G$  y  $\Delta\Delta G_{\text{unión}}$  y para su uso en la interpretación de variantes genéticas. Sin embargo, estas

estimaciones aún no son lo suficientemente buenas como para superar la capacidad predictiva de las propiedades de conservación (RF\_Bioinf) o las predicciones de herramientas preexistentes (RF\_Metap).

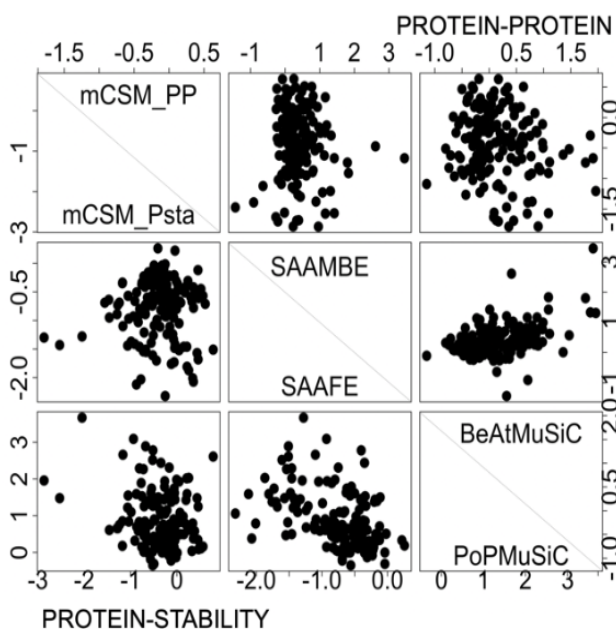


**Figura 5.1 Rendimiento de los predictores desarrollados en este capítulo de la tesis: RF\_Bphyslr, RF\_Bphys, RF\_Bioinf y RF\_Metap.** R\_Bhys se encuentran las dos versiones, usando una estructura de baja resolución RF\_Bphyslr y alta resolución RF\_Bphys. Las barras horizontales muestran, en diferentes colores (ver leyenda), los valores de cuatro parámetros de rendimiento relevantes para el despliegue clínico (sensibilidad, especificidad, AUC y MCC) se muestran en diferentes colores.

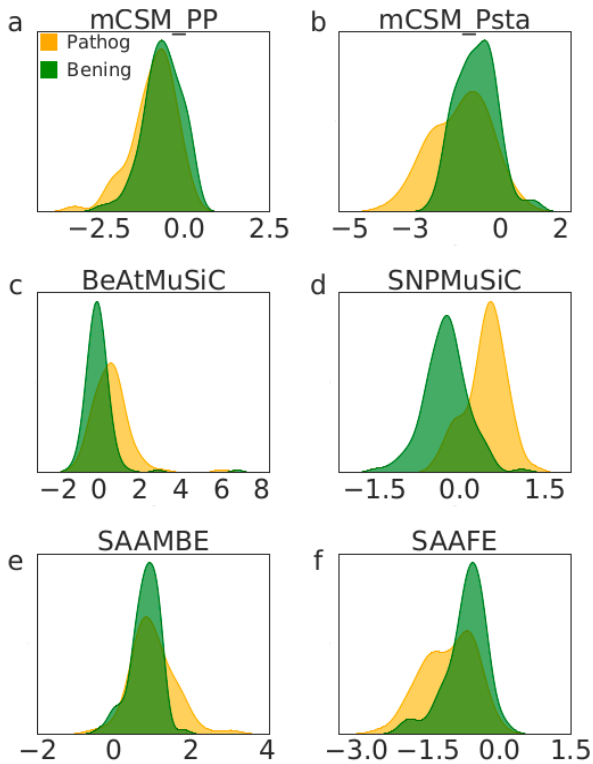




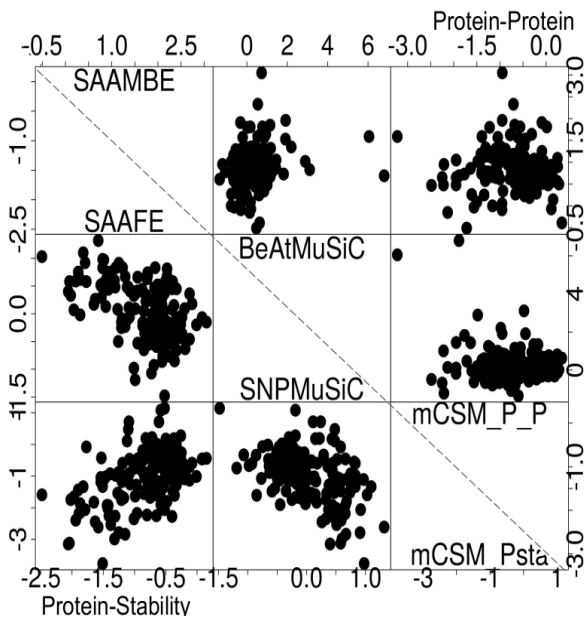
**Figura 5.2. Gráficos unidimensionales para análisis de la capacidad predictiva de RF\_Biophyslr.** (a)-(f) Características de entrenamiento RF\_Bphys. El elevado solapamiento entre las distribuciones de las variantes patogénicas y benignas explica la baja capacidad predictiva de esta herramienta. PDB:5NP0



**Figura 5.3. Consistencia de la estabilidad de la proteína y las energías de unión obtenidas de diferentes herramientas para la construcción de RF\_Biophyslr.** La matriz de diagramas de dispersión muestra la comparación entre los resultados de los programas utilizados en este trabajo para calcular el impacto de las variantes en la estabilidad de las proteínas (triángulo inferior) y las energías de unión (triángulo superior).



**Figura 5.4. Gráficos unidimensionales para análisis de la capacidad predictiva de RF\_Biophys.** (a)-(f) Características de entrenamiento RF\_Bphys. Los resultados obtenidos muestran una mayor capacidad discriminante para las energías derivadas de la nueva versión de la estructura experimental de ATM (PDB: 7SIC).



**Figura 5.3. Consistencia de la estabilidad de la proteína y las energías de unión obtenidas de diferentes herramientas para la construcción de RF\_Biophys.** La matriz de diagramas de dispersión muestra la comparación entre los resultados de los programas utilizados en este trabajo para calcular el impacto de las variantes en la estabilidad de las proteínas (triángulo inferior) y las energías de unión (triángulo superior).

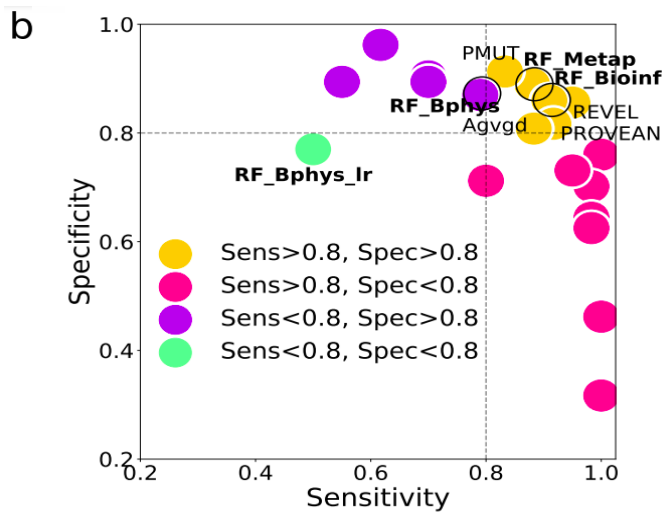
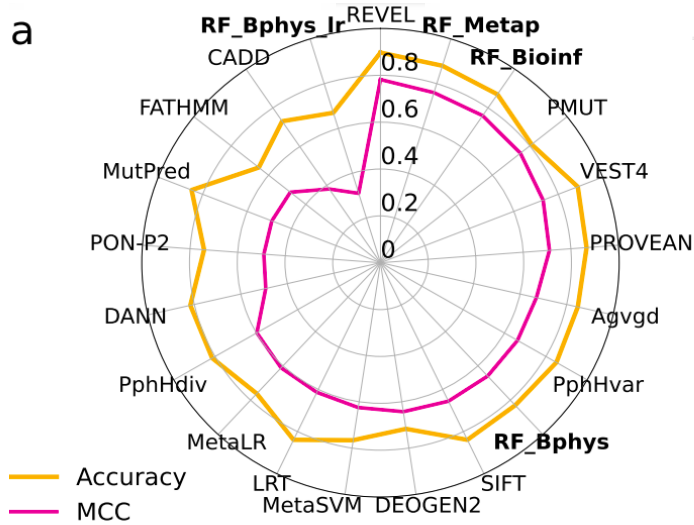
### 5.3.1.2 Comparación de predictores específicos de *ATM* con predictores generales

Para calibrar el valor de las herramientas generadas, comparamos su rendimiento con el de un conjunto de diecisiete predictores de patogenicidad conocidos. Vemos (Figura 5.6) que, independientemente de la medida considerada, RF\_Bioinf y RF\_Metap se sitúan siempre en las primeras posiciones. Por ejemplo, sus MCC son superados solo por REVEL (Figura 5.6a) y sus sensibilidades y especificidades están equilibradas (ambos valores están por encima del 80%) y se encuentran entre las más altas de todos los predictores (Figura 5.6b). RF\_Bphys tiene un rendimiento promedio, en comparación con los predictores generales (Figura 5.6). Sin embargo, hay que destacar que es más equilibrado que muchos predictores (Figura 5.6b) en su capacidad de identificar correctamente variantes benignas y patogénicas.

En resumen, nuestros resultados indican que los nuestros predictores RF\_Bioinf y RF\_Metap, tienen un alto poder de discriminación. Por esta razón, basamos nuestros análisis posteriores en RF\_Bioinf y RF\_Metap.

Finalmente, cabe señalar que el rendimiento obtenido para RF\_Metap y los predictores generales que se muestran en la Figura 5.6 puede ser algo optimista, porque no es posible establecer si parte de las variantes en nuestro conjunto de datos ya se utilizaron en el

entrenamiento de estos métodos. Esta consideración no afecta al rendimiento de RF\_Bioinf en el que hay un control absoluto sobre qué variantes están en el conjunto de entrenamiento y cuales en el conjunto de test.



**Figura 5.6 Comparación de la capacidad predictiva de nuestros tres predictores (RF\_Bphys, RF\_Bioinf y RF\_Metap) con el conjunto de diecisiete predictores publicados.** (a) El diagrama de radar muestra la acc (amarillo) y los valores de MCC (rosa); (b) Diagrama de dispersión que muestra los valores de Especificidad versus Sensibilidad para todos los predictores. Las líneas discontinuas dividen el gráfico en cuatro regiones, de las cuales la de arriba a la derecha corresponde a los predictores que muestran los valores de especificidad y sensibilidad más altos (>80 %) y más equilibrados.

### 5.3.1.3 Análisis de la fiabilidad de las predicciones

Asignamos un índice de fiabilidad (R) a cada predicción generada por RF\_Bioinf y RF\_Metap, siguiendo el procedimiento descrito en Materiales y Métodos (sección 5.2). Con esta información pasamos a verificar hasta qué punto R está relacionada con la idea intuitiva de fiabilidad de acuerdo con la cual una alta fiabilidad corresponde a una mayor probabilidad de que la predicción sea acertada.

Para caracterizar la relación entre la fiabilidad y la precisión de la predicción, utilizamos un valor de umbral específico (0.6) para dividir las predicciones según R. Observamos que para aquellas predicciones con  $R \leq 0.6$ , las precisiones de RF\_Bioinf y RF\_Metap son del 58 % y el 40 %, respectivamente. Por el contrario, para las predicciones con  $R > 0.6$  las precisiones son 98% y 97%, respectivamente. Ello confirma, dentro del conjunto de datos disponible, que R refleja la corrección de las predicciones en el sentido mencionado.

Es interesante señalar que no hay correlación total entre las escalas de fiabilidad de los dos predictores. Para algunas variantes, hay una buena coincidencia, como en el caso de D2987E es para la cual R es igual a 1.0 tanto para RF\_Bioinf como para RF\_Metap. Por el contrario, otras variantes presentan claras diferencias; por ejemplo, para F2827I tenemos 0.91 y 0.40, respectivamente. Este hecho se traduce en un coeficiente de correlación de Pearson de 0.41 (p-valor:  $1.57 \times 10^{-8}$ ) entre las dos escalas de fiabilidad.

### 5.3.2 Análisis gráfico de variantes

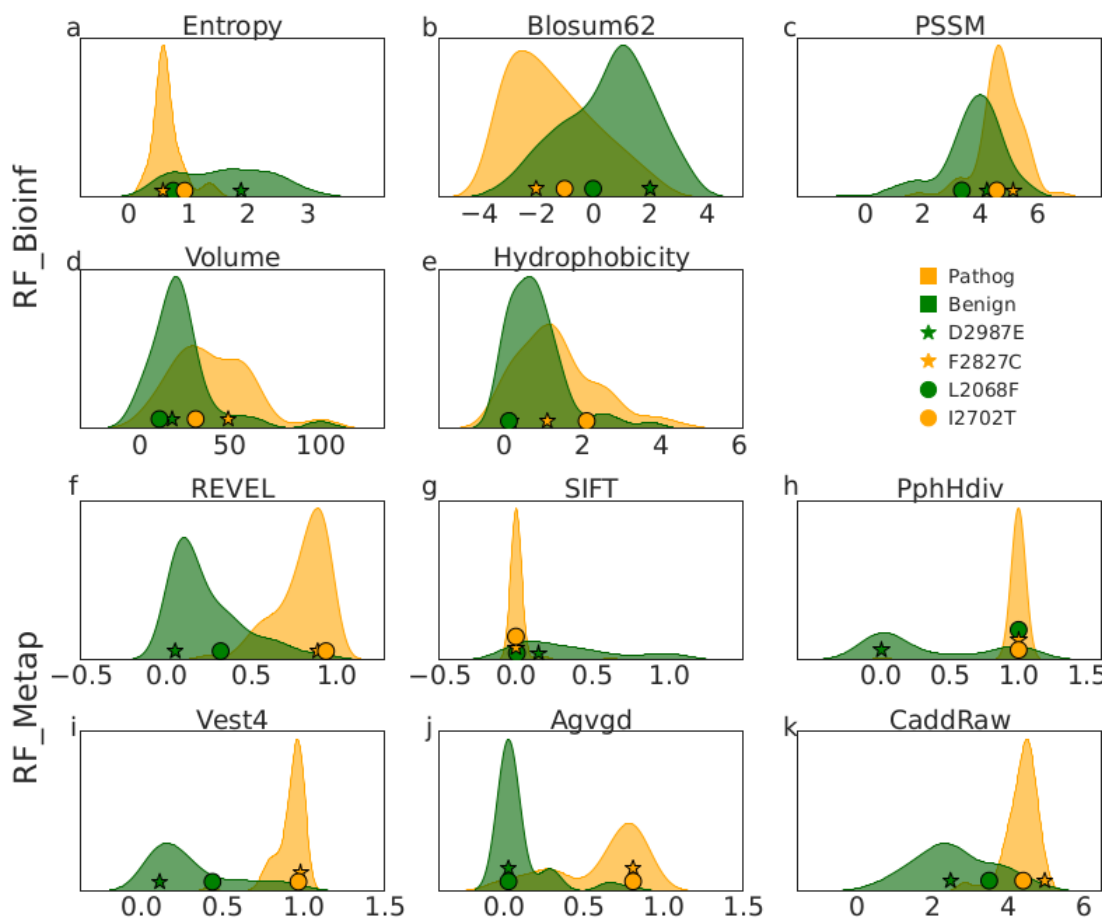
Presentamos una familia de representaciones gráficas para ayudar a la clasificación *in silico* de variantes missense de *ATM*. Estas representaciones nos permiten comparar las propiedades de una variante predicha con las de las variantes benignas y patogénicas conocidas. Entre las diferentes opciones posibles para este tipo de representación, en este trabajo hemos optado por utilizar gráficas de dos tipos: unidimensionales y bidimensionales. Las primeras son versiones suavizadas de los histogramas convencionales, y son fácilmente interpretables en términos de propiedades individuales. Por el contrario, las gráficas bidimensionales están basadas en un análisis PCA y utilizan mapas de contorno para representar las poblaciones de variantes benignas y patogénicas, con lo que permiten un análisis más rico de nuestra predicción. En las secciones siguientes además de presentar las gráficas, describimos su aplicación a tres problemas de la clasificación de variantes. Primero, respaldo de la evaluación de la predicción. Segundo, priorización de variantes de significado desconocido (VUS). Y, tercero, comparación de las dos adaptaciones a *ATM* de las guías ACMG/AMP (Feliubadaló. *et al.*2021; HBOPC VCEP. ClinGen; 2022) en su criterio de la evidencia computacional.

### 5.3.2.1 Graficas Unidimensionales

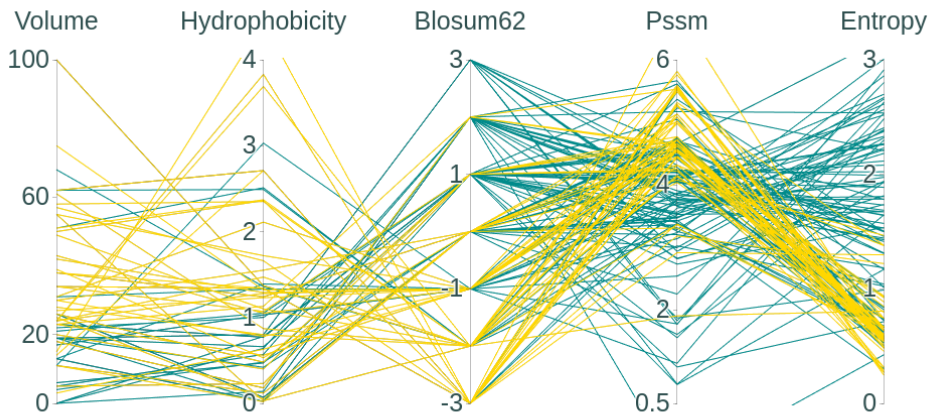
Al observar las representaciones unidimensionales (Figura 5.7), encontramos una coincidencia cualitativa general entre la ubicación de la variante y la fiabilidad numérica de la predicción. Por ejemplo, D2987E, que se predice benigna con alta fiabilidad (RF\_Bioinf: 0.93 y RF\_Metap: 0.78), tiende a ocurrir cerca del pico de la población benigna para la mayoría de las propiedades. Sin embargo, esta tendencia es desigual, siendo más marcada para las variantes benignas que para las patogénicas y para las propiedades de RF\_Bioinf que para las de RF\_Metap.

Un problema de las gráficas unidimensionales es que obligan al usuario a considerar diferentes gráficas a la vez para evaluar la predicción de su variante de interés. Para simplificar esta tarea, exploramos otras visualizaciones, haciendo hincapié en aquellas que permitieran observar simultáneamente todas las características de la variante. Por ejemplo, en la Figura 5.8 observamos uno de estos ensayos en los que se unifican todas las gráficas unidimensionales. En general, tanto para esta como para otras representaciones, llegamos a la conclusión de que su lectura es compleja lo que las hacía inviables para su uso cotidiano en un entorno clínico.





**Figura 5.7 Gráficos unidimensionales para análisis de predicciones.** (a)-(e) Gráficas correspondientes al predictor RF\_Bioinf. (f)-(k) Gráficas correspondientes al predictor RF\_Metap. Las poblaciones de variantes patogénicas y benignas se representan mediante histogramas suavizados en verde y amarillo, respectivamente. Se utilizan símbolos específicos para los cuatro ejemplos seleccionados.



**Figura 5.8 Gráficos unidimensionales relacionando las múltiples características de las variantes.** Representa las 5 características utilizadas en el predictor RF\_Bioinf. (b) representa las características utilizadas en el predictor RF\_Metap. Las poblaciones de variantes patogénicas y benignas se representan mediante líneas en amarillas y verdes, respectivamente.

### 5.3.2.2 Gráficos de contorno como herramientas complementarias para predictores de patogenicidad

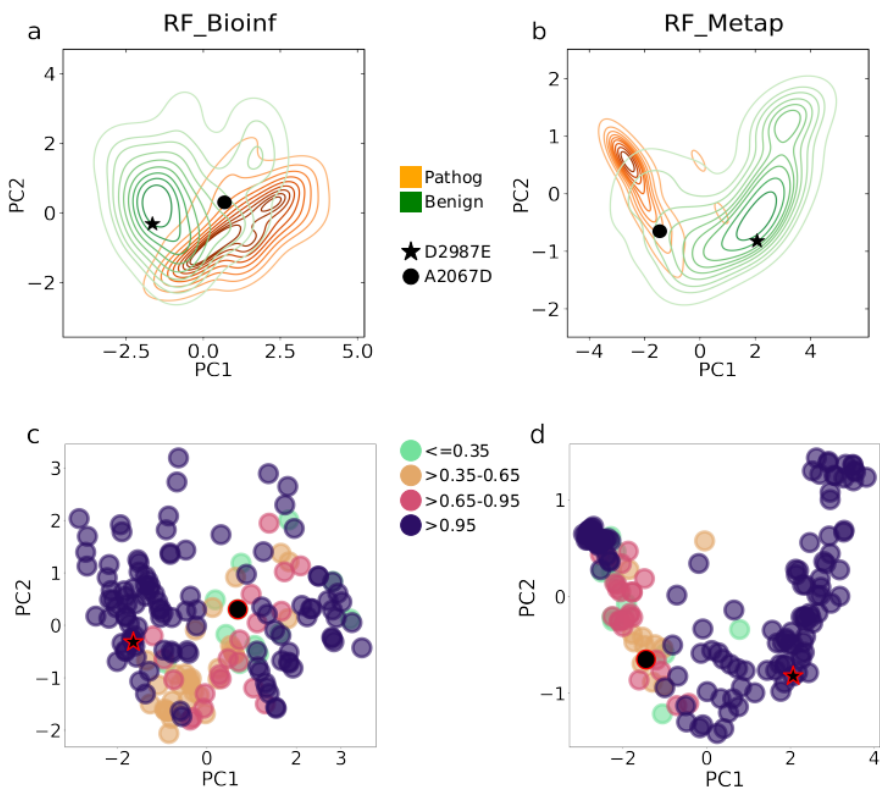
Tras realizar las pruebas que se indican en la sección anterior, decidimos utilizar otra aproximación al problema de representar toda la información en una única figura. Optamos por reducir la dimensionalidad del problema, es decir pasar de cinco o seis

características a dos, lo que permitiría reducir el problema de análisis de la información al estudio de una sola gráfica bidimensional. Para ello utilizamos la técnica de PCA que genera fácilmente proyecciones de la información multidimensional en un espacio bidimensional.

Representamos las poblaciones de variantes patogénicas y benignas utilizando mapas de curvas de nivel bidimensionales (Figuras 5.9). Las curvas de nivel se calculan después del análisis PCA (ver Materiales y métodos específicos). En nuestras gráficas un símbolo específico señala la variante de interés. Para ilustrar la generalidad de nuestra aproximación, mostramos los resultados obtenidos para dos conjuntos diferentes de propiedades: las utilizadas en RF\_Bioinf (Figura 5.9a) y las utilizadas en RF\_Metap (Figura 5.9b). En ambos casos, observamos algunos rasgos característicos: una región en la que se superponen las poblaciones, los máximos de cada población (alrededor de los cuales se concentran las curvas de nivel), y las regiones externas donde solo está presente un tipo de variante.

Al ver la posición relativa de una variante predicha respecto a estas regiones, los usuarios pueden obtener una confirmación visual de la información aportada por el índice de fiabilidad de la predicción. Por ejemplo, RF\_Bioinf predice que la variante benigna D2987E es benigna, con un alto R (1.0); su ubicación cerca del núcleo de la población benigna (Figura 5.9a) apoya la predicción. Por el contrario, A2067D, una variante patogénica predicha como benigna con R baja (0.02), cae en una región donde ambos tipos de variantes están presentes, lo que refuerza las sospechas que suscita la R baja.

Las observaciones anteriores se pueden generalizar representando todas las variantes de nuestro conjunto de datos y coloreándolas de acuerdo con su fiabilidad. Al analizar las figuras resultantes (Figuras 5.9c -5.9d) vemos que existe una coincidencia cualitativa entre las fiabilidades de las predicciones y las características principales de los gráficos de curvas de nivel: (i) las predicciones altamente fiables predominan en las regiones donde se encuentran la mayoría de las variantes patogénicas o benignas conocidas (máximos y lados externos); y (ii) las predicciones menos fiables predominan en la región donde se superponen las curvas de nivel de las dos poblaciones.



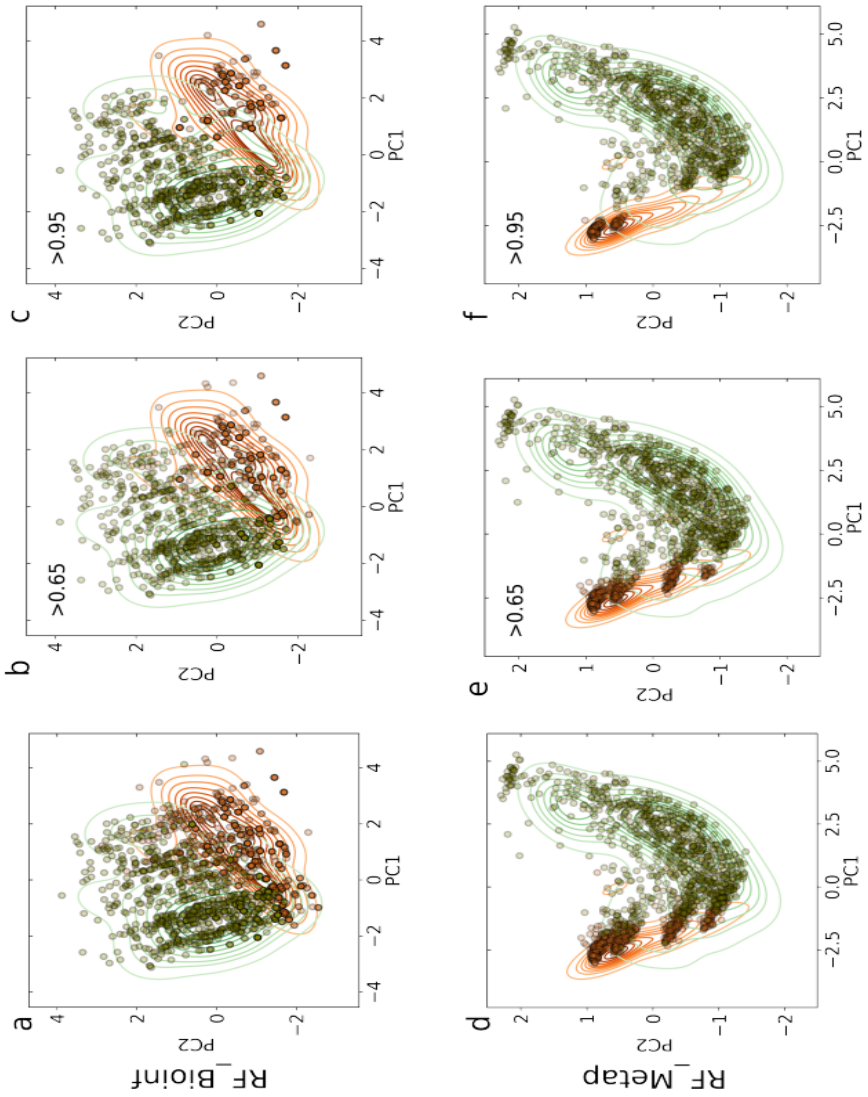
**Figura 5.9 Gráficas bidimensionales para análisis de predicciones.** (a)-(b) Gráficos de contorno que representan poblaciones de variantes patogénicas (verde) y benignas (amarillas). (a) y (b) se obtienen haciendo un PCA de las propiedades utilizadas en RF\_Bioinf y RF\_Metap (ver texto), respectivamente. PC1 y PC2 indican los componentes principales primero y segundo obtenidos en cada PCA. Se utilizan símbolos específicos para los cuatro ejemplos. (c)-(d) Variantes en nuestro conjunto de datos coloreadas según la fiabilidad de su predicción; en estas gráficas, las predicciones se obtienen con las herramientas RF\_Bioinf y RF\_Metap, respectivamente. Los ejes horizontal/vertical son los mismos que en (a) y (b).

### 5.3.2.3 Priorización de VUS

Usamos los gráficos de contorno para clasificar las variantes VUS de *ATM*. Con este fin, recuperamos de la base de datos ClinVar (noviembre de 2021) (HBOPC VCEP. ClinGen; 2022) 1572 VUS missense para este gen. Después de graficarlas en las representaciones asociadas con RF\_Bioinf y RF\_Metap (Figuras 5.10a y 5.10d), vemos que estas variantes se extienden por la mayor parte del gráfico. Algunas de ellas caen en regiones predominantemente patogénicas o benignas, mientras que otras caen en regiones intermedias, lo que indica que la población de VUS es heterogénea.

Las gráficas generadas sugieren la posibilidad de clasificar estas variantes para su estudio posterior para centrarnos, por ejemplo, en los casos benignos o patogénicos más claros. Aquí, mostramos un sencillo procedimiento de priorización basado en la correspondencia observada (Figuras 5.10c – 5.10d) entre las altas fiabilidades de predicción y las regiones predominantemente patogénicas/benignas de los gráficos de contorno. En este procedimiento, primero obtenemos las predicciones RF\_Bioinf para el conjunto de datos VUS; posteriormente eliminamos aquellos casos con índices de fiabilidad por debajo de cierto umbral (probamos 0.65 y 0.95); y, finalmente, representamos las variantes restantes. Las Figuras 5.10b – 5.10c muestran cómo los umbrales de fiabilidad cada vez más estrictos reducen el número de variantes en las regiones superpuestas, dejando principalmente aquellas que pueblan las regiones deseadas.

Se observa un resultado similar con la representación RF\_Metap (Figuras 5.10d – 5.10f). Proporcionamos la lista de variantes seleccionadas en la Tabla anexa 1 y 2.



**Figura 5.10 Aplicación de plots bidimensionales para la priorización de VUS en *ATM*.** Se recuperó un conjunto de 1360 VUS missense de ClinVar. Mostramos su ubicación en los gráficos bidimensionales para las propiedades utilizadas en RF\_Bioinf (a) y RF\_Metap (b). En (b) y (e) eliminamos de estos gráficos las variantes con  $R \geq 0.65$ ; en (c) y (f) aumentamos el umbral de corte a  $R \geq 0.95$ .



#### 5.3.2.4 Comparación de las guías

##### ACMG/AMP adaptadas a ATM

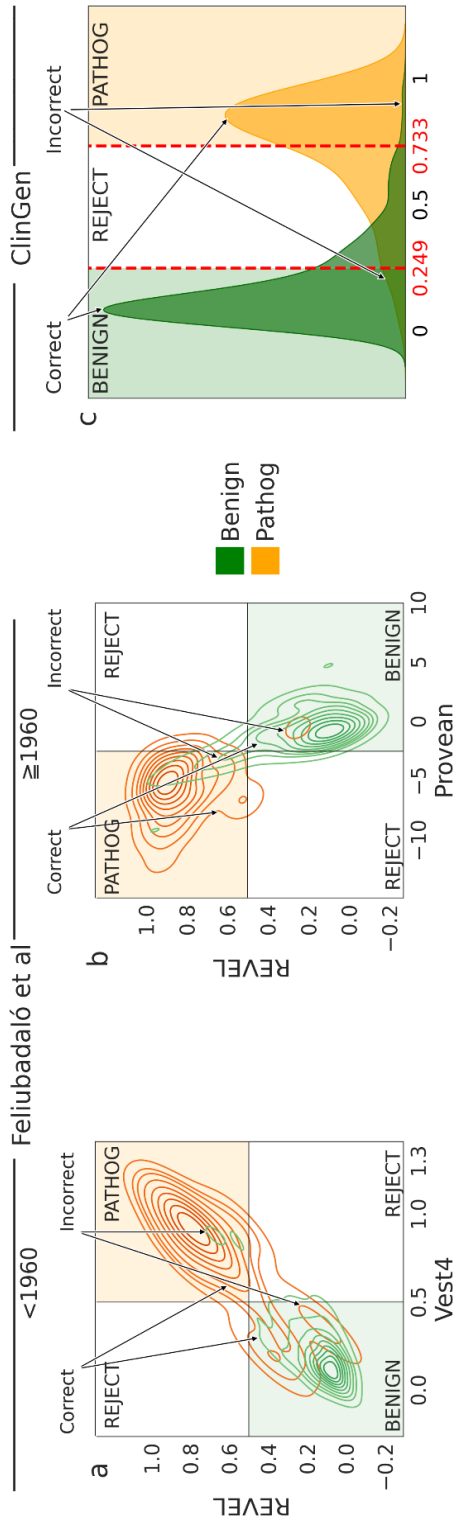
Las dos versiones de las guías adaptadas a ATM (Feliubadaló. *et al.* 2021; HBOPC VCEP. ClinGen; 2022 ) siguen diferentes criterios para aceptar/descartar la evidencia computacional, y puede ser útil para los usuarios comprender estas diferencias. Nuestros gráficos de contorno proporcionan un enfoque intuitivo para su comparación. A continuación describimos su aplicación a las dos versiones de las guías adaptadas.

En la versión de Feliubadaló *et al.* (Feliubadaló. *et al.*, 2021) se aplican dos predictores a las variantes: si sus resultados discrepan, se rechaza la evidencia computacional, y en caso contrario se acepta. En las Figuras 5.11a y 5.11b aplicamos nuestra representación a este criterio. Tres aspectos deben considerarse para su análisis. Primero, las curvas de nivel que atraviesan una región de aceptación de su mismo color corresponden a variantes que se clasificaron correctamente. En segundo lugar, las curvas de nivel que atraviesan una región de aceptación de un color diferente (por ejemplo, cuando las curvas verdes atraviesan la región naranja) corresponderán a variantes que se clasificarán incorrectamente. Y tercero, las curvas de nivel que atraviesan la región de rechazo corresponden a variantes cuyas predicciones no se tendrán en cuenta.

Consideremos ahora la adaptación de las guías ACMG/AMP realizada por el grupo de expertos de ClinGen (HBOPC VCEP. ClinGen; 2022). A

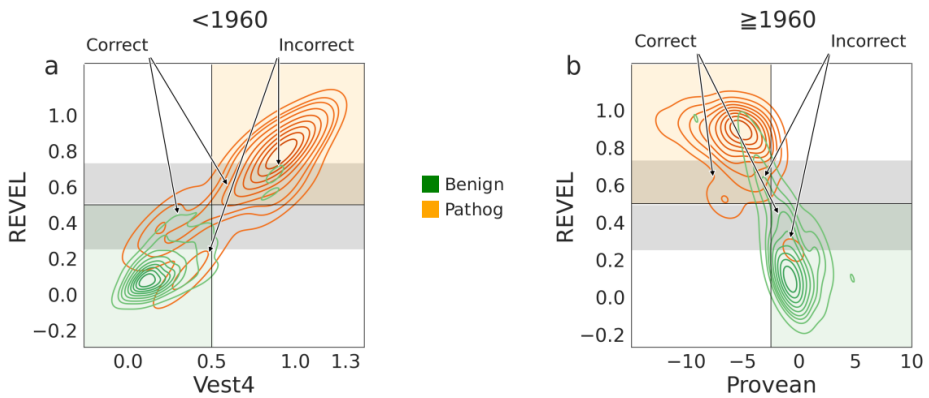
diferencia del caso anterior, estos autores sólo utilizan un predictor, REVEL, para la predicción de patogenicidad de las variantes y definen, como región de rechazo, los valores de REVEL comprendidos entre 0.249 y 0.733 (Figura 5.11c).

En la Figura 5.12 comparamos las dos guías adaptadas. La banda central gris corresponde a la región de rechazo de ClinGen. Esta región tiene un efecto principal respecto al criterio de Feliubadalo *et al.*: disminuye el número de variantes anotadas debido a una reducción en la región de aceptación. Este efecto tiene una consecuencia positiva: los errores de predicción disminuirán al aplicar el criterio de ClinGen. Sin embargo, también tiene un aspecto negativo: se rechazará una cantidad sustancial de predicciones correctas (esto afecta principalmente a las variantes patogénicas en posiciones anteriores a la 1960 -Figura 5.12a, o a las benignas en posiciones posteriores a la 1960 -Figura 5.12b). Otro efecto del criterio de ClinGen es la aceptación de una pequeña cantidad de predicciones, en su mayoría correctas, descartadas por los criterios de Feliubadalo *et al.*



**Figura 5.11 Representación visual de los criterios para la evidencia computacional en las guías ACMG/AMP adaptadas a ATM (Feliubadaló et al.)**

Estas guías utilizan dos predictores cuya contribución se acepta cuando las predicciones coinciden y se rechaza en caso contrario. Los predictores combinados varían según la ubicación de la variante: REVEL+VEST4 ( $<1960$ ) y REVEL+PROVEAN ( $\geq 1960$ ). La combinación de áreas coloreadas y líneas de contorno en (a) y (b) muestra cómo funcionan estos criterios y su precisión. Los cuadrados coloreados y blancos, respectivamente, corresponden a las regiones de aceptación y rechazo de los criterios. Las curvas de nivel que atraviesan un cuadrado del mismo color indican variantes que se anotarán correctamente; si los colores son opuestos, la anotación será incorrecta. Las flechas indican estas dos situaciones. Finalmente, las líneas de contorno que atraviesan un cuadrado blanco indican variantes para las cuales se rechazará la evidencia computacional. En (c), aplicamos el mismo esquema a las pautas adaptadas a ATM de ClinGen, que utilizan un solo predictor, REVEL, para la anotación *in silico* de variantes. Aquí, un solo histograma es suficiente para ilustrar el criterio de la guía. Las regiones de aceptación/rechazo están coloreadas como antes, e indicamos con flechas los posibles éxitos y errores de anotación de las regiones de aceptación.



**Figura 5.12 Comparación de las dos adaptaciones a ATM de las guías ACMG/AMP.** (a) y (b) Representan ambas versiones de las guías para ubicaciones de las variantes por debajo y por encima de 1960. Sobre los gráficos de contorno correspondientes a los criterios de Feliubadalo *et al.*, mostramos una banda (gris) correspondiente a la región de rechazo de ClinGen. Pueden ocurrir las siguientes situaciones. Cuando la banda se superpone con una región de aceptación coloreada, el criterio de ClinGen no producirá anotaciones para las variantes del interior. En consecuencia, la aplicación del criterio de ClinGen relativo al criterio de Feliubadalo *et al.* da como resultado la eliminación de algunos errores potenciales y la pérdida de un número sustancial de anotaciones correctas (ambos casos identificados con flechas). Cuando la banda gris se superpone a un cuadrado blanco, ambos criterios coinciden en la exclusión de la evidencia computacional. Finalmente, fuera de la banda gris, las curvas de nivel que atraviesan cuadrados blancos corresponden a anotaciones que serán rechazadas por el criterio de Feliubadalo *et al.* pero aceptadas por el criterio de ClinGen.

## 5.4 Discusión

En este trabajo, exploramos dos propuestas para mejorar la clasificación clínica *in silico* de variantes missense de *ATM* basadas en mejorar la interpretabilidad de los resultados computacionales. Una mayor interpretabilidad de estos resultados es deseable porque permite a los profesionales juzgar mejor los resultados computacionales y, por consiguiente, facilita la identificación de errores de predicción (Rudin; 2019). Esto es importante en un contexto como el de la clasificación de variantes de *ATM*, donde la precisión de las herramientas *in silico* está por debajo del 100 % y, por ello, se producen errores (Figura 1.17).

En nuestra primera propuesta, hemos desarrollado tres predictores de patogenicidad específicos de *ATM* en los que las características de entrenamiento cubren el rango de interpretabilidad: alto (RF\_Bphys), promedio (RF\_Bioinf) y bajo (RF\_Metap). Dos de estas herramientas (RF\_Bioinf y RF\_Metap) son competitivas en relación a los métodos conocidos (Figura 5.6), mientras que RF\_Bphys, la más interpretable, tiene un rendimiento promedio. Sin embargo, las diferencias no son demasiado grandes (ver MCC y precisión en la Figura 5.6a), y RF\_Bphys tiene un balance de sensibilidad/especificidad más equilibrado que la mayoría de las herramientas convencionales (Figura 5.6b). Ello indica que, con precaución, las estimaciones de energía se pueden utilizar para evaluar la anotación *in silico* de

variantes missense en *ATM* (para ubicaciones  $\geq 1960$ ). Esta aplicación podría además mejorar a medida que se disponga de estructuras *ATM* de mayor resolución (como se ve al comparar RF\_Bphys y RF\_Bphys\_lr). Una consecuencia adicional de estos resultados es que abren el camino para complementar las descripciones binarias (patogénicas/benignas) de las variantes con estimaciones continuas y más realistas de su impacto, en la línea de trabajos recientes en el campo de la predicción (Masica, 2016; Gray, 2018).

La segunda propuesta que presentamos para mejorar la clasificación de variantes *in silico*, se basa en atacar el problema desde un ángulo diferente, reemplazando las métricas numéricas (puntuaciones de predicción/fiabilidades) con herramientas gráficas que promuevan y faciliten el papel de los profesionales clínicos. Este enfoque está en línea con la visión cada vez más extendida (AI-HLEG-group, 2019 & Rudin, 2019) de acuerdo con la cual los expertos humanos deben poder supervisar las aplicaciones de machine learning en campos de alto riesgo. Nuestros gráficos favorecen este objetivo al proporcionar representaciones fáciles de interpretar de la variante objetivo en el contexto de poblaciones de variantes patogénicas/benignas conocidas. En las gráficas podemos observar las características de las variantes y relacionarlas con el valor de fiabilidad de la predicción (Figuras 5.9c-d). Ello apoya el uso de nuestras representaciones para evaluar predicciones visualmente, ampliando así la información proporcionada por los índices de fiabilidad estándar (Lopez-Ferrando, 2017 & Ferrer-Costa, 2004).

La naturaleza general de nuestras representaciones nos permite usarlas para fines distintos a la evaluación de predicciones. Por ejemplo, para ayudar a anotar la gran población de VUS del gen *ATM*. El número de estas variantes en genes de interés clínico es un cuello de botella en el desarrollo de la medicina genómica (Shendure, 2019) y, con 1572 VUS (ClinGen, noviembre de 2021), *ATM* no es una excepción (Feliubadalo *et al.*, 2021). En este contexto, las estrategias para comprender estas variantes y reducir su número son una prioridad. Aquí, exploramos la posibilidad de usar nuestras representaciones para producir una comprensión y clasificación rápida de VUS del gen *ATM* en relación con su impacto molecular. Encontramos (Figuras 5.10a-d) que la población VUS cubre todo el rango patógeno-benigno. Para ayudar a seleccionar los casos más claros para estudios posteriores, definimos regiones de aceptación/rechazo en función del índice de fiabilidad de las predicciones de patogenicidad de las variantes. En las Figuras 5.10b-c y 5.10e-f (ver también Tablas anexas 1 y 2) se ve cómo esta sencilla estrategia permite la selección de variantes en las regiones deseadas.

En una tercera aplicación de nuestras representaciones, mostramos cómo los gráficos de curvas de nivel permiten comparar los criterios de uso de la evidencia computacional en las guías adaptadas a *ATM* (Feliubadaló. *et al.*2021; HBOPC VCEP. ClinGen; 2022). Encontramos que los criterios de Clingen tienen una tasa de clasificación errónea menor, aunque al precio de rechazar un número sustancial de

predicciones correctas. En cambio, ocurre lo contrario con los criterios de Feliubadalo *et al.* A priori no se puede establecer qué criterios son preferibles; la respuesta probablemente dependerá del contexto de la aplicación. Por ejemplo, si la evidencia disponible sobre una variante es sólida, parece preferible aplicar los criterios de Clingen, ya que el posible rechazo de la evidencia computacional tendrá un impacto bajo. Sin embargo, si la evidencia disponible es escasa o débil, el uso de los criterios más inclusivos de Feliubadalo *et al.* puede dar pistas útiles sobre la naturaleza de la variante.

Las herramientas presentadas en este trabajo se han desarrollado utilizando los conjuntos de variantes descritos en Feliubadalo *et al.* A medida que aumente el número de variantes disponibles en la literatura las iremos actualizando, lo cual puede aumentar el rendimiento de los predictores de patogenicidad. Sin embargo, esperamos un impacto menor en los gráficos de curvas de nivel, que reflejan el comportamiento general de las poblaciones de variantes. En general, no creemos que un mayor número de variantes redunde en cambios sustanciales en las principales conclusiones de nuestro trabajo, como son el valor de usar estimaciones de energía para validar anotaciones o la aplicabilidad de los gráficos de curvas de nivel a problemas de anotación.

Finalmente, nos gustaría mencionar que nuestro trabajo se centra en *ATM*. Sin embargo, debido a su simplicidad, nuestra metodología también puede extenderse a otros genes de interés. El único requisito es que haya suficientes variantes clasificadas para producir los



gráficos de curvas de nivel y los predictores asociados.

## 6 DISCUSIÓN GENERAL

En el contexto clínico, la correcta clasificación de las variantes tiene un impacto directo sobre la atención médica que recibirán sus portadores (Merino Bonilla JA. *et al.* 2017). En el caso particular del cáncer de mama, el determinar si una variante del gen *ATM* que se encuentra en una paciente es patogénica, permitirá buscar el tipo de tratamiento más acorde con la genética de la portadora, garantizando así una mayor esperanza de vida (Bernstein *et al.*, 2010).

La contribución de esta tesis se dirige a mejorar el proceso de clasificación de variantes centrándonos en la contribución de los métodos de clasificación *in silico* actuales. Para ello nos hemos centrado en dos aspectos complementarios: (i) un mejor uso de las herramientas pre-existentes, y (ii) desarrollo de nuevas herramientas que facilitan la evaluación humana de las predicciones *in silico*.

El esfuerzo destinado a mejorar el uso de las herramientas ya existentes se sitúa en el contexto de la adaptación de las guías ACMG/AMP al gen *ATM* de Feliubadalo *et al.* (Feliubadaló. *et al.*, 2021) , un trabajo que ha involucrado al 'Consortio español para la interpretación de las variantes genéticas en el gen *ATM*'. Nuestro principal aporte fue el análisis de cada mitad de la proteína (N y C-terminal) por separado, implementando una combinación de predictores para cada segmento. Con nuestro esfuerzo se ha obtenido una capacidad predictiva superior al uso de predictores individuales convencionales, permitiendo que la clasificación sea más confiable. (Figura 4.6 y 4.7 )

Este método de clasificación *in silico* sugerido para las guías adaptadas, a pesar de su buen rendimiento, presenta limitaciones por la falta de interpretabilidad. Además, también puede ocurrir que los valores de las métricas obtenidas estén sobreestimados, ya que no es posible determinar si nuestras variantes o parte de ellas fueron usadas durante el proceso de entrenamiento de los predictores que fueron seleccionados. Este problema, es conocido como problema de circularidad (Grimm *et al.*, 2015), y es relativamente frecuente en el desarrollo de herramientas *in silico*. Por ello decidimos desarrollar nuestra propia tecnología, basada en nuestro trabajo anterior en predictores específicos (Riera *et al.*, 2016) y que se hiciese eco, además, de las nuevas tendencias en interpretación de herramientas de inteligencia artificial (Rudin, 2019).

En la dirección de mejorar los predictores *in silico* para ATM, obtuvimos 3 predictores específicos: RF\_Biophys, RF\_Bioinf y RF\_Metap. El enfoque que primó en el desarrollo de estos predictores fue la interpretabilidad de sus resultados, combinada con una capacidad predictiva competitiva, acorde con las nuevas tendencias para el desarrollo de los predictores en el uso clínico (Petch *et al.*, 2022). Un aspecto interesante de los resultados obtenidos (Figura 5.6), además de la capacidad predictiva obtenida, es que confirman la

conjetura de Rudin (Rudin, 2019) de acuerdo con la cual predictores de baja complejidad pueden tener capacidades predictivas comparables a las de predictores 'black box'. Aún así, el uso de nuestros predictores entrañaba la aplicación de tecnología Random Forest, que no es fácilmente comprensible (Musolf *et al.* 2022). Por ello decidimos desarrollar una familia de representaciones gráficas que complementan a nuestros predictores y facilitan su uso por parte de los profesionales de la salud.

Obtuvimos nuestras representaciones visuales a partir de la distribución de las poblaciones de variantes patogénicas y benignas conocidas, respecto a las cuales ubicamos nuestra variable de interés (Figura 5.9). La sencillez de las herramientas desarrolladas permite su aplicación a diferentes problemas en la anotación de variantes de *ATM*. Entre ellos destaca la priorización de variantes para su posterior estudio funcional, un problema relevante debido a la gran cantidad de variantes VUS para este gen (Federici G. *et al.*, 2020). Aquí presentamos un sencillo protocolo de priorización, que ilustramos con la totalidad de las variantes VUS que se encuentran en la plataforma de ClinGen (anexo tabla 1 y 2) (Figura 5.10). Los resultados obtenidos contribuyen a estratificar la población de estas variantes, identificando aquellas que muestran un carácter más definido y para

las que, por lo tanto, es más fácil de establecer su naturaleza. Lo anterior es una necesidad porque permite optimizar los recursos de investigación y mejora en la atención a cáncer de mama principalmente en países con menos recursos económicos (Wojtyla C. *et al.*, 2017).

También realizamos un comparativo de las estrategias de clasificación *in silico* de las adaptaciones para las guías ACMG/AMP a ATM. Con ayuda de las herramientas visuales aquí desarrolladas, revisamos las ventajas y desventajas de los métodos usados. Encontramos que el método propuesto por nosotros puede ser ventajoso ya que al dividir la proteína se obtiene una mejor resolución de la distribución de las variantes para cada mitad, efecto que no se puede observar con el uso de una misma metodología para todas las variantes.

Finalmente, en este trabajo hemos avanzado hacia una mejor interpretabilidad de los predictores de patogenicidad de manera general, ya que las herramientas visuales desarrolladas pueden ser transversales siendo adaptables a otros genes. Nuestras herramientas presentan una capacidad predictiva competente, y entre ellas destacaremos el predictor RF\_Bioinf. Este predictor no presenta problemas de circularidad, siendo un predictor completamente *de*

*novo* y construido con características tanto evolutivas como biofísicas. Todo ello lo hace un sistema sencillo e interpretable, con una capacidad de predictiva alta. Así nuestros resultados contribuyen a la comunidad interesada en *ATM* para la anotación de variantes de este gen.

## 7 CONCLUSIONES



### **Son conclusiones de esta tesis:**

1. Para el desarrollo de un predictor específico de *ATM* la mejor estrategia es considerar por separado las dos mitades N-terminal y C-terminal de la proteína, ya que es una proteína grande y con dominios funcionales muy diferentes entre estas dos mitades.
2. Para la clasificación de variantes de *ATM* según las guías adaptadas por Feliubadaló *et al.* (Feliubadaló. *et al.*2021) se estableció un sistema de dos predictores en cada mitad de la proteína: REVEL+ VEST4 Y REVEL+ PROVEAN .
3. La construcción de predictores con características evolutivas han mostrado ser un sistema muy eficiente en la clasificación de variantes *ATM*. También, el desarrollo de un metapredictor ha resultado una buena estrategia de clasificación para *ATM*, aunque puede ser un sistema potencialmente sobrevalorado.
4. Gracias a las estructuras de alta resolución se han podido estimar con bastante precisión los cambios de energía libre asociados a las variantes missense. Ello ha permitido desarrollar un predictor con una buena capacidad de acierto.
5. Asociado a los predictores presentados en este trabajo, se ha desarrollado un sistema visual para la interpretación de los resultados obtenidos por los predictores.

6. Nuestro sistema de representaciones gráficas presenta unos rasgos principales asociados a la fiabilidad de las predicciones.
7. Se ha aplicado nuestras gráficas a la priorización de variantes VUS para futuros estudios primarios funcionales.
8. Se ha comparado el tratamiento de la evidencia computacional en las dos versiones de las guías ACMG/AMP adaptadas a *ATM* utilizando nuestras representaciones gráficas. Ello ha permitido identificar las regiones complementarias en el tratamiento de la evidencia computacional entre estas adaptaciones.
9. El sistema de interpretabilidad desarrollado en este trabajo es transversal y, por tal razón, puede ser adaptado a otros predictores u otros genes.

## 8 REFERENCIAS

Abdel-Razeq H, Tamimi F, Abujamous L, Abdel-Razeq R, Abunasser M, Edaily S, Abdulelah H, Khashabeh RA, Bater R. Rates of Variants of Uncertain Significance Among Patients With Breast Cancer Undergoing Genetic Testing: Regional Perspectives. *Front Oncol*. 2022 Mar 25;12:673094. doi: 10.3389/fonc.2022.673094. PMID: 35402282; PMCID: PMC8989924.

AI-HLEG-group. Ethics Guidelines for Trustworthy AI. European Commission. Eur. Comm. 2019

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., ... Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4), 248–249. <https://doi.org/10.1038/nmeth0410-248>

Auton A, Abecasis GR, Altshuler DM, et al (2015) A global reference for human genetic variation. *Nature* 526:68–74. <https://doi.org/10.1038/nature15393>

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5), 412–424. <https://doi.org/10.1093/bioinformatics/16.5.412>

Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., ... Zhang, J. (2017a). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, 45(D1), D158–D169. <https://doi.org/10.1093/nar/gkw1099>

Bakkenist CJ, Kastan MB. DNA damage activates *ATM* through intermolecular autophosphorylation and dimer dissociation. *Nature*. 2003 Jan 30;421(6922):499-506. doi: 10.1038/nature01368. PMID: 12556884.

Bernstein JL; WECARE Study Collaborative Group, Concannon P. *ATM*, radiation, and the risk of second primary breast cancer. *Int J Radiat Biol*. 2017 Oct;93(10):1121-1127. doi: 10.1080/09553002.2017.1344363. Epub 2017 Jul 27. PMID:

28627265; PMCID: PMC6113688.

Bernstein JL, Haile RW, Stovall M, Boice JD Jr, Shore RE, Langholz B, Thomas DC, Bernstein L, Lynch CF, Olsen JH, Malone KE, Mellemkjaer L, Borresen-Dale AL, Rosenstein BS, Teraoka SN, Diep AT, Smith SA, Capanu M, Reiner AS, Liang X, Gatti RA, Concannon P; WECARE Study Collaborative Group. Radiation exposure, the *ATM* Gene, and contralateral breast cancer in the women's environmental cancer and radiation epidemiology study. *J Natl Cancer Inst.* 2010 Apr 7;102(7):475-83. doi: 10.1093/jnci/djq055. Epub 2010 Mar 19. PMID: 20305132; PMCID: PMC2902825.

Breiman, L. Random Forests. *Machine Learning* 45, 5–32 (2001). <https://doi.org/10.1023/A:1010933404324>

Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics.* 2013;14 Suppl 3.

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining*, 10(1), 1–17. <https://doi.org/10.1186/s13040-017-0155-3>

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino Acid substitutions and indels. *PLoS One.* 2012;7:e46688. 18.

Christopher Bishop, *Pattern Recognition and Machine Learning* , 2006.

Chun S, Fay JC (2009) Identification of deleterious mutations within three human genomes. *Genome Res* 19:1553–1561. <https://doi.org/10.1101/gr.092619.109>

Couch FJ, Shimelis H, Hu C, Hart SN, Polley EC, Na J, Hallberg E, Moore R, Thomas A, Lilyquist J, Feng B, McFarland R, Pesaran T, Huether R, LaDuca H, Chao EC, Goldgar DE, Dolinsky JS. Associations Between Cancer Predisposition Testing Panel Genes and Breast

Cancer. *JAMA Oncol.* 2017 Sep 1;3(9):1190-1196. doi: 10.1001/jamaoncol.2017.0424. PMID: 28418444; PMCID: PMC5599323.

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory*. USA: Wiley-Interscience.

Currie G, Hawk KE, Rohren E, Vial A, Klein R. Machine Learning and Deep Learning in Medical Imaging: Intelligent Imaging. *J Med Imaging Radiat Sci.* 2019 Dec;50(4):477-487. doi: 10.1016/j.jmir.2019.09.005. Epub 2019 Oct 7. PMID: 31601480.

Dong C, Wei P, Jian X, et al (2015) Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet* 24:2125–2137. <https://doi.org/10.1093/hmg/ddu733>

Dehouck Y, Kwasigroch JM, Gilis D, Rooman M (2011) PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics* 12:151

Dehouck Y, Kwasigroch JM, Rooman M, Gilis D (2013) BeATMuSiC: Prediction of changes in protein-protein binding affinity on mutations. *Nucleic Acids Res* 41:W333–W339. <https://doi.org/10.1093/nar/gkt450>

Ernst, C., Hahnen, E., Engel, C., Nothnagel, M., Weber, J., Schmutzler, R. K., & Hauke, J. (2018). Performance of *in silico* prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics. *BMC Medical Genomics*, 11(1), 1–10. <https://doi.org/10.1186/s12920-018-0353-y>

Erfani P, Bhangdia K, Stauber C, Mugunga JC, Pace LE, Fadelu T. Economic Evaluations of Breast Cancer Care in Low- and Middle-Income Countries: A Scoping Review. *Oncologist.* 2021 Aug;26(8):e1406-e1417. doi: 10.1002/onco.13841. Epub 2021 Jun 5. PMID: 34050590; PMCID: PMC8342576.

Fauchere J, Pliska V (1983) Hydrophobic parameters of amino acid side-chains from the partitioning of N-acetyl-amino-acid amides.

- Federici G, Soddu S. Variants of uncertain significance in the era of high-throughput genome sequencing: a lesson from breast and ovary cancers. *J Exp Clin Cancer Res*. 2020 Mar 4;39(1):46. doi: 10.1186/s13046-020-01554-6. PMID: 32127026; PMCID: PMC7055088.
- Feliubadaló L, Moles-Fernández A, Santamariña-Pena M, Sánchez AT, López-Novo A, Porras LM, Blanco A, Capellá G, de la Hoya M, Molina IJ, Osorio A, Pineda M, Rueda D, de la Cruz X, Diez O, Ruiz-Ponte C, Gutiérrez-Enríquez S, Vega A, Lázaro C. A Collaborative Effort to Define Classification Criteria for *ATM* Variants in Hereditary Cancer Patients. *Clin Chem*. 2021 Mar 1;67(3):518-533. doi: 10.1093/clinchem/hvaa250. PMID: 33280026.
- Ferrer-Costa, C., Orozco, M., & De La Cruz, X. (2004). Sequence-based prediction of pathological mutations. *Proteins: Structure, Function and Genetics*, 57(4), 811–819. <https://doi.org/10.1002/prot.20252>
- Filippini SE, Vega A. Breast cancer genes: beyond BRCA1 and BRCA2. *Front Biosci (Landmark Ed)*. 2013 Jun 1;18(4):1358-72. doi: 10.2741/4185. PMID: 23747889.
- Fortuno C, Lee K, Olivier M, Pesaran T, Mai PL, de Andrade KC, Attardi LD, Crowley S, Evans DG, Feng BJ, Foreman AKM, Frone MN, Huether R, James PA, McGoldrick K, Mester J, Seifert BA, Slavin TP, Witkowski L, Zhang L, Plon SE, Spurdle AB, Savage SA; ClinGen TP53 Variant Curation Expert Panel. Specifications of the ACMG/AMP variant interpretation guidelines for germline TP53 variants. *Hum Mutat*. 2021 Mar;42(3):223-236. doi: 10.1002/humu.24152. Epub 2020 Dec 25. PMID: 33300245; PMCID: PMC8374922.
- Gray VE, Hause RJ, Luebeck J, Shendure J, Fowler DM. Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis

Data. Cell Syst. 2018;6:116–24.

Grimm DG, Azencott CA, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, Cooper DN, Stenson PD, Daly MJ, Smoller JW, Duncan LE, Borgwardt KM. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum Mutat.* 2015 May;36(5):513-23. doi: 10.1002/humu.22768. Epub 2015 Mar 26. PMID: 25684150; PMCID: PMC4409520.

González-Santiago S, Ramón y Cajal T, Aguirre E, Alés-Martínez JE, Andrés R, Balmaña J, *et al.* SEOM clinical guidelines in hereditary breast and ovarian cancer (2019). *Clin Transl Oncol.* 2020;22:193–200

Hannan Z, Yu S, Mamtani R, Reiss KA. Clinical Characteristics of Patients with Pancreatic Cancer and Pathogenic *ATM* Alterations. *JNCI Cancer Spectr.* 2021;5:pkaa121.

Hall MJ, Bernhisel R, Hughes E, Larson K, Rosenthal ET, Singh NA, Lancaster JM, Kurian AW. Germline Pathogenic Variants in the Ataxia Telangiectasia Mutated (*ATM*) Gene are Associated with High and Moderate Risks for Multiple Cancers. *Cancer Prev Res (Phila).* 2021 Apr;14(4):433-440. doi: 10.1158/1940-6207.CAPR-20-0448. Epub 2021 Jan 28. PMID: 33509806; PMCID: PMC8026745.

Hart SN, Hoskin T, Shimelis H, Moore RM, Feng B, Thomas A, Lindor NM, Polley EC, Goldgar DE, Iversen E, Monteiro ANA, Suman VJ, Couch FJ. Comprehensive annotation of BRCA1 and BRCA2 missense variants by functionally validated sequence-based computational prediction models. *Genet Med.* 2019 Jan;21(1):71-80. doi: 10.1038/s41436-018-0018-4. Epub 2018 Jun 8. PMID: 29884841; PMCID: PMC6287763.

HBOPC VCEP. ClinGen Hereditary Breast, Ovarian and Pancreatic Cancer Expert Panel Specifications to the ACMG/AMP Variant Interpretation Guidelines for *ATM* Version 1.1. 2022.



- Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22), 10915–10919. Retrieved from [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list\\_uids=1438297](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=1438297)
- Hu C, Hart SN, Gnanaolivu R, Huang H, Lee KY, Na J, *et al.* A Population-Based Study of Genes Previously Implicated in Breast Cancer. *N Engl J Med*. 2021;384:440–51.
- Jerzak KJ, Mancuso T, Eisen A. Ataxia-telangiectasia gene (*ATM*) mutation heterozygosity in breast cancer: a narrative review. *Curr Oncol*. 2018 Apr;25(2):e176-e180. doi: 10.3747/co.25.3707. Epub 2018 Apr 30. PMID: 29719442; PMCID: PMC5927797.
- Kaur H, Salles DC, Murali S, Hicks JL, Nguyen M, Pritchard CC, *et al.* Genomic and Clinicopathologic Characterization of *ATM*-deficient Prostate Cancer. *Clin Cancer Res*. 2020;26:4869–81.
- Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*. 2009;4:1073–81.
- Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., ... Maglott, D. R. (2016). ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1), D862–D868. <https://doi.org/10.1093/nar/gkv1222>
- Liu X, Wu C, Li C, Boerwinkle E (2016) dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat* 37:235–241. <https://doi.org/10.1002/humu.22932>
- Loannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, *et al.* REVEL: An Ensemble Method for Predicting the

Pathogenicity of Rare Missense Variants. *Am J Hum Genet.* American Society of Human Genetics; 2016;99:877–85. 17.

López-Ferrando V, Gazzo A, De La Cruz X, et al (2017b) PMut: A web-based tool for the annotation of pathological variants on proteins, 2017 update. *Nucleic Acids Res* 45:W222–W228. <https://doi.org/10.1093/nar/gkx313>

Lu Q, Hu Y, Sun J, et al (2015) A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci Rep* 5:10576. <https://doi.org/10.1038/srep10576>

Masica DL, Karchin R. Towards Increasing the Clinical Relevance of *in silico* Methods to Predict Pathogenic Missense Variants. *PLoS Comput Biol.* 2016;12:e1004725.

Merino Bonilla JA, Torres Tabanera M, Ros Mendoza LH. Breast cancer in the 21st century: from early detection to new therapies. *Radiologia.* 2017 Sep-Oct;59(5):368-379. English, Spanish. doi: 10.1016/j.rx.2017.06.003. Epub 2017 Jul 14. PMID: 28712528.

Minion LE, Dolinsky JS, Chase DM, Dunlop CL, Chao EC, Monk BJ. Hereditary predisposition to ovarian cancer, looking beyond BRCA1/BRCA2. *Gynecol Oncol.* 2015 Apr;137(1):86-92. doi: 10.1016/j.ygyno.2015.01.537. Epub 2015 Jan 23. PMID: 25622547.

Moslemi M, Moradi Y, Dehghanbanadaki H, Afkhami H, Khaledi M, Sedighimehr N, Fathi J, Sohrabi E. The association between *ATM* variants and risk of breast cancer: a systematic review and meta-analysis. *BMC Cancer.* 2021 Jan 5;21(1):27. doi: 10.1186/s12885-020-07749-6. PMID: 33402103; PMCID: PMC7786920.

Musolf AM, Holzinger ER, Malley JD, Bailey-Wilson JE. What makes a good prediction? Feature importance and beginning to open the black box of machine learning in genetics. *Hum Genet.* 2022

Sep;141(9):1515-1528. doi: 10.1007/s00439-021-02402-z. Epub 2021 Dec 4. PMID: 34862561; PMCID: PMC9360120.

Niroula A, Urolagin S, Vihinen M (2015) PON-P2: Prediction Method for Fast and Reliable Identification of Harmful Variants. PLoS One 10:e0117380. <https://doi.org/10.1371/journal.pone.0117380>

Özkan S, Padilla N, Moles-Fernández A, et al (2021) The computational approach to variant interpretation: principles, results, and applicability. In: Lázaro C, Lerner-Ellis J, Spurdle A (eds) Clinical DNA Variant Interpretation: Theory and Practice. Elsevier Inc./Academic Press, San Diego, pp 89–119

Padilla, N., Moles-Fernández, A., Riera, C., Montalban, G., Özkan, S., Ootes, L., ... de la Cruz, X. (2019). BRCA1- and BRCA2-specific *in silico* tools for variant interpretation in the CAGI 5 ENIGMA challenge. *Human Mutation*, 40(9), 1593–1611. <https://doi.org/10.1002/humu.23802>

Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn : Machine Learning in Python. J Mach Learn Res 12:2825–2830.

Petch J, Di S, Nelson W. Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology. Can J Cardiol. 2022 Feb;38(2):204-213. doi: 10.1016/j.cjca.2021.09.004. Epub 2021 Sep 14. PMID: 34534619.

Petracci E, Decarli A, Schairer C, Pfeiffer RM, Pee D, Masala G, Palli D, Gail MH. Risk factor modification and projections of absolute breast cancer risk. J Natl Cancer Inst. 2011 Jul 6;103(13):1037-48. doi: 10.1093/jnci/djr172. Epub 2011 Jun 24. PMID: 21705679; PMCID: PMC3131219.

Pintelas E, Liaskos M, Livieris IE, Kotsiantis S, Pintelas P. Explainable Machine Learning Framework for Image Classification Problems: Case Study on Glioma Cancer Prediction. J Imaging. 2020 May 28;6(6):37. doi: 10.3390/jimaging6060037. PMID: 34460583; PMCID: PMC8321040.

- Piruzan E, Vosoughi N, Mahdavi SR, Khalafi L, Mahani H. Target motion management in breast cancer radiation therapy. *Radiol Oncol*. 2021 Oct 8;55(4):393-408. doi: 10.2478/raon-2021-0040. PMID: 34626533; PMCID: PMC8647788.
- Raimondi D, Tanyalcin I, FertCrossed JSD, et al (2017) DEOGEN2: Prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Res* 45:W201–W206. <https://doi.org/10.1093/nar/gkx390>
- Rentzsch P, Witten D, Cooper GM, et al (2019) CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* 47:D886–D894. <https://doi.org/10.1093/nar/gky1016>
- Reva B, Antipin Y, Sander C (2011) Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Res* 39:e118. <https://doi.org/10.1093/nar/gkr407>
- Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL; ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015 May;17(5):405-24. doi: 10.1038/gim.2015.30. Epub 2015 Mar 5. PMID: 25741868; PMCID: PMC4544753.
- Riera, C., Lois, S., & De la Cruz, X. (2014). Prediction of pathological mutations in proteins: The challenge of integrating sequence conservation and structure stability principles. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. <https://doi.org/10.1002/wcms.1170>
- Riera, C., Padilla, N., & de la Cruz, X. (2016). The Complementarity Between Protein-Specific and General Pathogenicity Predictors for Amino Acid Substitutions. *Human Mutation*, 37(10),

1013–1024. <https://doi.org/10.1002/humu.23048>

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>

Shendure J, Findlay GM, Snyder MW. Genomic Medicine—Progress, Pitfalls, and Promise. *Cell*. 2019;177:45–57.

Schroeder C, Faust U, Sturm M, et al (2015) HBOC multi-gene panel testing: comparison of two sequencing centers. *Breast Cancer Res Treat* 152:129–136. <https://doi.org/10.1007/s10549-015-3429-9>

Shihab HA, Gough J, Cooper DN, et al (2013) Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Hum Mutat* 34:57–65. <https://doi.org/10.1002/humu.22225>

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shaw, K., & Cooper, D. N. (2012). The human gene mutation database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Current Protocols in Bioinformatics*, 39, 1.13.1-1.13.20. <https://doi.org/10.1002/0471250953.bi0113s39>

Stucci LS, Internò V, Tucci M, Perrone M, Mannavola F, Palmirotta R, Porta C. The *ATM* Gene in Breast Cancer: Its Relevance in Clinical Practice. *Genes (Basel)*. 2021 May 13;12(5):727. doi: 10.3390/genes12050727. PMID: 34068084; PMCID: PMC8152746.

Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., Kondrashov, A. S., & Bork, P. (2001). Prediction of deleterious human alleles. *Human Molecular Genetics*, 10(6), 591–597. <https://doi.org/10.1093/hmg/10.6.591>

Tavtigian, S. V., Greenblatt, M. S., Lesueur, F., & Byrnes, G. B. (2008). *in*

*silico* analysis of missense substitutions using sequence-alignment based methods. *Human Mutation*, 29, 1329–1336. <https://doi.org/10.1002/humu.20892>

Vihinen, M. (2012). How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics*, 13, S2. <https://doi.org/10.1186/1471-2164-13-S4-S2>

Warren C, Pavletich NP. Structure of the human *ATM* kinase and mechanism of Nbs1 binding. *Elife*. 2022;11:e74218.

Waskom M (2021) seaborn: statistical data visualization. *J Open Source Softw* 6 (60):3021. <https://doi.org/10.21105/joss.03021>

Wojtyla C, Bertuccio P, Ciebiera M, La Vecchia C. Breast Cancer Mortality in the Americas and Australasia over the Period 1980-2017 with Predictions for 2025. *Biology (Basel)*. 2021 Aug 23;10(8):814. doi: 10.3390/biology10080814. PMID: 34440046; PMCID: PMC8389642.

Zhu SY, Yu KD. Breast Cancer Vaccines: Disappointing or Promising? *Front Immunol*. 2022 Jan 28;13:828386. doi: 10.3389/fimmu.2022.828386. PMID: 35154149; PMCID: PMC8831788.

## 9 ANEXOS

## **Anexo 1. Variantes Vus Priorizadas según RF\_Metap**

### **Variantes Patogénicas**

D2016V	W2205R	D2395G	W2491C	D2672Y	R2723I
G2020R	W2205C	D2395V	K2515T	D2672A	Q2724R
G2020D	L2255R	Q2397P	M2520R	G2694R	D2725H
G2020V	L2293P	Q2397R	Y2521C	G2695R	D2725G
W2104S	W2300R	Y2398C	L2523W	G2695C	D2725V
R2105W	W2300C	L2417P	R2526G	G2695D	V2727A
E2164K	A2308T	R2436G	R2526T	G2695A	Q2729P
E2164A	L2312P	Y2437C	L2544P	G2695V	Q2730R
E2164G	Y2360H	R2443P	L2557W	K2700N	Q2730L
E2164V	Y2360S	L2445P	R2598L	I2702T	V2731A
Y2170C	Y2360C	D2458V	A2602T	I2702R	F2732S
L2176W	L2361P	Y2470C	L2623H	C2704Y	K2756E
E2181G	R2392P	L2474S	Y2627C	K2717Q	L2760P



L2182P	F2393C	D2481G	D2634V	G2718S	G2765D
G2765V	Y2852H	H2872L	G2897S	C2930G	S3046R
W2769R	S2855N	N2875T	G2897D	C2930Y	G3051R
C2770R	V2856E	I2878T	V2906A	C2930F	G3051E
C2770G	A2857T	L2885H	R2912G	C2930W	G3051A
C2770F	S2860C	H2887P	R2912T	E2932G	W3052R
I2776N	G2863V	I2888T	D2913H	M2938K	W3052S
L2780P	I2865R	I2888R	I2914T	E2950K	W3052C
C2824Y	G2867E	D2889H	V2915A	L2952P	W3055S
P2829L	G2869S	D2889G	G2922A	Y2954C	
V2830G	G2869R	L2890P	G2922V	P2956L	
F2831S	G2869D	G2891S	V2923D	L2957R	
R2832G	G2869A	G2891C	G2925S	W2960C	
F2834C	D2870Y	G2891D	G2925R	L3026R	
D2841H	D2870G	G2891V	G2925D	G3030V	
D2841Y	R2871G	A2893T	G2925A	N3044I	

R2849Q R2871T Q2896P G2925V L3045P

**Variantes Benignas**

D1963N L1975I K1992R S2000I Q2028K M2041V  
K1964R A1976T S1993R D2003N Q2028E M2041T  
K1964N A1976V S1993N D2003E P2029R G2043S  
K1965N E1978D S1993I L2004I P2029L G2043V  
S1966N G1980R K1994Q L2004F I2030L L2046V  
M1967V Q1982R K1994R L2006V T2031P V2047I  
D1968G T1984S E1995K E2007A T2031I V2047L  
Q1970R T1984K E1996G E2007D L2033P V2047E  
E1971V T1984R T1997P I2008V T2035A V2047G  
E1971D T1985A T1997I G2022S E2037Q L2051I  
K1972E T1985I G1998E G2024R H2038N L2051F  
K1972T S1988R G1998V G2024W H2038Y E2052G  
R1973G E1991K I1999L G2024E A2040P A2054T  
R1973K E1991G I1999V K2025N A2040E A2054S  
S1974N E1991D S2000N M2026I A2040V I2055V  
P2056S Q2066K K2082E D2090A D2110E E2126K

P2056L	Q2066E	K2082N	D2090E	H2111P	L2128F
S2057P	Q2066R	G2083E	W2091G	H2111R	N2130H
S2057L	Q2066H	G2083A	C2092G	C2112S	L2132V
S2058L	A2067T	L2084V	P2093S	C2112Y	L2135V
T2059A	A2067S	D2085H	L2095V	C2112F	R2136K
T2059K	L2068F	D2085Y	E2096G	T2113A	D2137N
R2060S	L2071V	D2085E	E2097Q	S2114Y	D2137H
R2060C	G2072R	Y2086H	E2097D	V2115I	D2137E
R2060P	H2075P	Y2086C	H2099D	V2115L	R2138G
R2060L	H2075R	Y2086F	H2099R	E2118V	T2142A
Q2061P	I2076V	N2088D	A2102T	E2120D	T2142I
A2062E	L2077V	N2088S	A2102S	T2122S	Y2144C
G2063A	S2078P	K2089N	A2102V	S2123G	E2145A
I2064T	S2078F	D2090N	M2107T	S2123T	E2145G
I2065T	L2081F	D2090Y	Q2108H	S2123I	S2146N
K2148T	L2163V	I2185T	I2203T	M2224L	Q2243K

K2148I	S2165T	I2185S	K2204Q	M2224V	Q2243P
V2152I	Y2167C	L2188I	Q2206E	I2230T	R2244K
K2153Q	L2169V	L2188F	Q2206H	E2232D	E2245Q
V2155A	L2169F	R2191G	K2207R	I2233L	E2245G
E2156K	L2173V	R2191T	H2208Y	I2233M	C2246G
E2157D	S2174R	V2193I	H2208Q	M2235V	C2246Y
M2158T	S2174N	T2194I	Q2210P	K2237T	I2247L
M2158I	R2175K	H2195Y	L2211F	K2237R	I2247V
C2159R	R2175M	H2195P	L2212I	E2238V	I2247M
C2159G	E2183K	H2195R	S2215G	M2239L	K2248E
C2159Y	S2184G	Q2197P	S2215T	D2240N	K2248T
R2161G	S2184C	S2199Y	F2217Y	D2240H	D2249N
R2161C	S2184T	E2200K	F2219L	N2241T	D2249Y
R2161H	I2185F	V2201A	I2223L	N2241S	D2249G
S2162G	I2185N	Y2202H	I2223V	S2242P	D2249V
L2251V	F2265C	N2282D	K2302E	D2320G	L2332F

T2252S	K2266Q	N2282S	K2302R	A2321P	L2332R
K2253Q	K2266R	S2283L	K2303T	A2321D	T2333R
K2253T	Q2269K	V2284F	K2303M	S2322N	Y2334H
K2253R	Q2269E	C2286Y	S2306N	S2322I	T2335P
K2253N	P2271S	C2286F	L2307V	C2323Y	L2338V
L2255F	R2273K	G2287R	L2309V	C2323F	R2339K
V2256I	A2274S	G2287E	S2310G	A2325V	V2340I
V2256E	I2275L	V2288D	S2310N	N2327D	V2340F
V2256A	I2275V	S2289F	I2311V	N2327I	N2343S
E2257D	I2275T	E2290K	Q2314H	P2328T	L2345V
S2259C	I2275M	E2290Q	M2315I	P2328A	P2353A
I2260V	F2276L	E2290A	K2317T	P2328S	A2354T
I2260M	Q2277E	E2290G	K2318Q	P2328L	A2354E
L2261V	I2278M	V2298I	K2318E	S2329T	A2354V
A2262G	Y2281H	A2301T	L2319M	L2332V	V2355D
I2356V	Y2371D	L2390F	K2421T	N2435K	C2464G

M2357T	Y2371C	Y2398H	K2421I	V2439L	C2464Y
Q2358R	Y2371F	R2400I	E2422K	V2439A	V2467I
Q2358H	D2372V	R2400S	E2422G	V2439G	I2471V
T2359N	E2374A	I2401V	V2424A	E2446K	N2472D
T2359I	E2374D	M2405L	G2425V	E2446G	N2472S
L2361I	S2375G	F2410Y	R2428G	L2450V	L2475S
L2361V	S2375N	N2412D	R2428W	A2451T	S2476N
A2364V	S2376R	N2412T	R2428K	A2451S	E2479Q
V2365I	D2377N	N2412S	H2430N	L2452V	H2480Y
E2366K	D2377Y	Q2414K	H2430D	R2453C	D2481N
V2367I	L2379V	A2415G	H2430R	R2453P	M2482L
V2367L	G2382A	A2415V	I2432V	A2454P	M2482V
A2368V	M2384T	L2416F	I2432N	A2454E	M2482I
G2369E	A2386G	K2418T	Q2433L	E2457A	W2483R
N2370D	S2389L	K2421E	N2435D	R2461H	V2484L
S2495F	M2509V	N2543D	T2573A	K2589R	T2608A

G2496E	M2509R	N2543S	T2573S	Q2590E	I2609V
G2496A	M2509I	M2550L	K2574Q	Q2590R	I2609T
S2498A	P2512S	M2550V	P2575Q	S2591R	R2612K
E2499G	T2513A	M2550I	E2576G	S2591N	P2614S
V2500I	Y2514H	D2551N	V2577I	E2596K	P2614H
N2501S	Y2514F	L2557F	A2578T	D2597H	P2614L
G2502D	M2520I	I2559V	A2578P	T2599I	Q2615E
G2502V	M2532T	A2562S	R2579T	E2600K	V2617F
M2503L	M2532I	L2563I	R2580K	N2603D	R2618G
M2503V	G2533E	L2563V	S2581R	N2603S	R2618S
M2503I	G2534S	A2564T	S2581N	R2604S	S2619G
R2506G	E2539K	A2566G	R2582G	I2605V	V2620L
R2506K	E2539D	R2568G	I2583K	I2606V	V2620F
D2507E	N2542S	D2569N	V2587M	I2606M	E2621Q
G2508R	N2542K	F2571S	K2589E	C2607R	A2622T
A2622G	Q2637R	T2654S	L2680M	K2710R	T2751A

L2623I	W2638G	T2654I	I2683L	K2710M	T2751S
L2623F	K2639N	K2655T	I2683V	R2712K	I2752V
I2628L	K2643E	L2656F	I2683M	R2713K	C2753Y
I2628V	K2643N	K2657E	Q2684R	V2716I	C2753F
I2629V	G2644S	N2658Y	A2688T	R2719L	T2754A
I2629M	G2644R	E2660D	E2689G	Q2733R	T2754I
L2630I	I2645V	V2662I	R2691S	M2734I	L2760V
L2630V	I2645M	V2664F	R2691G	N2736S	L2760F
N2632K	N2646H	M2667L	R2691H	T2737R	R2763G
L2633V	P2648A	M2667V	V2696I	R2741G	R2763Q
D2634E	P2648S	M2667I	V2696A	T2743R	V2766I
A2635D	A2649V	E2668Q	D2703N	T2743M	T2771A
A2635G	D2650N	H2673Y	D2703V	E2744A	T2771I
A2635V	D2650A	E2676K	V2705I	K2747R	V2774A
T2636I	I2653V	G2678R	V2705L	K2749R	G2777S
E2778A	Q2802R	S2812C	Y2833H	Q2874H	N2940D



F2779Y	K2804Q	F2813L	C2835Y	I2878L	N2940K
D2785G	M2805I	F2813S	K2838Q	N2879D	S2941A
D2785V	M2806T	E2815K	K2838R	N2879S	T2944I
K2789R	M2806I	K2816Q	L2840S	Q2881R	L2945V
N2794K	E2807K	Y2817C	L2840F	Q2881H	I2948V
F2796L	E2807V	V2819G	I2844L	S2882L	V2949I
S2797G	V2808M	M2821V	I2844S	V2892I	L2968F
S2797N	V2808A	M2821T	F2846I	I2899V	Y2969H
A2798S	Q2809E	M2821R	F2846L	I2899M	Q2972R
A2798D	K2810Q	D2822N	V2856I	T2902I	R2973G
A2798V	K2810E	C2824S	I2861V	T2911S	R2973M
F2799I	K2810T	Q2825R	I2865V	I2914M	R2973S
F2799V	S2812P	N2826S	L2866V	I2920V	P2974T
Q2800P	S2812A	Q2828E	V2873I	M2935I	P2974S
Q2800H	S2812Y	Q2828R	V2873A	E2936Q	E2975Q
E2979Q	D2987Y	S3001N	K3018E	K3043R	

E2979D	D2987G	S3001T	V3020L	R3047Q
L2980I	Q2989E	F3002Y	G3023S	P3050L
L2980V	Q2989R	K3004E	T3024P	K3053Q
H2981N	C2991R	V3005I	T3024N	A3054T
H2981Y	C2991Y	V3005A	T3024I	A3054P
H2981R	C2991F	V3009G	V3025M	A3054G
H2981Q	R2993Q	M3011L	V3025L	V3056M
P2982T	N2994T	M3011V	V3028I	
T2983P	L2995I	M3011K	G3029V	
T2983N	S2996N	M3011T	L3034W	
T2983I	D2997N	M3011R	I3036V	
L2984M	I2998V	Q3014R	Q3037H	
L2984V	I2998N	E3015K	I3040V	
N2985K	D2999V	E3015Q	I3040T	
D2987H	Q3000E	E3015G	K3043E	

## Tabla anexo2. Variantes Vus Priorizadas según RF\_Bioinf

### Variantes Patogénicas

I1960N	G2024E	G2063V	E2164G	W2291R	R2392W
F1977S	G2024R	G2072E	Y2170C	L2293P	F2393C
L1989W	G2024W	G2072R	L2176W	W2300C	D2395G
G1998E	R2032I	L2073P	E2181G	W2300R	Y2398C
G1998V	Y2036C	Y2080C	L2182P	L2312P	Y2404C
L2004P	Y2036D	G2083E	W2205C	I2316N	S2408W
Y2009N	E2037G	D2090Y	W2205R	L2319W	E2422G
G2013R	E2039G	W2104S	F2219S	E2351G	R2436G
P2015L	W2042C	R2105G	L2226P	Y2360C	Y2437N
G2020R	W2042R	R2105W	A2262D	Y2360S	L2445P
G2020V	A2045D	N2106I	F2265C	L2361P	L2455P
C2021R	Y2049C	H2125L	E2272A	A2386E	R2459C
C2021Y	Y2049D	L2135P	E2272G	R2392L	R2459G

G2023E	Y2049K	E2164A	W2291C	R2392P	A2466E
Y2470C	L2544P	D2672G	R2713I	P2759L	D2795Y
G2477R	R2547G	D2672Y	G2718R	G2765V	C2824Y
D2481G	D2551G	F2690C	D2720G	V2766G	P2829L
W2483R	P2553L	G2694R	D2721G	L2767P	V2830G
C2488F	P2553R	G2695C	D2721Y	W2769R	F2831S
C2488R	H2554L	G2695R	R2723I	C2770F	F2834C
S2489F	L2557W	G2695V	D2725G	C2770G	D2841Y
S2489Y	A2562D	N2697Y	V2727G	C2770R	P2842R
W2491C	D2569G	P2699L	V2731G	G2772E	R2854G
F2516C	R2598L	P2699R	F2732S	G2772R	R2854L
P2518L	C2624G	I2702R	E2744A	P2775L	R2854P
Y2521C	Y2627C	C2704R	C2753F	G2777V	G2863V
L2523W	P2652L	C2704Y	C2753Y	L2780P	G2867E
R2526G	P2665L	D2708Y	Y2755S	G2786C	G2869R

G2533E	D2672A	G2709C	V2757G	D2795A	D2870G
D2870Y	V2926G	G3030V			
R2871G	C2930F	P3042L			
H2872L	C2930G	N3044I			
V2886G	C2930W	L3045P			
I2888R	C2930Y	P3050L			
D2889G	C2931R	G3051E			
L2890P	E2932G	G3051R			
G2891C	L2952P	W3052C			
G2891V	Y2954C	W3052R			
P2901L	P2956L	W3052S			
P2907R	W2960C				
R2912G	V3009G				
G2922V	L3017P				
G2925R	G3023V				

G2925V G3030E

### **Variantes Benignas**

D1963N	L1975I	K1994R	Q2028E	A2054S	A2067S
K1964N	A1976T	E1995K	Q2028K	A2054T	L2071V
K1964R	A1976V	E1996G	I2030L	I2055V	L2073V
K1965N	E1978D	I1999L	T2031I	P2056S	I2076V
S1966N	G1980R	I1999V	T2031P	S2057L	L2077V
M1967V	T1984K	S2000N	E2037Q	S2057P	S2078P
D1968G	T1984S	D2003E	M2041T	S2058L	K2082E
Q1970R	T1985A	D2003N	M2041V	T2059A	K2082N
E1971D	T1985I	L2004I	G2043S	A2062E	L2084V
E1971V	E1991D	L2006V	L2046V	I2065T	D2085E
K1972E	E1991K	E2007D	V2047E	Q2066E	D2085H
K1972T	K1992R	I2008V	V2047I	Q2066H	D2085Y
R1973K	S1993N	E2014Q	V2047L	Q2066K	Y2086C
S1974N	K1994Q	M2026I	L2051I	Q2066R	Y2086F

Y2086H	H2099R	V2115L	Y2144C	E2164Q	I2185N
N2088D	Q2101K	E2118V	E2145A	S2165T	I2185S
N2088S	A2102S	E2120D	E2145G	Y2167C	I2185T
K2089N	Q2108H	T2122S	S2146N	L2169F	L2188I
D2090A	Q2108P	S2123T	K2148I	L2169R	V2193I
D2090E	Q2108R	Y2129H	K2148T	L2169V	H2195P
D2090N	D2110E	N2130H	V2152I	L2173V	H2195R
P2093S	H2111P	L2132V	K2153Q	S2174N	H2195Y
L2095V	H2111R	L2135V	E2156K	S2174R	Q2197P
E2096G	C2112F	R2136K	E2157D	R2175K	S2199Y
E2097D	C2112S	D2137E	M2158I	E2183K	E2200K
E2097Q	C2112Y	D2137H	R2161C	S2184C	V2201A
L2098P	T2113A	D2137N	R2161G	S2184G	Y2202H
L2098R	S2114Y	T2142A	R2161H	S2184T	I2203T

H2099D	V2115I	T2142I	L2163V	I2185F	K2204Q
Q2206E	I2233M	E2245Q	K2253R	I2275L	S2289P
Q2206H	L2234Q	C2246G	V2256A	I2275M	E2290K
K2207R	M2235V	C2246Y	V2256E	I2275T	E2290Q
H2208Q	K2237R	I2247L	V2256I	I2275V	V2298I
H2208Y	K2237T	I2247M	E2257D	F2276L	A2301P
Q2210P	M2239L	I2247V	S2259C	Q2277E	A2301T
L2212I	D2240H	K2248E	I2260M	I2278M	K2302R
D2216N	D2240N	K2248T	I2260V	Y2281H	K2303M
I2223L	N2241S	D2249N	L2261V	N2282D	K2303T
I2223V	N2241T	L2251V	K2266Q	N2282S	S2306N
M2224L	S2242P	T2252N	K2266R	S2283L	L2307H
M2224V	Q2243K	T2252P	N2267S	V2284F	L2307V
V2229D	Q2243P	T2252S	T2268A	C2286F	L2309V
E2232D	R2244K	K2253N	P2271S	C2286Y	S2310G



I2233L	E2245G	K2253Q	R2273K	V2288D	S2310N
Q2314H	P2328T	A2354E	Y2371D	N2412S	Q2433L
M2315I	S2329T	A2354T	Y2371F	N2412T	N2435D
K2317T	L2332F	A2354V	D2372V	Q2414K	T2438P
L2319M	L2332R	T2359I	E2374A	A2415G	V2439A
D2320G	L2332V	T2359N	E2374D	A2415V	V2439L
A2321D	T2333R	L2361I	S2375G	K2418T	V2441F
A2321P	Y2334H	L2361V	S2375N	K2421E	Q2442H
S2322I	T2335P	V2365I	S2376R	K2421I	Q2442R
S2322N	C2337Y	E2366K	D2377N	L2427R	E2444D
C2323F	L2338V	V2367I	D2377Y	R2428K	E2446G
C2323Y	R2339K	V2367L	L2379V	H2430D	E2446K
A2325V	V2340F	A2368V	G2382A	H2430N	L2447V
N2327D	V2340I	G2369E	M2384T	H2430R	D2448N
P2328A	N2343S	N2370D	F2410Y	K2431E	L2450V

P2328S	L2345V	Y2371C	N2412D	I2432V	A2451S
L2452Q	M2482L	R2506G	M2532I	R2568G	K2585E
L2452V	M2482V	R2506K	G2534S	F2571S	V2587M
R2453C	V2484L	D2507E	H2538Q	T2573A	K2589E
R2453P	S2495F	G2508R	E2539D	T2573S	K2589R
A2454E	G2496A	M2509I	E2539K	K2574Q	Q2590E
A2454P	G2496E	M2509R	N2542S	P2575Q	Q2590R
C2464Y	S2498A	M2509V	N2543D	E2576G	S2591N
V2467I	E2499G	P2512S	N2543S	V2577I	S2591R
N2472D	V2500I	P2512T	L2544V	A2578P	E2596K
N2472S	N2501S	Y2514F	M2550I	A2578T	T2599I
L2475S	G2502D	Y2514H	M2550L	R2579T	E2600K
S2476N	G2502V	L2519I	M2550V	R2580K	N2603D
E2479Q	M2503I	L2519V	T2556N	S2581N	N2603S
H2480Y	M2503L	M2520I	L2563I	S2581R	R2604S

M2482I	M2503V	M2531R	L2563V	I2583K	I2605V
I2606M	V2620F	A2635G	I2653V	V2681L	V2696A
I2606V	V2620L	T2636I	T2654I	V2681M	V2696I
C2607R	E2621Q	Q2637R	T2654S	T2682A	D2703N
T2608A	A2622G	K2643E	K2655T	T2682S	V2705I
I2609T	A2622T	K2643N	L2656R	I2683L	V2705L
I2609V	L2623I	G2644R	E2660D	I2683M	S2707T
R2612K	L2623V	G2644S	V2662I	I2683V	K2710M
P2614H	D2625A	I2645M	M2667I	Q2684R	K2710R
P2614L	D2625H	N2646H	M2667L	A2688T	R2713K
P2614S	I2629M	P2648S	M2667V	E2689G	L2715I
Q2615E	I2629V	P2648T	E2668Q	R2691C	V2716I
M2616V	L2630I	D2650A	H2673Y	R2691G	K2717R
R2618G	L2630V	D2650N	E2676K	R2691H	R2719H
R2618S	L2633V	Q2651K	G2678R	R2691P	R2719L

S2619G	A2635D	I2653F	L2680M	R2691S	L2722V
V2731I	T2771A	F2799V	F2813L	K2838R	Q2881H
Q2733R	T2771I	Q2800H	E2815K	F2839I	Q2881R
M2734I	F2779Y	Q2800P	K2816Q	F2839L	L2890I
N2736D	D2785G	M2805I	V2819G	I2844L	L2890V
N2736S	D2785V	M2806I	M2821R	I2844S	V2892I
T2743M	K2789R	E2807K	M2821T	L2850V	E2895Q
K2747R	N2794K	V2808A	M2821V	S2855N	I2899M
K2749R	F2796L	V2808M	D2822N	V2856I	I2899V
L2750I	S2797G	K2810E	Q2825R	S2859T	L2900I
T2751A	S2797N	K2810Q	N2826S	L2866V	E2904Q
T2751S	A2798D	K2810T	Q2828E	L2868I	V2906I
K2756R	A2798F	S2812A	Q2828R	L2868V	I2914M
L2760V	A2798S	S2812C	Y2833H	V2873I	V2915M
R2763Q	A2798V	S2812P	M2836V	N2879D	D2916N

V2766I	F2799I	S2812Y	K2838Q	N2879S	E2924D
--------	--------	--------	--------	--------	--------

V2926I	E2979D	C2991Y	Q3014H	I3040V
E2936Q	E2979Q	R2993Q	Q3014R	K3043R
N2940D	L2980I	N2994T	E3015K	L3045V
S2941A	L2980V	L2995I	E3015Q	L3048V
L2945V	H2981N	S2996N	T3024N	F3049L
V2949I	H2981Q	D2997N	T3024P	K3053Q
V2951L	H2981R	I2998N	S3027T	V3056M
L2952V	H2981Y	I2998V	V3028I	
Y2969H	P2982T	D2999V	V3032L	
R2973G	L2984M	Q3000E	V3032M	
R2973M	L2984V	S3001N	N3033S	
R2973S	N2985K	F3002Y	L3034W	
P2974S	D2987H	V3005I	L3035I	
P2974T	Q2989E	M3011L	Q3038H	
E2975Q	Q2989R	M3011V	I3040T	

