Universidad Autónoma de Madrid

Facultad de Medicina

M.U. Bioinformática y Biología Computacional

Trabajo Fin de Máster

Primeros modelos de *machine learning* para la predicción de la patogenicidad de variantes *missense* en canales $K_v7.2$

Autor: Alba Sáez Matía

Tutor: Álvaro Villarroel Muñoz

Ponente: Modesto Redrejo Rodríguez

Febrero 2022

Abstract

Los rápidos avances en las tecnologías de secuenciación de ADN han puesto en evidencia las limitaciones de la validación experimental de variantes proteicas así como los obstáculos de su diagnóstico clínico. En el presente estudio se ha demostrado como los modelos computacionales con capacidad predictiva constituyen un enfoque valioso para evaluar la patogenicidad de las variantes, puesto que acelerarían su diagnóstico y el tratamiento de las patologías asociadas. Para ello, se han diseñado los primeros modelos de aprendizaje automático capaces de predecir la patogenicidad de variantes *missense* en el canal de potasio dependiente de voltaje $K_v7.2$, asociado a diversos tipos de epilepsias neonatales. Los resultados de especificidad y sensibilidad obtenidos alcanzaron los valores de 87.17 % y 82.53 %, respectivamente. A pesar de que las mutaciones recopiladas de $K_v7.2$ se caracterizaron mediante descriptores de secuencia y de estructura secundaria, no pudieron diseñarse descriptores tridimensionales a causa de las dificultades encontradas. No obstante, el cálculo de la inestabilidad proteica provocada por una mutación ($\Delta\Delta G$) permitió observar el potencial clínico de los modelos entrenados. El análisis de las reclasificaciones de variantes mostró que el 9.82 % de las mutaciones de KCNQ2 anotadas en ClinVar fueron reclasificadas en un periodo de 2 años, siguiendo una tendencia habitual respecto a estudios anteriores.

Palabras Clave: KCNQ2, epilepsia, predicción del efecto de variantes, patogenicidad, modelos de *machine learning.*

Índice General

1. Introducción

1.1. Era genómica, diagnóstico clínico y métodos computacionales

Los rápidos avances en las tecnologías de secuenciación de ADN han hecho realidad la genómica clínica a gran escala, donde la identificación de variantes y sus costes ya no constituyen los pasos limitantes [1–3]. Como consecuencia, en la última década laboratorios de todo el mundo han secuenciado genomas completos de miles de pacientes en la búsqueda de mutaciones causantes de enfermedades [4]. Sin embargo, identificar las pocas variantes fenotípicamente causales entre las numerosas variantes presentes en los genomas humanos sigue siendo un reto importante, especialmente en el caso de las enfermedades raras y patologías complejas en los que la información genética por sí sola suele ser insuficiente [1,5]. Asimismo, la interpretación fiable de variantes múltiples y *de novo* detectadas mediante secuenciación requiere de una validación adicional [6,7].

Por todo ello, las tecnologías moleculares no relacionadas con la secuenciación siguen siendo cruciales en la interpretación clínica de las variantes [8]. No obstante, y a pesar de su importancia, los ensayos funcionales y la validación experimental de todas las posibles variantes patológicas resultantes de la secuenciación a gran escala sigue siendo, a día de hoy, muy laboriosa y costosa [5,9].

Ante estos problemas, la predicción computacional de la clínica de las variantes genéticas ha cobrado importancia en los últimos años, especialmente los métodos basados en aprendizaje automático [10,11]. Estos métodos constituyen una herramienta muy valiosa ya que permiten procesar fácilmente el gran número de mutaciones que generan las nuevas herramientas de secuenciación proporcionando información útil, casi sin coste, sobre su carácter patológico [12]. Asimismo, estos algoritmos permiten inferir automáticamente la patogenicidad de nuevas mutaciones que pudieran aparecer, en base a descriptores de secuencia (Ej.: cambio de aminoácido, conservación evolutiva del residuo, etc.) o estructurales (Ej.: estructura secundaria, cálculos energéticos de estabilidad, etc.), siendo condición indispensable para estos últimos la estructura de la proteína en donde se estuvieran estudiando las variantes [13,14].

En definitiva, estos métodos computacionales consiguen reducir enormemente los costes que suponen a los laboratorios clínicos los ensayos funcionales de variantes puesto que permiten

acotar el número de muestras a validar, estudiando solo aquellas que se hubieran predicho por el algoritmo como patológicas. Así, se conseguiría validar más rápidamente y con menores costes las mutaciones con verdadera clínica patológica, siendo esta su principal motivación.

1.2. El dilema de la estructura proteica

En 2009 se estimó que, en general, la estructura se conservaba entre 3 y 10 veces más que la secuencia de las proteínas [15]. Además, se acepta que el conocimiento de la estructura tridimensional de las proteínas es crucial para responder a un vasto número de cuestiones biológicas entre las que se encuentra la patogenicidad de las variantes proteicas [16, 17].

Desde entonces, y gracias a los esfuerzos de laboratorios y a iniciativas dedicadas a la genómica estructural, se han depositado más de 50.000 estructuras de proteínas humanas en Protein Data Bank (PDB) [16]. No obstante, se ha estimado que solo el 35 % de las proteínas humanas cuenta con una entrada de PDB y, en muchos casos, la estructura únicamente cubre un fragmento de la secuencia proteica [18].

Con los rápidos avances de las técnicas de secuenciación de ADN se generó una avalancha de nuevas secuencias cuya estructura era imposible de determinar mediante técnicas experimentales por estar repletas de obstáculos y limitaciones [16, 19]. En ausencia de estructura experimental, los métodos computacionales de modelado por homología (*homology modelling*) y *ab-initio* se utilizaron durante décadas para pronosticar de alguna manera la estructura tridimensional de las proteínas [20]. Aun así, la predicción computacional de la estructura de las proteínas a partir de la secuencia de aminoácidos se siguió considerando un problema fundacional de la bioquímica y uno de los retos más difíciles de la bioinformática actual [21].

La predicción de estructuras proteicas ha experimentado un progreso sustancial en los últimos años gracias a la aparición de AlphaFold2 [22] y RosseTTAFold [23], dos métodos basados en el aprendizaje profundo (*deep learning*) [24].

Por un lado, AlphaFold2 es un algoritmo de inteligencia artificial desarrollado por Deepmind que ha demostrado predecir con mayor precisión la estructura tridimensional (3D) de una proteína en comparación con otros métodos conocidos hasta la fecha. AlphaFold2 se basa en alineamientos múltiples de secuencias (MSA) de homólogos de la proteína objetivo y en una red neuronal basada en el aprendizaje profundo para predecir con precisión las distancias entre pares de aminoácidos [22]. El código fuente de AlphaFold2 ha sido publicado y, además, las

estructuras predichas por el *software* han sido puestas a disposición de la comunidad científica gracias al Instituto Europeo de Bioinformática (EMBL-EBI) a partir de la Base de Datos de Estructuras proteicas de AlphaFold.

Por el otro lado, RoseTTAFold es un algoritmo de aprendizaje profundo con características inspiradas en AlphaFold2 [13]. Para su predicción, RoseTTAFold emplea una red neuronal de tres capas que combina tanto información a nivel de secuencia como de mapas de distancia y coordenadas atómicas [23]. Este *software* ha sido desarrollado y publicado por el grupo Baker y se ha puesto a disposición de todos al liberar el código fuente y como parte del Robetta Server.

1.3. ClinVar como fuente de información

ClinVar, en el Centro Nacional de Información Biotecnológica (NCBI), es un archivo de libre acceso para la interpretación de la significancia clínica de las variantes genéticas conocidas. Según la información pública disponible en ClinVar Miner, actualmente ClinVar registra más de 1.900.000 variantes, que afectarían a más de 34.200 genes.

Como la significancia clínica es reportada directamente por laboratorios distribuidos por todo el mundo, el Colegio Americano de Genética Médica y Genómica y la Asociación de Patología Molecular (ACMG-AMP) diseñaron en 2015 unas directrices básicas que ayudaban a estandarizar la clínica de las variantes genéticas entre los distintos laboratorios [4, 25]. Así, establecieron cinco categorías básicas en las que se podían clasificar las variantes genéticas: patológicas (*pathogenic*), probablemente patológicas (*likely pathogenic*), variantes de significado incierto (VUS), probablemente benignas (*likely benign*) y benignas (*benign*). Las VUS son variantes genéticas con un impacto desconocido en la salud, lo que hace que tanto los pacientes como los proveedores no estén seguros de su clínica real [26]. Más adelante, ClinVar añadió una categoría adicional para hacer referencia a aquellas mutaciones cuya clínica era conflictiva entre miembros de un mismo consorcio, las variantes de interpretación conflictiva [27].

Por último, las variantes genéticas no siempre presentan una clínica estática ya que pueden reclasificarse con el tiempo a medida que se descubren nuevos conocimientos sobre sus efectos fenotípicos o los investigadores presentan nuevas pruebas [5,28]. Este hecho se ha hecho especialmente visible con los avances en las pruebas genéticas de los últimos años [29]. Además, estas reclasificaciones de variantes pueden generar circunstancias difíciles tanto para los pacientes como para los médicos desde que pueden ocasionar cambios en el tratamiento [26, 28].

1.4. Trastornos epilépticos y canales $K_v7.2$

La epilepsia es un trastorno neurológico crónico que alcanza una incidencia anual de más de 67 personas por cada 100.000 habitantes [30]. Según la Organización Mundial de la Salud (OMS), unos 50 millones de personas padecen esta enfermedad en todo el mundo, lo que la convierte en uno de los trastornos neurológicos más comunes [31]. Dentro de los trastornos epilépticos, nos enfocamos en la Epilepsia Neonatal Benigna Familiar (BFNE) y la Encefalopatía Epiléptica (EE).

Por un lado, la BFNE es un síndrome raro de herencia dominante con una incidencia de 1 de cada 100.000 personas y con una penetrancia de hasta el 85 % [32, 33]. Esta condición se caracteriza por la aparición precoz de convulsiones generalizadas en el neonato que desaparecen tras los primeros meses de vida. No obstante, alrededor del 15 % de los pacientes pueden tener convulsiones recurrentes más adelante así como presentar características clínicas más severas [34, 35]. Por el otro lado, la EE es una forma grave de epilepsia neonatal también de herencia dominante. Por lo general, la EE se manifiesta en recién nacidos durante la primera semana de vida con convulsiones a menudo acompañadas de sacudidas clónicas o de un comportamiento motor más complejo [36]. Entre las causas que generan ambos trastornos, las mutaciones en los canales neuronales $K_v7.2$ son y han sido recurrentes [37–40].

El canal $K_v7.2$ es codificado por el gen KCNQ2 y pertenece a la familia de los canales de potasio dependientes de voltaje, cuyos miembros cumplen un papel crucial en la excitabilidad neuronal y cardíaca ($K_v7.1$-5) [41]. No obstante, $K_v7.2$ se caracteriza por expresarse mayoritariamente en el sistema nervioso, donde desempeña un papel fundamental en la excitabilidad neuronal al mediar la "corriente M", ayudando a la hiperpolarización del potencial de acción [42–44].

Estructuralmente, y al igual que el resto de miembros de su familia, el canal $K_v7.2$ es una proteína tetramérica de membrana (ver Figura 1). Topológicamente, cada monómero presenta seis segmentos transmembrana (S1-S6). Dentro de estos segmentos, se distingue el dominio sensor de voltaje (S1-S4) y el dominio funcional del poro (S5-S6). Además, los extremos N y C-terminal son intracelulares, siendo este último de mucha mayor longitud. El extremo C-terminal es clave en la regulación y función del canal, puesto que contiene el dominio de interacción de subunidades (SID), que permite la formación de los tetrámeros de $K_v7.2$; así como el dominio de interacción con calmodulina (*CaM interaction domain*), indispensable para su correcto fun-

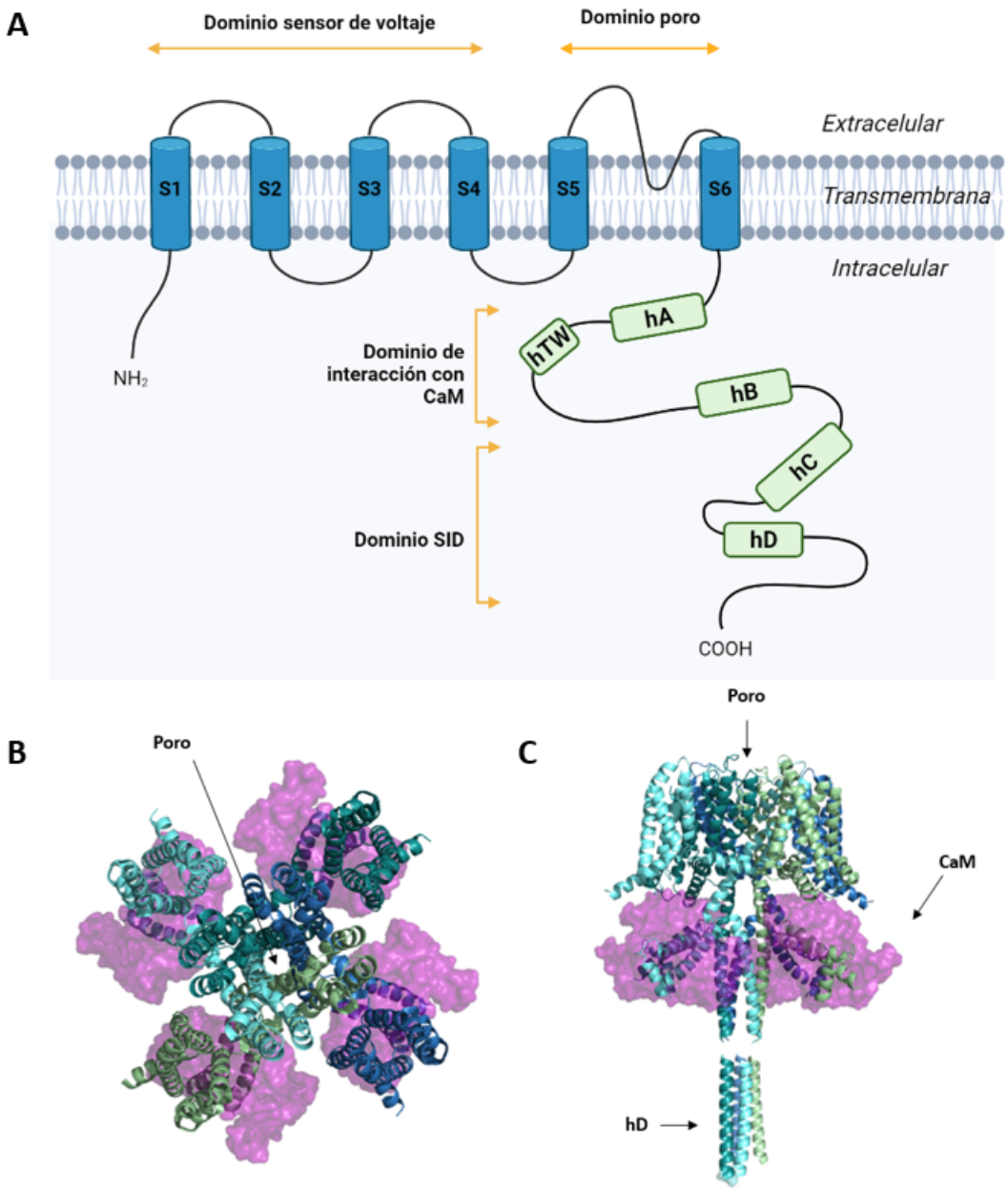

Figura 1. Estructura del canal K_v7.2. **(A)** Principales dominios topológicos y funcionales de un monómero del canal. En azul se observan los seis segmentos transmembrana (S1-S6) y en verde las cinco hélices intracelulares. El dominio funcional del sensor de voltaje se compone de los segmentos S1-S4, el dominio del poro de los segmentos S5-S6, el dominio de interacción de subunidades (SID) de las hélices C-D y el dominio de interacción con calmodulina (CaM) de las hélices A, TW y B. **(B)** Estructura tridimensional del tetrámero de K_v7.2 visto desde arriba y **(C)** desde un lateral. Ambas figuras fueron creadas a partir del fichero PDB 7CR3 mediante Pymol [45]. Cada una de las subunidades del tetrámero se encuentra coloreada en tonos azules y verdes. La calmodulina (CaM), esencial para el funcionamiento del canal, se encuentra representada en color morado y permanece unida a los dominios de interacción de CaM de cada uno de los monómeros. Las hélices D (hD) han sido diseñadas a partir de las predicciones realizadas por AlphaFold2 al no estar experimentalmente resueltas para K_v7.2.

cionamiento [39]. El dominio de interacción con CaM se compone de las hélices intracelulares A-TW-B, mientras que el dominio SID de las hélices C-D.

1.5. Justificación y objetivos del trabajo

A pesar de la incidencia de la epilepsia, el número de mutaciones patológicas conocidas y registradas en bases de datos públicas es muy inferior a lo que se esperaría para una enfermedad de esta magnitud. Este hecho perjudica directamente al diagnóstico de los pacientes que la padecen, así como a la comprensión de la enfermedad o a estudios derivados que se nutren de esta información. En consecuencia, existe un creciente interés en caracterizar y clasificar las mutaciones de los canales $K_v7.2$ según su patogenicidad. Sin embargo, las aproximaciones tradicionales también se han ido alejando de ese objetivo que pretendían perseguir al volverse cada vez más costosas y laboriosas.

El presente trabajo pretende establecer las bases para poner fin a esta situación aplicando métodos computacionales basados en *machine learning* en la predicción de la patogenicidad de mutaciones en KCNQ2, permitiendo así resolver conflictos de variantes ya depositadas en las bases de datos públicas así como predecir mutaciones no conocidas hasta la fecha. A pesar de que estas técnicas son completamente novedosas para el canal $K_v7.2$, estudios similares han demostrado ya su potencial.

Anteriormente, Li *et al.* [46] construyeron una red neuronal para predecir el efecto de variantes de significancia desconocida en el canal de potasio dependiente de voltaje $K_v7.1$, el símil cardíaco de $K_v7.2$. Más recientemente, Xenakis *et al.* [47] diseñaron un modelo computacional para predecir la patogenicidad de variantes en el canal de sodio dependiente de voltaje $Na_v1.7$ donde se incorporaba información atómica de la biomolécula al modelo. Por último, Larrea-Sebal *et al.* [10] consiguieron un nuevo modelo que mejoraba la precisión de otros métodos de diagnóstico disponibles en la predicción de la clínica de variantes con cambio de sentido (*missense variants*) en el receptor de lipoproteínas de baja densidad (LDLr).

Por todo ello, los objetivos propuestos para el presente estudio son: (a) el diseño de una base de datos para KCNQ2 con el mayor número de mutaciones *missense* conocidas hasta la fecha, caracterizándolas mediante descriptores de secuencia y de estructura; (b) la cuantificación de las dificultades del diagnóstico clínico de las epilepsias mediante el estudio de las reclasificaciones de KCNQ2 en ClinVar; (c) el desarrollo de modelos de *machine learning* capaces de predecir

con éxito la patogenicidad de mutaciones en el canal de potasio dependiente de voltaje $K_v7.2$, y (d) otorgar una interpretación biológica a las predicciones de los modelos, ya sea mediante información conocida en la bibliografía o Rosetta [48], un *software* que permite estudiar la estabilidad de mutaciones presentes en macromoléculas biológicas mediante cálculos físicos y funciones energéticas.

2. Materiales y métodos

2.1. Obtención del conjunto de datos de trabajo

Hasta la fecha, más de 1.200 variantes de KCNQ2 han sido anotadas en ClinVar (16/09/2021). Estas variantes están divididas en 6 subclases en base a sus consecuencias moleculares. Estos 6 tipos de mutaciones son: *frameshift*, *missense*, *nonsense*, *splice site*, ncRNA y UTR (*untranslated region*).

Con el objetivo de desarrollar un modelo de *machine learning* capaz de predecir con precisión la patogenicidad de variantes en KCNQ2 es necesario analizar de manera independiente cada una de las subclases. Esto se debe a que un determinado descriptor o característica puede ser relevante para una subclase pero no para el resto, al ser tan diferentes entre ellas [10]. Además de esto, hay que tener en cuenta otras dos consideraciones:

- Algunas subclases están representadas con un bajo número de variantes, por ejemplo, las mutaciones *splice site* para KCNQ2 solo cuentan con 41 muestras. Este reducido espacio muestral dificulta el diseño de un modelo de *machine learning* por ser métodos que funcionan mejor con grandes conjuntos de datos [49].

- La propia naturaleza de algunas subclases de mutaciones impide el diseño de un modelo de clasificación de *machine learning* por no existir más de una clase en su clínica. Por ejemplo, las mutaciones *frameshift* y *nonsense* conllevan un efecto deletéreo en la mayoría de sus variantes, es decir, casi siempre serían clasificadas como variantes "patológicas". Por lo tanto, en estos casos no existiría la clase "benigna". La situación opuesta ocurre con las variantes ncRNA y UTR, donde casi siempre su efecto es "benigno" y no existiría la clase "patológica".

No obstante, la naturaleza y la representación de las mutaciones *missense* de KCNQ2 anotadas en ClinVar no se ven afectadas por ninguno de estos problemas.

Hasta la fecha, 551 mutaciones *missense* han sido registradas en esta base de datos. Estas variantes han sido clasificadas de acuerdo a su patogenicidad de la siguiente manera: 6 benignas o B (*benign*), 12 probablemente benignas o LB (*likely benign*), 254 variantes de significado desconocido o VUS (*uncertain significance*), 38 variantes de interpretación conflictiva o CI (*conflicting interpretation*), 122 probablemente patológicas o LP (*likely pathogenic*) y 141 patológicas o P (*pathogenic*). Sin embargo, algunas de estas variantes se encuentran incluidas en más de una etiqueta (Ej.: la variante R871S se encuentra incluida al mismo tiempo en las etiquetas "*benign*" y "*likely benign*" o la variante R560W en las etiquetas "*pathogenic*" y "*likely pathogenic*").

En primer lugar, y con el objetivo de aumentar el número de variantes en cada clase y facilitar la clasificación, se agruparon las variantes "*benign*" y "*likely benign*" bajo la etiqueta de "*benign_variant*"; y las mutaciones "*pathogenic*" y "*likely pathogenic*" bajo la etiqueta de "*pathogenic_variant*". Así, se obtuvo un primer conjunto de datos formado por 12 variantes benignas y 240 variantes patológicas.

En segundo lugar, y con el fin de aumentar el conjunto de datos de trabajo, se realizó una exhaustiva búsqueda bibliográfica que pretendía resolver alguna de las variantes que habían sido clasificadas como VUS e incorporar algunas adicionales. De manera complementaria, se buscaron nuevas mutaciones *missense* en otras 4 bases de datos: Global Variome Shared, SFARI Gene, Human Gene Mutation Database (HGMD) y RIKEE database (ver Figura S1 del Material Suplementario). De esta manera, se consiguió una base de datos de trabajo formada por 39 variantes benignas y 314 mutaciones patológicas de KCNQ2.

En tercer lugar, las 38 variantes de interpretación conflictiva se separaron para formar parte del conjunto de datos a predecir por los modelos de *machine learning* (ver 2.4.4. Predicción).

Finalmente, se diseñaron las trece características que fueron empleadas para caracterizar las variantes de KCNQ2 recopiladas: el aminoácido previo a la mutación (*initial_aa*), el aminoácido resultante de la mutación (*final_aa*), el dominio topológico de $K_v7.2$ donde tiene lugar la mutación (*topological_domain*), el dominio funcional de $K_v7.2$ donde ocurre la mutación (*functional_domain*), el valor de conservación evolutiva del residuo mutado (*residue_conserv*), la estructura secundaria que afecta la mutación (*secondary_str*); así como los cambios de tamaño

(*d_size*), hidrofobicidad (*d_hf*), volumen (*d_vol*), accesibilidad media al solvente (*d_msa*), carga (*d_charge*), polaridad (*d_pol*) y aromaticidad (*d_aro*) que tienen lugar entre el aminoácido original y el resultante de la mutación (ver Apartado *"Diseño de los descriptores de la base de datos"* del Material Suplementario).

2.2. Estudio de las reclasificaciones de KCNQ2 en ClinVar

Se analizaron únicamente las reclasificaciones de ClinVar por ser la base de datos que más contribuyó al conjunto de datos de trabajo, concretamente un 71.39 % (252 mutaciones de 353). Las variantes se analizaron siguiendo el procedimiento facilitado por Andrew Sharo [5].

Para este fin, es necesario conocer y acceder al servidor FTP de ClinVar. Dentro de él, se encuentran los archivos VCF que contienen una instantánea de todas las variantes existentes en ClinVar y su clasificación en una fecha específica. Se emplea el genoma humano GRCh38 y los años 2019 y 2021 para el análisis. No se seleccionó el año 2020 para el estudio de las reclasificaciones por no haber seguido una tendencia habitual con motivo de la pandemia de COVID-19, según información obtenida de ClinVar Miner.

Una vez descargados los ficheros de ambos años, se filtraron las variantes específicas de KCNQ2. A continuación, se emparejaron las mutaciones de 2019 y 2021. Posteriormente, se comparó la clínica entre las mutaciones comunes en ambas versiones y se estimó la tasa de reclasificación entre estas últimas. Por último, se representaron los datos empleando la librería floWeaver (v2.0.0) [50]. El procedimiento completo se muestra en la Figura S3 del Material Suplementario.

2.3. Obtención de modelos estructurales para el monómero de K_v7.2

La estructura del canal K_v7.2 no está completamente resuelta a día de hoy ya que solo cubre un fragmento de su secuencia proteica (ver Figura 2). Esto supone un gran problema puesto que para caracterizar las mutaciones recopiladas mediante descriptores estructurales (Ej.: estructura secundaria, inestabilidad proteica a partir del estudio de los cambios de energía libre provocados por una mutación o $\Delta\Delta$G) es necesario conocer la estructura en la que se encuentra cada una de ellas.

El empleo de programas informáticos en la obtención de un modelo estructural completo

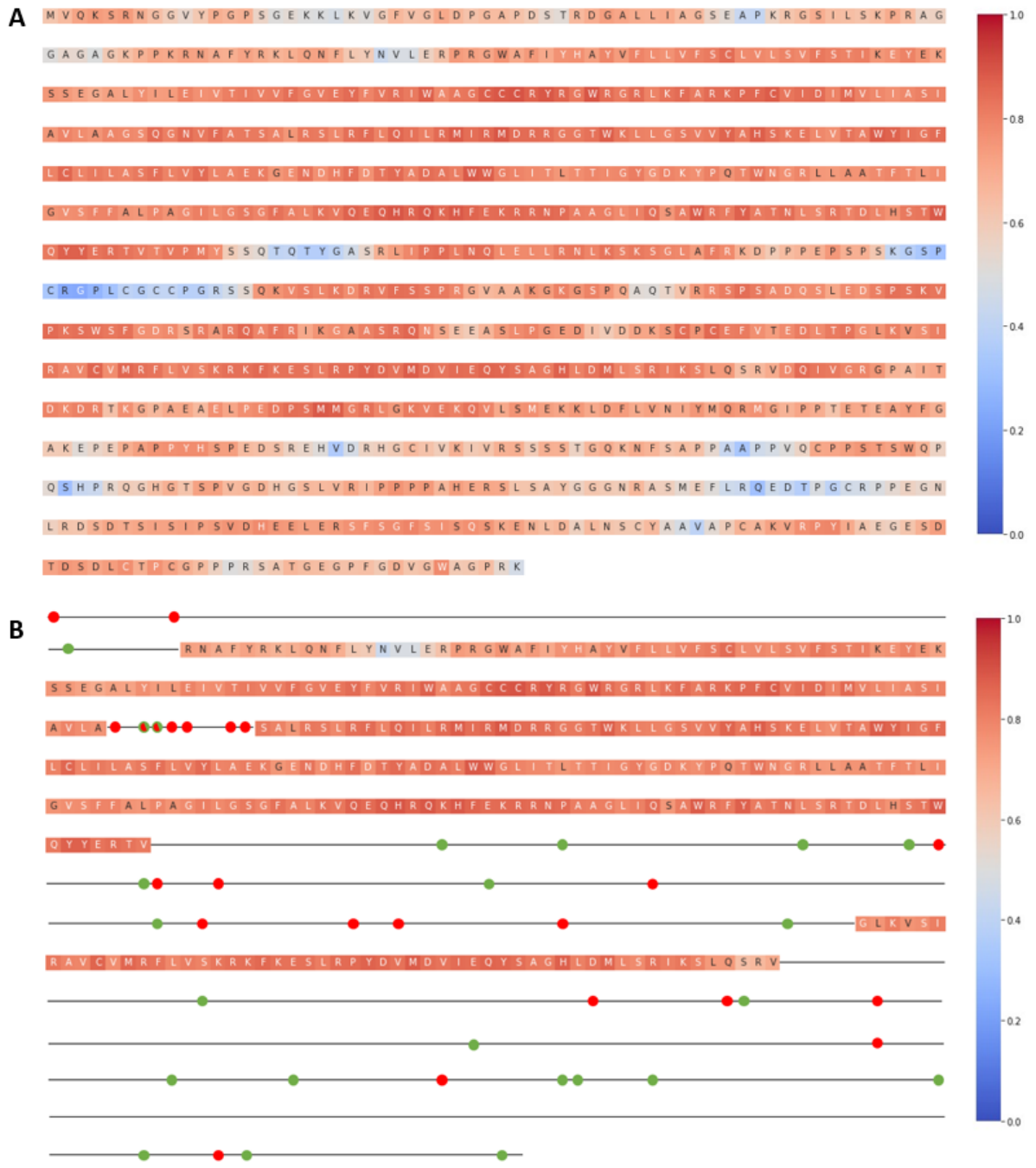

Figura 2. Secuencia completa del canal K_v7.2 frente a la secuencia real de la que se conoce su estructura. **(A)** Secuencia completa de la isoforma 1 de KCNQ2 (UniProt ID: O43526) desde el residuo 1 al 872. Cada uno de los residuos se encuentra coloreado con su valor de conservación evolutiva, correspondiendo el rojo a residuos más conservados (valores próximos a 1) y el azul a residuos menos conservados (valores próximos a 0). **(B)** Secuencia real del canal de la que se conoce su estructura, usando como referencia el PDB 7CR3. Los residuos de los que se conoce su estructura se encuentran coloreados siguiendo el mismo patrón que en (A) y las líneas negras simulan los fragmentos de secuencia de estructura desconocida. Sobre estas últimas, se colorean en verde las posiciones de las 23 mutaciones benignas y, en rojo, las 29 variantes patológicas ubicadas en regiones sin estructura experimental.

para el canal $K_v7.2$ cobra mayor relevancia desde que el 14.73 % (52 de 353) de las mutaciones recopiladas se encuentran en zonas estructuralmente desconocidas que, en términos de clase, corresponden al 58.97 % de las mutaciones benignas (23 de 39) y al 8.92 % (28 de 314) de las mutaciones patológicas. Ante el gran porcentaje de mutaciones benignas ubicadas en estas zonas se desestima la eliminación de estas variantes del conjunto de datos de trabajo ya que se reduciría aún más la baja representatividad de esta clase. Ante esta problemática, se propone modelar mediante programas bioinformáticos aquellas regiones cuya estructura no se ha resuelto experimentalmente.

Para obtener la estructura secundaria de las zonas estructuralmente desconocidas de $K_v7.2$ fue necesario combinar la configuración tridimensional de KCNQ2 resuelta hasta la fecha (PDB ID: 7CR3) con programas bioinformáticos de predicción estructural. Aunque existen diversos programas para tal fin, se decidió trabajar con el metaserver PROTEUS2 [51] por permitir recibir como *input* el fichero PDB y modelar a partir de él mediante homología. A esta predicción se aplica una minimización de la energía por estar relacionada con mejores resultados de precisión.

Por último, para la predicción de la estructura tridimensional (3D) de $K_v7.2$ se emplearon los *softwares* por excelencia en este ámbito: AlphaFold2 [22] y RoseTTAFold [23]; así como el resto de herramientas que sus creadores han puesto a disposición de la comunidad científica.

2.4. Modelos de *machine learning*

Los distintos procedimientos que se muestran a continuación fueron realizados empleando Python (v3.8.8) [52] y las librerías pandas (v1.2.4) [53], NumPy (v1.20.1) [54] y scikit-learn (v0.24.1) [55]. Las visualizaciones se realizaron empleando la librería Plotly (v5.5.0) [56].

2.4.1. Preprocesamiento de los datos

En primer lugar, y debido a las múltiples fuentes que se emplearon para la obtención del conjunto de datos de trabajo, fue necesario homogeneizar los formatos de las variantes así como eliminar posibles duplicados. Adicionalmente, se realizó un control de valores atípicos (*outliers*) y valores perdidos (*missing values*). Como consecuencia, se eliminaron 9 mutaciones patológicas de la base de datos original.

A continuación, las 344 mutaciones restantes se dividieron de manera aleatoria en un con-

junto de datos de entrenamiento (*training set*) y un conjunto de datos de prueba (*test set*) en una proporción 70-30 %, respectivamente. Así, el *training set* estuvo formado por 240 variantes (26 benignas y 214 patológicas) y el *test set* por 104 mutaciones (13 benignas y 91 patológicas).

Como el conjunto de datos de trabajo era bastante desequilibrado (tan solo un 11.34 % de los datos totales son variantes benignas frente al 88.66 % de las patológicas) se decidió aplicar una técnica conocida como sobremuestreo (*oversampling*). Esta técnica funciona mejor cuando el conjunto de datos es pequeño y consiste en aumentar el número de instancias de la clase minoritaria produciendo sintéticamente nuevas muestras [57]. Para ello, se empleó la función "*resample*" del módulo utils de scikit-learn. Tras llevarla a cabo, 81 muestras sintéticas se crearon a partir de las variantes benignas ya presentes en el conjunto de datos de entrenamiento. De esta manera, se consiguió un conjunto de datos de entrenamiento con 321 mutaciones (107 benignas y 214 patológicas).

A continuación, se realizó una codificación de los descriptores cualitativos (Ej.: *d_charge*, *topological_domain*... etc.) así como de la etiqueta. En primer lugar, se realizó una transformación de los valores de estos descriptores empleando la clase "*LabelEncoder*" del módulo preprocessing de scikit-learn. Posteriormente, se aplicó un esquema de codificación de esos números mediante la clase "*OneHotEncoder*" también del módulo preprocessing de scikit-learn con el fin de evitar relaciones numéricas inexistentes. Por último, la etiqueta de cada variante fue codificada de manera ordinal de tal manera que la clase "benigna" fuese igual a 0 y la "patológica" igual a 1.

Por último, se realizó una visualización de los datos en 2D y 3D mediante un Análisis de Componentes Principales (PCA) empleando la clase "*PCA*" del módulo decomposition de scikit-learn y el módulo express de Plotly.

2.4.2. Diseño, entrenamiento y optimización de los modelos

Se decidió entrenar tres modelos sencillos que, por bibliografía, funcionaban correctamente en la predicción de variantes proteicas [58]. Estos modelos son la Regresión Logística con Penalización Lasso (L1), el *Support Vector Classifier* (SVC) y *Random Forest*. Todos estos modelos fueron construidos y entrenados empleando scikit-learn.

Para entrenar cada uno de los modelos se creó un *pipeline* con la clase "*Pipeline*" del módulo pipeline, que permite aplicar de manera secuencial una serie de transformaciones y un estimador final. Sobre este *pipeline* se añadió la clase "*StandardScaler*" del módulo preprocessing, que

permite estandarizar los datos que se introducen al modelo para evitar sesgos en la clasificación. Los parámetros modificados de cada uno de los modelos se recopilan en la Tabla 1. En todos ellos se aseguró la reproducibilidad de los resultados mediante un "*random_state*".

El ajuste de los hiperparámetros se realizó mediante la clase "*GridSearchCV*" del módulo model_selection. Para ello, se dividieron los datos en 10 subconjuntos y se repitió el proceso en 10 ocasiones. Previamente, se realizó un filtrado de características basado en un F-score (ANOVA) mediante la clase "*SelectKBest*" del módulo feature_selection implementado también en scikit-learn. Los *pipelines* finales para cada uno de los modelos así como el número de características seleccionadas (k) se muestran en la Tabla S8 del Material Suplementario.

Modelos	Parámetros
L1	`solver = 'liblinear', penalty = 'l1', fit_intercept = False, C = 2.2, class_weight = {0: 3, 1:1}, random_state = 1`
SVC	`kernel = 'sigmoid', coef0 = 0.3, class_weight = {0: 3, 1:2}, probability = True, random_state = 0`
RF	`criterion = 'entropy', max_depth= 3, min_samples_split = 3, class_weight = {0: 9, 1:1}, n_estimators = 145, max_features = 'sqrt', max_samples =15, bootstrap = True, random_state = 12`

Tabla 1. Parámetros modificados de los modelos empleados. En la tabla se recogen los parámetros modificados para el modelo de Regresión Lineal con penalización Lasso (L1), el modelo de *Support Vector Classifier* (SVC) y el de *Random Forest* (RF). El resto de parámetros que no se muestran mantuvieron su valor por defecto.

2.4.3. Evaluación de los modelos

Como métricas de evaluación y comparación de los modelos se emplearon el área bajo la curva ROC (*Receiver Operating Characteristic curve*) o AUC-ROC (*Area Under the ROC curve*); y las matrices de confusión, al ser útiles en situaciones de desequilibrio de clases. A partir de estas últimas se calcularon la especificidad y la sensibilidad de los modelos.

Por un lado, la sensibilidad es la métrica que evalúa la capacidad del algoritmo para predecir correctamente los verdaderos positivos (ver Ecuación 1). En el caso particular del estudio sería la capacidad del modelo para distinguir las variantes patológicas. Por el otro lado, la especificidad

es la métrica que evalúa la capacidad del algoritmo para diferenciar correctamente los verdaderos negativos (ver Ecuación 2). En este caso, la capacidad del modelo para distinguir las variantes benignas.

$$Sensibilidad = \frac{VP}{VP + FN} * 100 \quad (1)$$

$$Especificidad = \frac{VN}{VN + FP} * 100 \quad (2)$$

Siendo *"VP"*, *"FP"*, *"VN"*, *"FN"* los verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos, respectivamente.

Por último, la curva ROC de un modelo es un gráfico que permite ver cómo varía a distintos umbrales su Tasa de Verdaderos Positivos (TVP), es decir, la capacidad para distinguir los VP (sensibilidad); frente a su Tasa de Falsos Positivos (TFP o $1 - sensibilidad$) [59]. El AUC-ROC sería el valor numérico que resume la capacidad predictiva de ese modelo y se obtiene al integrar el área bajo la curva ROC. El AUC-ROC puede tomar valores entre 0 y 1, donde un valor de 0 indica un modelo completamente inexacto y un valor de 1 refleja una prueba perfectamente precisa [60]. En general, un AUC-ROC de 0.5 sugiere que el modelo no tiene discriminación (es decir, no distinguiría entre mutaciones benignas y patológicas), de 0.7 a 0.8 se considera aceptable, de 0.8 a 0.9 se considera excelente, y más de 0.9 se considera sobresaliente [61].

2.4.4. Predicción

Treinta y ocho variantes de interpretación conflictiva de KCNQ2 anotadas en ClinVar fueron seleccionadas para poner a prueba la precisión de los modelos entrenados. Con el fin de obtener resultados más robustos, se decidió combinar la predicción de todos los modelos.

2.5. Interpretación biológica de las predicciones y cálculo de ΔΔG

Aquellas variantes de interpretación conflictiva donde las predicciones de los tres modelos fueran idénticas y cuya estructura fuera conocida se seleccionaron para analizarlas con más detalle.

En primer lugar, se recopiló información biológica sobre las localizaciones de las variantes (dominios funcionales y topológicos de $K_v7.2$ afectados) así como la conservación evolutiva de sus posiciones. Posteriormente, y debido a que se seleccionaron las variantes con estructura conocida, se pudo calcular la ΔΔG mediante Rosetta (v3.13) [48], un *software* de modelización

macromolecular que evalúa la plausibilidad física de las biomoléculas. Este cálculo permite conocer la inestabilidad o estabilidad proteica que genera una mutación a partir de la variación del cambio en la energía libre de Gibbs, denominada ΔΔG (ver Ecuación 3). Desde que menor energía implica mayor estabilidad del sistema y desde que ΔΔG se define como:

$$\Delta\Delta G = \Delta G_{mutante} - \Delta G_{nativo} \tag{3}$$

se considera que una mutación es desestabilizante cuando ΔΔG es positiva, es decir, cuando $\Delta G_{mutante} > G_{nativo}$; y estabilizante cuando es negativa, $G_{mutante} < G_{nativo}$ [62].

Para ello, se eligieron los paquetes "*Flex_ddG*" de RosettaScripts [63] y "*MPddG*" de PyRosetta [64]. Sin embargo, se empleó una versión mejorada de "*MPddG*" diseñada previamente en el grupo de investigación puesto que el protocolo original sobreestimaba las interacciones de Van der Vaals [62]. Asimismo, para el cálculo final de la ΔΔG se siguió el mismo protocolo detallado por la autora.

Finalmente, y teniendo en cuenta toda la información considerada, se estudió la factibilidad biológica de las predicciones de los modelos de *machine learning* entrenados.

3. Resultados

3.1. Análisis de las reclasificaciones de KCNQ2

Entre el 19 de diciembre de 2019 y el 4 de diciembre de 2021 existen 784 variantes comunes de KCNQ2 en ClinVar. Estas mutaciones han sido anotadas siguiendo la clínica recomendada por los estándares de la ACGM-AMP [4]. Entre ambos años, solo el 9.82 % (77 de 784) de estas variantes han sido reclasificadas y actualizadas en ClinVar (ver Figura 3). Entre las reclasificaciones más frecuentes se encuentran los cambios de LP a CI (11 de 77), VUS a CI (11 de 77), P a LP (10 de 77) y LB a B (10 de 77). En la Tabla S6 del Material Suplementario se desglosan por etiqueta clínica todas las reclasificaciones de KCNQ2 entre 2019 y 2021.

Teniendo en cuenta la certeza de las reclasificaciones, el 41.56 % (32 de 77) de las variantes reclasificadas de KCNQ2 se actualizaron a una categoría de mayor certeza (VUS a P/LP, VUS a LB, LP a P, LB a B, CI a B/LB o CI a P/LP), el 55.84 % (43 de 77) a una categoría de menor

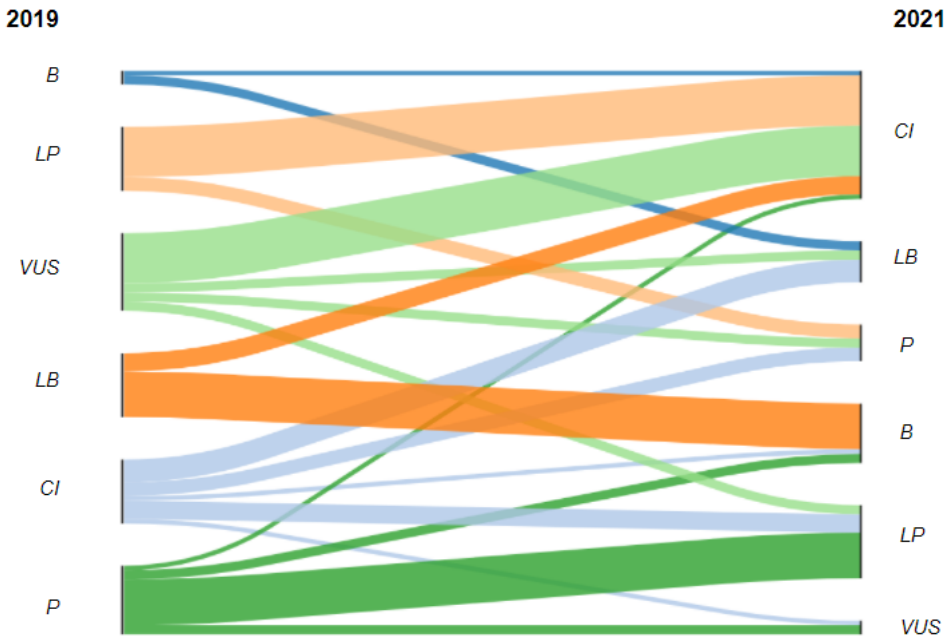

Figura 3. Diagrama de Sankey de las variantes de KCNQ2 reclasificadas en ClinVar entre los años 2019 y 2021. En la columna de la izquierda se representan las etiquetas clínicas de las variantes de 2019 y en la columna de la derecha las etiquetas clínicas de las variantes de 2021. El tamaño de las barras es proporcional al número de veces que sucede ese cambio. La etiqueta P hace referencia a las variantes patológicas, LP a las probablemente patológicas, VUS a las variantes de significado desconocido, CI a las variantes de interpretación conflictiva, LB a las variantes probablemente benignas y B a las benignas.

certeza (B a LB/CI, P a LP/CI/VUS, LP a CI, LB a CI, VUS a CI y CI a VUS) y solo un 2.60 % (2 de 77) a una categoría opuesta (P a B).

Por último, y siguiendo el mismo procedimiento que para KCNQ2, se estimaron las tasas de reclasificación (tr) de las variantes anotadas en ClinVar para el resto de miembros de la familia de los canales de potasio dependientes de voltaje (K_v7.1-5). Estos canales se codifican a partir de los genes KCNQ1-KCNQ5, de manera respectiva. Los resultados obtenidos muestran que la tasa de reclasificación es mayor para KCNQ5 (18.75 %) y KCNQ1 (15.15 %) que para KCNQ2 (9.82 %). Sin embargo, la tasa de reclasificación de KCNQ2 es mayor que para KCNQ3 (9.05 %) y KCNQ4 (6.54 %). Dicho de otra manera:

$$tr_{KCNQ5} > tr_{KCNQ1} > tr_{KCNQ2} > tr_{KCNQ3} > tr_{KCNQ4}$$

Sin embargo, es importante destacar que el número de variantes comunes entre 2019 y 2021 a partir del cual se han estimado las tasas de reclasificación es muy diferente para cada miembro de la familia KCNQ, al no estar igual de representados en ClinVar (ver Tabla 2 y Material

Gen	N.º total de variantes	Clínica conservada	Reclasificadas
KCNQ1	944	84.85 % (801/944)	15.15 % (143/944)
KCNQ2	784	90.18 % (707/784)	9.82 % (77/784)
KCNQ3	486	90.95 % (442/486)	9.05 % (44/486)
KCNQ4	107	93.46 % (100/107)	6.54 % (7/107)
KCNQ5	16	81.25 % (13/16)	18.75 % (3/16)

Tabla 2. Variantes reclasificadas y no reclasificadas en ClinVar de los miembros de la familia KCNQ entre los años 2019 y 2021. Se puede observar como los canales codificados por los genes KCNQ1 y KCNQ2 ($K_v7.1$ y $K_v7.2$, respectivamente) han sido los más estudiados y anotados en ClinVar. Sin embargo, el resto de miembros de la familia KCNQ han sido investigados a niveles inferiores. Las mayores tasas de reclasificación se encuentran en $K_v7.1$ y $K_v7.5$.

Suplementario).

3.2. Interpretación de los modelos estructurales obtenidos: estructura secundaria y terciaria de la secuencia completa de KCNQ2

Por un lado, la predicción de la estructura secundaria de la secuencia completa de KCNQ2 mediante PROTEUS2 se obtuvo mediante el empleo de 6 secuencias homólogas (PDB IDs: 1ORQ, 1BL8, 2A79, 2A9H, 1F6G y 2P7T). Todas las secuencias empleadas para el modelado por homología corresponden a estructuras experimentales de canales de potasio. La confianza media de la predicción fue del 81.22 %, relacionándose con una calidad media - alta. De los 872 residuos, el 59.86 % (522 de 872) fue predicho como hélices superenrolladas (*coiled coils*), el 21.44 % (187 de 872) como α-hélices, el 16.40 % (143 de 862) como hélices transmembrana y el 2.30 % (20 de 872) como β-láminas (ver Figura S2 del Material Suplementario).

Debido a la confianza media de la predicción, así como a las puntuaciones individuales de los residuos ubicados en zonas estructuralmente desconocidas, se decidió incorporar la estructura secundaria como descriptor estructural de las variantes de KCNQ2 recopiladas.

Por el otro lado, y a diferencia de lo que ocurría con PROTEUS2, la predicción de la estructura tridimensional de la secuencia completa de KCNQ2 mediante AlphaFold2 y RoseTTAFold otorgó unos resultados de calidad media-baja. Mientras que el mejor modelo de AlphaFold2 presentaba una confianza del 59.07 %, la mejor opción con RoseTTAFold contaba con una confianza del 34.99 %. Al analizar el mejor modelo de AlphaFold2 en detalle se observó que las

zonas de interés se predecían, en su gran mayoría, con una prueba de diferencia de distancia local predicha (pLDDT) baja ($70 > pLDDT > 50$) o muy baja ($pLDDT < 50$) de igual manera a como estaban ya predichas para O43526 en la Base de Datos de Estructuras proteicas de AlphaFold [65]. A pesar de que los modelos de RoseTTAFold se descartaron en el primer intento, se intentó predecir la estructura tridimensional mediante las herramientas disponibles en Robetta Server. No obstante, ninguna de ellas consiguió mejorar las confianzas obtenidas mediante el propio *software*.

Por lo tanto, y debido a la calidad de los modelos obtenidos a partir AlphaFold2 y RoseTTAFold, se desestimó la caracterización tridimensional de las mutaciones de KCNQ2 al no poder emplear sus modelos en el diseño de descriptores estructurales fiables.

3.3. Modelos computacionales

3.3.1. Calidad de los modelos y visualización espacial de los datos

El modelo de Regresión Logística con penalización Lasso (L1) y selección de características mediante F-score (ANOVA) ha clasificado correctamente el 89.72 % (96 de 107) de las variantes benignas y el 77.10 % (165 de 214) de las variantes patológicas durante el entrenamiento; y el 84.62 % (11 de 13) de las variantes benignas y el 76.92 % (70 de 91) de las variantes patológicas en el *test*. De manera similar, el modelo de *Support Vector Classifier* (SVC) con selección de características mediante F-score (ANOVA) ha clasificado correctamente el 80.37 % (86 de 107) de las variantes benignas y el 75.23 % (161 de 214) de las variantes patológicas durante el entrenamiento; y el 76.92 % (10 de 13) de las variantes benignas y el 78.02 % (71 de 91) de las variantes patológicas en el *test*. Por último, el modelo de *Random Forest* (RF) ha clasificado correctamente el 81.32 % (87 de 107) de las variantes benignas y el 83.74 % (179 de 214) de las variantes patológicas durante el entrenamiento; y el 76.92 % (10 de 13) de las variantes benignas y el 81.32 % (74 de 91) de las variantes patológicas en el *test*. En términos de sensibilidad y especificidad, el modelo L1 ha sido el método con mayor capacidad de distinguir las mutaciones benignas, al presentar una especificidad media del 87.17 %; y el modelo RF el método con mayor capacidad de distinguir las mutaciones patológicas, al presentar una sensibilidad media del 82.53 %. Los valores de AUC-ROC calculados, situados próximos a 0.8, muestran la calidad de los modelos (ver Tabla 3).

Al estudiar la relevancia de los descriptores en la toma de decisiones de los tres modelos

Modelo	Set	Especificidad	Sensibilidad	AUC-ROC
L1 (k=15)	*Train*	89.72 %	77.10 %	0.83
	Test	84.62 %	76.92 %	0.81
SVC (k=10)	*Train*	80.37 %	75.23 %	0.78
	Test	76.92 %	78.02 %	0.77
RF	*Train*	81.32 %	83.74 %	0.82
	Test	76.92 %	81.32 %	0.79

Tabla 3. Tabla resumen de los valores de especificidad, sensibilidad y AUC-ROC obtenidos para los modelos entrenados. Se observa como el modelo de Regresión Logística con penalización Lasso (L1) y selección de características ($k = 15$) es el método con mayor especificidad, es decir, capacidad para diferenciar las variantes benignas; mientras que el modelo de *Random Forest* (RF) es el modelo con mayor sensibilidad, es decir, capacidad para diferenciar las variantes patológicas. El modelo de *Support Vector Classifier* (SVC) con selección de características ($k = 10$) presenta unos valores de especificidad y sensibilidad intermedios entre ambos. Los valores de AUC-ROC obtenidos reflejan la calidad de los modelos.

se observa que 6 características se repiten entre todos ellos y son: el valor de la conservación evolutiva del residuo mutado (*residue_conserv*), que la estructura secundaria afectada por la mutación sea una hélice de membrana (*secondary_str = "membrane_helix"*), que el cambio de polaridad consecuencia de la mutación sea de aminoácido polar a no polar (*d_pol = "p_to_np"*) así como de no polar a no polar (*d_pol = "np_to_np"*), que topológicamente la mutación se encuentre en el citoplasma o en el dominio del poro (*topological_domain = "Cytoplasmic"/"Pore"*).

3.3.2. Visualización espacial de los datos mediante PCA

Con el fin de entender los resultados obtenidos de los modelos entrenados, se decide representar las variantes de KCNQ2 empleadas en el estudio mediante un Análisis de Componentes Principales (PCA).

Por un lado, la visualización de las dos componentes principales explica el 77.10 % de la variabilidad total de los datos. Sin embargo, esta representación no permite distinguir espacialmente las clases 0 (variantes benignas) y 1 (variantes patológicas) puesto que la mayoría de sus instancias se encuentran superpuestas. Por el otro lado, la visualización de las tres componentes principales explica el 95.90 % de la variabilidad total de los datos. A diferencia de la primera representación, se pueden apreciar dos clústeres ligeramente señalados en el espacio tridimensional. Sin embargo, y de manera similar a como ocurría con el primer PCA, existe una serie

de mutaciones benignas que se encuentran ubicadas en el clúster espacial donde se encuentran la mayoría de las variantes patológicas (ver Figura 4).

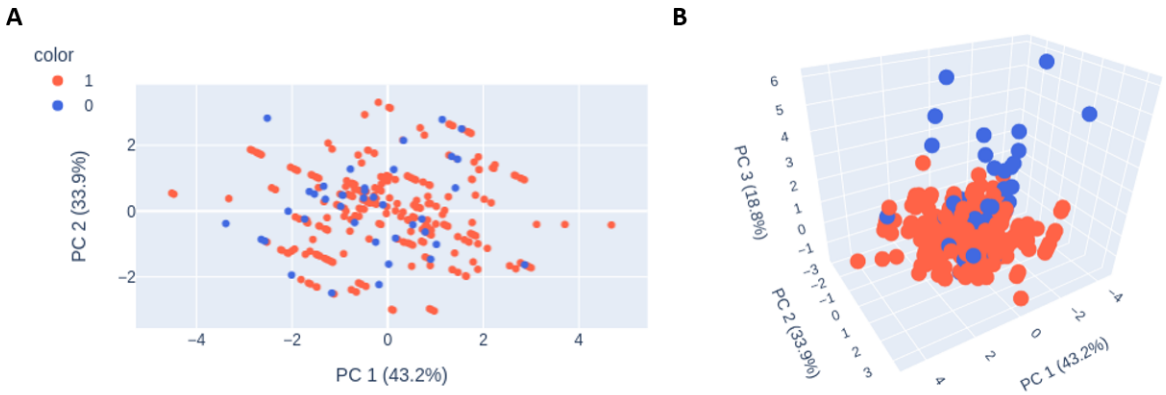

Figura 4. PCA de las variantes benignas y patológicas de KCNQ2. **(A)** La visualización de las dos componentes principales (PC1 y PC2) explica un 77.10 % de la variabilidad de los datos, mientras que **(B)** la visualización de las tres componentes principales (PC1, PC2 y PC3) explica el 95.90 % de la variabilidad total de los datos. Sin embargo, tanto en (A) como en (B) se observa el mismo problema y es que a pesar de que algunas mutaciones benignas (coloreadas en azul) se separan espacialmente de las variantes patológicas (coloreadas en rojo) existe un porcentaje de estas que se sitúan espacialmente junto a la clase opuesta.

Por lo tanto, y a pesar de que el 95.90 % de la variabilidad total de las variantes de KCNQ2 recopiladas se explica con solo tres componentes principales (PC1, PC2 y PC3), existe una serie de mutaciones benignas que no podrían diferenciarse, a partir de los descriptores considerados en el estudio, de ciertas variantes patológicas. Por ende, los modelos de *machine learning* entrenados tampoco serían capaces de distinguir este tipo de mutaciones.

3.4. Predicción de la patogenicidad de las variantes conflictivas, cálculo de ΔΔG e interpretación biológica de los resultados.

De las 38 variantes conflictivas que se predijeron, los modelos entrenados tomaban una decisión inequívoca en 24 de ellas. De estas 24 variantes, solo 16 se ubicaban en zonas estructuralmente conocidas y podía calcularse la ΔΔG (ver Tablas S9-S11 del Material Suplementario). Por último, y cumpliendo uno de los objetivos establecidos en el presente trabajo, se seleccionaron 4 variantes con el fin de otorgarles una interpretación biológica más extendida (ver Figura 5).

En primer lugar, la mutación V182M se localiza en el dominio topológico del segmento S3

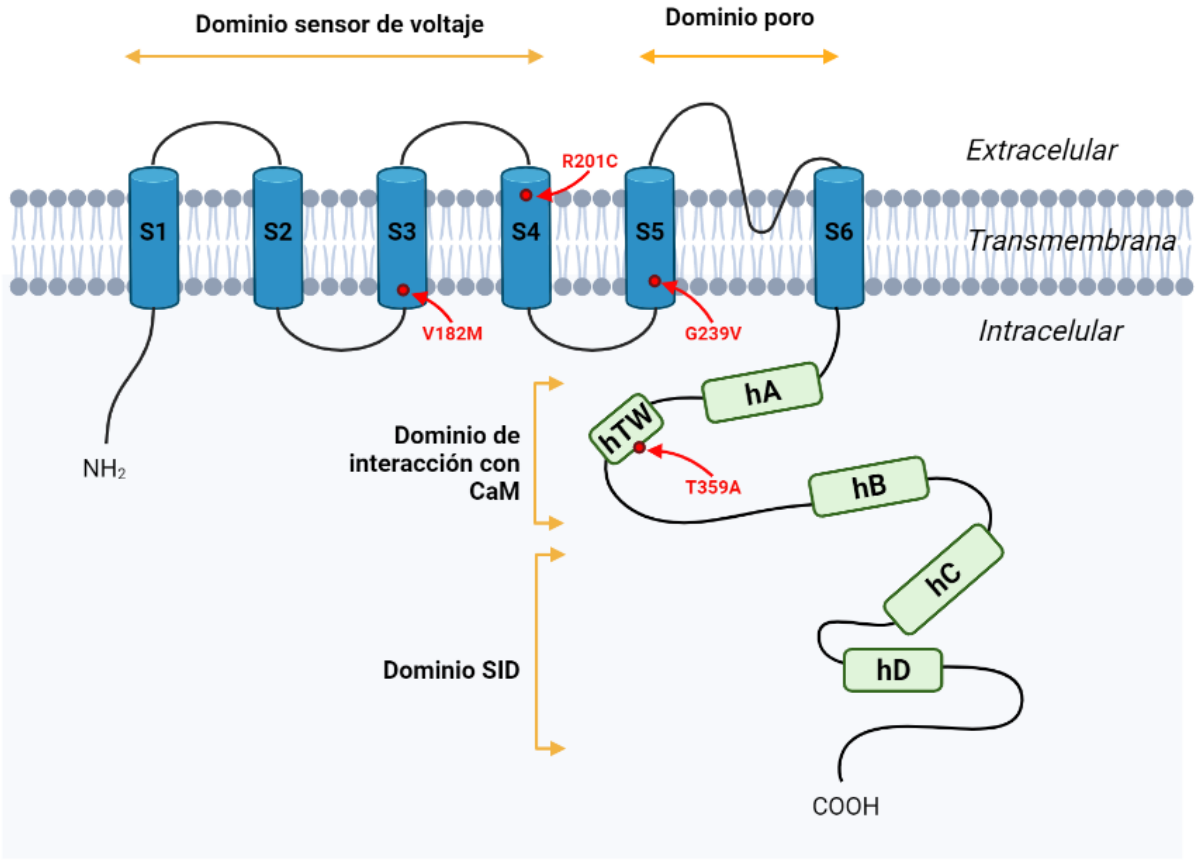

Figura 5. Localización de las variantes sobre las que se realiza la interpretación biológica y el cálculo de ΔΔG. Las variantes elegidas para su interpretación son: V182M, predicha como benigna (clase 0) por los tres modelos y situada en el segmento transmembrana S3; y R201C, G239V y T359A, todas ellas predichas como variantes patológicas (clase 1) y situadas en S4, S5 y hTW de manera respectiva.

como parte del sensor de voltaje. A pesar de la conservación evolutiva de la posición donde ocurre la mutación (con_{182} = 0.79708), los tres algoritmos entrenados (L1, SVC y RF) han predicho que se trata de una variante benigna de KCNQ2. Al interpretar el valor de ΔΔG estimado, se observa que V182M tendría un efecto estabilizante en su entorno tridimensional ($\Delta\Delta G_{V182M}$ = -7.06). Además, V182M no experimenta ningún cambio de polaridad ni de carga tras la mutación, desde que la valina (V) y la metionina (M) son aminoácidos neutros y no-polares. Del mismo modo, la mutación R201C se localiza en el dominio funcional del sensor de voltaje en una posición aún más conservada que la anterior (con_{201} = 0.84353). A diferencia de V182M, R201C se encuentra en el dominio topológico del segmento S4 y es predicha de manera inequívoca por los tres algoritmos como una variante patológica de KCNQ2. Además, esta variante experimenta un cambio de carga tras la mutación, al ser la arginina (R) un aminoácido positivo y la cisteína (C) un residuo neutro. Al analizar el valor de ΔΔG estimado, se observa

Mutación	Patogenicidad predicha	Dominio topológico	Dominio funcional	Conservación evolutiva	ΔΔG
V182M	Benigna	S3	Sensor de voltaje	0.79708	-7.06
R201C	Patológica	S4	Sensor de voltaje	0.84353	2.69
G239V	Patológica	S5	Poro	0.83092	31.13
T359A	Patológica	hTW	Interacción con CaM	0.83093	0.23

Tabla 4. Información biológica de las mutaciones seleccionadas y su valor de ΔΔG. En la tabla se recogen las 4 mutaciones elegidas para el análisis, así como la etiqueta con la que fueron predichas de manera inequívoca por los tres algoritmos. Los valores de conservación evolutiva se representan en una escala del 0 al 1, siendo 1 el valor de mayor conservación. En la columna de ΔΔG, valores positivos informan de una mutación desestabilizante, mientras que valores negativos de una mutación estabilizante. El valor de ΔΔG para la variante T359A se considera de un efecto neutral.

que R201C tendría un efecto desestabilizante en su entorno tridimensional ($\Delta\Delta G_{R201C} = 2.69$) lo que provocaría una clínica patológica al desestabilizar el sensor de voltaje en el segmento S4.

Por último, las variantes G239V y T359A también han sido predichas como mutaciones patológicas por los tres algoritmos. Además, ambas variantes conllevan un cambio de polaridad desde que la glicina (G) y la treonina (T) son aminoácidos polares y la valina (V) y la alanina (A) son residuos no-polares. Mientras que G239V parece tener una interpretación biológica sencilla al encontrarse en el dominio topológico del S5 en una posición altamente conservada ($con_{239} = 0.83092$) y desestabilizando el poro en gran medida ($\Delta\Delta G_{G239V} = 31.13$); la variante T359A tiene un efecto neutral en la estabilidad del sistema ($\Delta\Delta G_{T359A} = 0.23$). No obstante, esta variante es predicha como mutación patológica en KCNQ2 porque, posiblemente, al encontrarse en una posición altamente conservada de la hTW ($con_{359} = 0.83093$) se vería afectada la interacción con CaM. En la Tabla 4 se muestran los resultados presentados.

4. Discusión

Los recientes avances y la reducción de los costes de las tecnologías de secuenciación masiva han permitido su uso en aplicaciones clínicas, entre ellas la detección de nuevas variantes genéticas [66–68]. Sin embargo, solo una pequeña fracción de las mutaciones detectadas por secuenciación son causantes de enfermedad lo que obliga a validarlas experimentalmente [69, 70].

Debido a que los ensayos funcionales serían muy laboriosos y costosos, los métodos de *machine learning* para la predicción de la patogenicidad de las variantes proteicas han cobrado importancia en los últimos años [5, 71, 72]. Esta situación aparece también en el canal de potasio dependiente de voltaje $K_v7.2$ (KCNQ2), donde mutaciones en sus subunidades se relacionan directamente con la Epilepsia Neonatal Benigna Familiar (BFNE) y la Encefalopatía Epiléptica (EE) [38]. A pesar del tiempo que se lleva estudiando el canal $K_v7.2$, todavía no se había desarrollado un modelo de *machine learning* para la predicción de la patogenicidad de sus variantes. En este contexto donde la interpretación clínica de las variantes es la columna vertebral de la medicina actual y donde el diagnóstico temprano de BFNE/EE permitiría prever un resultado favorable en los pacientes y un riesgo bajo de epilepsia posterior, surge la motivación principal del presente estudio [33, 73].

En primer lugar, el presente trabajo tuvo como objetivo el diseño de una base de datos para KCNQ2 con el mayor número posible de mutaciones *missense*, por ser el tipo de mutación más común de variantes genéticas codificantes y por ser de vital importancia en los estudios genéticos y el diagnóstico clínico [69].

Para cumplir el primer objetivo, se combinó información de varias bases de datos con fuentes bibliográficas. Una vez recopiladas el mayor número posible de mutaciones *missense*, cada variante fue caracterizada mediante descriptores de secuencia y estructurales. Como descriptores de secuencia se emplearon el aminoácido previo a la mutación, el aminoácido resultante de la mutación, el dominio topológico y funcional de $K_v7.2$ donde tiene lugar la mutación, el valor de conservación evolutiva del residuo mutado; así como los cambios de tamaño, hidrofobicidad, volumen, accesibilidad media al solvente, carga, polaridad y aromaticidad. Se emplearon estos descriptores debido a que están bien establecidos y no necesitan de *software* externo a excepción de la conservación evolutiva [10].

Adicionalmente, se propuso caracterizar las mutaciones de KCNQ2 mediante descriptores estructurales debido a que, previamente, se había demostrado que pueden mejorar la capacidad predictiva de los modelos [74]. Por un lado, y debido a la calidad de las predicciones de PROTEUS2, se pudo incorporar como descriptor la estructura secundaria afectada por cada una de las variantes de $K_v7.2$ recopiladas. Por el otro lado, no se consiguieron diseñar descriptores tridimensionales de las variantes de KCNQ2 debido a la falta de estructura completa del canal y a la baja calidad de las predicciones de AlphaFold2 y RoseTTAFold, cuyos mejores modelos obtuvieron una confianza del 59.07 % y 34.99 %, respectivamente. A pesar de que se

analizó cada predicción de forma individual, ninguna de ellas permitió completar alguna de las regiones estructuralmente desconocidas de KCNQ2 debido a la baja o muy baja calidad de las predicciones en esas zonas.

Las "malas" predicciones de los dos *softwares* por excelencia en el ámbito de la bioinformática estructural podría explicarse si las zonas de $K_v7.2$ cuya estructura se desconoce fueran regiones intrínsecamente desordenadas (IDRs). De hecho, en los modelos de AlphaFold2 estas IDRs se identifican por un $pLDDT < 50$ y, a menudo, se representan gráficamente como largos filamentos [20]. Este es el escenario ante el que pudo haberse encontrado según los modelos obtenidos. Sin embargo, la única región que parece ser un IDR en los canales KCNQ según estudios experimentales sería la región que precede a la hélice D, es decir, una pequeña fracción de la totalidad de residuos ubicados en zonas estructuralmente desconocidas del canal $K_v7.2$ [75–77]. Por lo tanto, debido a la falta de información sobre IDRs en KCNQ2 y a que continúan siendo un interrogante en términos de funcionalidad, se decidió desestimar los modelos de AlphaFold2 y RoseTTAFold hasta futuros avances en el campo de la biología estructural [78].

Por lo tanto, un total de 353 variantes de KCNQ2 fueron caracterizadas empleando información de la secuencia de sus aminoácidos así como información de la estructura secundaria. Sin embargo, y a pesar de la relevancia de la estructura terciaria en la función biológica de las proteínas, no se consiguió incorporar ningún descriptor 3D de las variantes debido al escenario estructural que experimenta el canal $K_v7.2$ [79].

En segundo lugar, el presente trabajo tuvo como objetivo la cuantificación de las dificultades del diagnóstico clínico de las epilepsias mediante el estudio de las reclasificaciones de KCNQ2 en ClinVar.

Por un lado, los resultados mostraron que en un rango de dos años el 9.82 % de las variantes de KCNQ2 han sido reclasificadas, habiéndose considerado como rangos habituales unas tasas de reclasificación de entre el 6.40 % y el 15.00 % [80]. Por esta razón, las tasas de reclasificación estimadas para KCNQ2 y el resto de miembros de su familia se han considerado como normales, a excepción de KCNQ5 cuyo valor excede dicho rango ($tr_{KCNQ5} = 18{,}75\,\%$). Sin embargo, la tasa de reclasificación en KCNQ5 no debería de tenerse en cuenta puesto que solo presenta 16 variantes anotadas en ClinVar entre los años 2019 y 2021. No obstante, estudios previos han registrado situaciones más extremas donde las tasas de reclasificación superaban el 35 % de los casos en un rango de 5 y 10 años [29, 81].

Por el otro lado, y teniendo en cuenta la etiqueta clínica, las variantes clasificadas como LP y VUS tuvieron la mayor tasa de reclasificación en KCNQ2 (14.30 % en ambos casos). En primer lugar, los resultados concuerdan con estudios anteriores donde las variantes VUS fueron las mayormente reclasificadas con el tiempo [26, 80, 81]. La mayor tasa de reclasificación de las variantes VUS se debe a que, con frecuencia, es difícil establecer con certeza la patogenicidad de variantes recién descubiertas hasta que no se obtienen más pruebas clínicas o ensayos experimentales y, como consecuencia, se clasifican en un primer momento como VUS a menos que sean muy recurrentes [5, 69]. Asimismo, se recomienda que los pacientes diagnosticados con una variante VUS realicen un seguimiento cada 2-3 años con su laboratorio para conocer el estado actual de su variante, favoreciéndose aún más este tipo de reclasificaciones [26]. En segundo lugar, las reclasificaciones de las variantes LP entre 2019 y 2021, así como el resto de etiquetas clínicas reclasificadas, se encuentran dentro de la normalidad según la bibliografía consultada.

Por lo tanto, y a pesar de los desafíos diagnósticos de la epilepsia, las tasas de reclasificación estimadas para KCNQ2 se encuentran dentro de los rangos habituales [82]. No obstante, estos valores son extremadamente dependientes del tipo de proteína afectada y los años seleccionados para estudio [80]. Como última aclaración, no se deberían de interpretar las reclasificaciones de variantes como un aspecto negativo en el diseño de un modelo de *machine learning* desde que son esenciales para seguir desarrollando nuestra comprensión de los genes y sus condiciones asociadas [26].

En tercer lugar, y una vez generada la base de datos de trabajo, el presente estudio tuvo como objetivo el desarrollo de modelos de *machine learning* capaces de predecir con éxito la patogenicidad de mutaciones en el canal de potasio dependiente de voltaje $K_v7.2$.

Debido a que la aplicación de técnicas de *machine learning* en la predicción de variantes patológicas era una meta novedosa para KCNQ2, se diseñaron tres modelos sencillos para proporcionar una primera aproximación en este ámbito. Cabe destacar que se desestimó la idea de aplicar técnicas de *deep learning* sobre las variantes recopiladas ya que requieren de mayor cantidad de datos que los métodos de aprendizaje automático tradicionales y, en el caso del presente estudio, el conjunto de datos iniciales se consideró bastante reducido [83].

Para cumplir tal objetivo, se entrenaron tres algoritmos que actúan correctamente en la predicción de variantes proteicas: Regresión Logística con penalización Lasso (L1), *Support Vector Classifier* (SVC) y *Random Forest* (RF) [58]. De los modelos entrenados, L1 fue el método

con mayor capacidad para distinguir las mutaciones benignas al presentar una especificidad media del 87.17 % (AUC-ROC = 0.82); y RF el modelo con mayor capacidad para distinguir las mutaciones patológicas al presentar una sensibilidad media del 82.53 % (AUC-ROC = 0.80). No obstante, todos los modelos entrenados cuentan con valores de especificidad y sensibilidad próximos al 80 %, mostrando así la calidad de todos ellos. Previamente, Li *et al.* [46] diseñaron una red neuronal para predecir la patogenicidad de variantes en KCNQ1 obteniendo unos valores de AUC-ROC muy próximos a los obtenidos en el presente trabajo, concretamente de 0.85 y 0.87. Además de este estudio, no se encontraron trabajos similares para el resto de miembros de la familia KCNQ.

En primer lugar, los valores de especificidad y sensibilidad obtenidos pudieron deberse a que se observó un conjunto de mutaciones benignas situadas muy próximas a mutaciones patológicas en el espacio tridimensional. Como consecuencia, los modelos entrenados tendrían dificultades a la hora de diferenciar este tipo de mutaciones a partir de los descriptores considerados. Curiosamente, y a pesar de estar empleando descriptores idénticos a otros estudios similares, los valores de precisión para los modelos de KCNQ2 son más bajos [10]. Esta situación podría estar informando de mecanismos moleculares particulares en el caso de las variantes del canal $K_v7.2$ o la necesidad de un mayor tamaño muestral para el entrenamiento de estos modelos. Sin embargo, la información de KCNQ2 en ClinVar y otras bases de datos es muy reducida en comparación con otras enfermedades, lo que dificulta la ampliación del conjunto de datos. Este hecho no es sorprendente desde que algunas enfermedades raras han recibido más atención que otras, a menudo debido a la incidencia de la enfermedad, oportunidades científicas o factores fortuitos [5].

En segundo lugar, mencionar que la estructura secundaria fue considerada por los tres modelos entrenados como un descriptor esencial a la hora de predecir variantes patológicas en KCNQ2, reafirmando una vez más su importancia biológica [16,17]. Otros descriptores esenciales fueron el cambio de polaridad, los dominios topológicos y funcionales de KCNQ2 afectados, así como el valor de conservación evolutiva del residuo mutado.Estos últimos eran esperados debido a su aplicación en múltiples ámbitos de la biología [84].

Por lo tanto, en el presente estudio se han desarrollado los primeros modelos de *machine learning* capaces de predecir la patogenicidad de variantes en KCNQ2 con cierta precisión, lo que facilitaría el diagnóstico clínico y el pronóstico de los pacientes que sufren BFNE/EE [33,73]. Además, este tipo de modelos permitiría reducir tiempo y gastos a la hora de caracterizar la

patogenicidad de las variantes de KCNQ2, dejando clara la importancia de seguir desarrollando este tipo de herramientas. Asimismo, ha permitido reafirmar la importancia de la estructura en la patogenicidad de las variantes proteicas.

Como objetivo final de este estudio, se decidió otorgar una interpretación biológica de las predicciones de los modelos mediante información ya conocida en la bibliografía y los cálculos de estabilidad/inestabilidad de Rosetta. Con este objetivo, se pretendía justificar las predicciones de los modelos así como resaltar su utilidad en el ámbito clínico.

Por un lado, la mutación V182M es predicha por los tres algoritmos entrenados como una mutación benigna. Biológicamente, la predicción de los tres modelos es plausible debido a que se trata de un reemplazo conservativo de valina a metionina donde no existe un cambio de carga ni de polaridad [85]. Además, la variante V182M se localiza en el dominio topológico del segmento S3 formando parte del sensor de voltaje y, según los cálculos de Rosetta, estabilizando su entorno tridimensional. Al estabilizar el sensor de voltaje, un dominio esencial en la apertura y cierre del canal, no derivaría en una clínica patológica [86]. Contrariamente, la variante R201C es predicha por los tres modelos entrenados como una variante patológica de KCNQ2 a pesar de estar formando parte del sensor de voltaje como V182M. Sin embargo, R201C se encuentra en una posición más conservada del segmento S4 y, según los cálculos de Rosetta, estaría desestabilizando su entorno tridimensional. De nuevo, los conocimientos experimentales respaldan la predicción de R201C como mutación patológica en KCNQ2 porque, a pesar de estar formando parte del sensor de voltaje como V182M, se sabe que el segmento S4 es mucho más importante en el sensor de voltaje de los canales KCNQ que los segmentos S1-S3 [87]. Esto se debe a que el segmento S4 es el sensor principal de la despolarización en el canal [88, 89]. Además, la variante R201C experimenta un cambio de carga con la mutación lo que le llevaría a perder una de las cargas positivas esenciales en el dominio del sensor de voltaje [87]. Asimismo, estudios experimentales han demostrado que mutaciones en el dominio S4 provocan grandes cambios en la activación de $K_v7.2$ y se relacionarían con una clínica patológica [38, 90]. Por lo tanto, y a pesar de que ambas mutaciones se ubican en el sensor de voltaje, la predicción de los tres modelos como variante benigna y patológica en el caso de V182M y R201C, respectivamente, se encuentra respaldada tanto por cálculos físicos como por conocimientos experimentales.

Por el otro lado, las variantes G239V y T359A han sido predichas como mutaciones patológicas por los tres algoritmos. En primer lugar, la variante G239V parece tener una interpretación biológica sencilla al encontrarse en el dominio topológico del S5, en una posición altamente con-

servada y, según los cálculos de Rosetta, desestabilizando el poro en gran medida. Este hecho se encuentra respaldado por ensayos experimentales donde se ha demostrado que mutaciones en el dominio topológico del S5 reducen globalmente las amplitudes de las corrientes de $K_v7.2$, al afectar directamente el poro del canal [38, 90]. Por último, la variante T359A tiene un efecto neutral en la estabilidad del sistema, pero, aun así, es predicha como mutación patológica por los tres algoritmos. La decisión de los tres modelos también es biológicamente plausible desde que T359A estaría afectando una posición altamente conservada de la hTW donde tiene lugar la interacción con CaM. Además, la interacción con CaM podría verse afectada en el espacio tridimensional debido a la sustitución de la treonina (T), un aminoácido polar, por la alanina (A), un residuo no-polar. Como consecuencia del cambio de polaridad en esa posición, la proteína tendería a ocultar dicho residuo hidrofóbico evitando su exposición al agua [78, 91]. Estructuralmente, ese cambio de polaridad podría conllevar a un efecto conformacional lo suficientemente relevante como para que la interacción con CaM se viera afectada en T359A y, por ende, derivar en una clínica patológica. Asimismo, estudios experimentales en KCNQ2 han demostrado la importancia de los residuos que se extienden desde T359 hasta Y362 en la unión con CaM y como la reducción de dicha interacción sería uno de los mecanismos patogénicos de $K_v7.2$ [92, 93].

Por lo tanto, las predicciones de los modelos entrenados para las cuatro variantes seleccionadas son biológicamente factibles gracias a la bibliografía consultada y a los cálculos de $\Delta\Delta G$ de Rosetta. En consecuencia, los modelos entrenados toman decisiones que podrían representar un asesoramiento clínicamente relevante sobre variantes de interpretación conflictiva en KCNQ2. No obstante, una validación experimental de la patogenicidad predicha para las cuatro variantes seleccionadas permitiría conocer, en mayor profundidad, la utilidad real de los algoritmos diseñados.

5. Conclusiones

En el presente trabajo:

1) Se ha creado una base de datos de mutaciones *missense* de KCNQ2 combinando información unificada de fuentes bibliográficas y cinco bases de datos clínicas. Además, las mutaciones recopiladas han sido caracterizadas mediante descriptores de secuencia y de

estructura secundaria.

2) Se ha intentado diseñar descriptores estructurales 3D de las variantes anotadas sin éxito debido a la falta de estructura completa de KCNQ2 y a las calidades de los modelos estructurales obtenidos mediante AlphaFold2 y RoseTTAFold.

3) Se ha observado como las tasas de reclasificación de KCNQ2 son las habituales según estudios similares, a pesar de las dificultades diagnósticas de las epilepsias.

4) Se han diseñado, según la bibliografía consultada, los primeros modelos de *machine learning* capaces de predecir la patogenicidad de mutaciones *missense* en el canal $K_v7.2$, alcanzando una especificidad y sensibilidad del 87.17 % y 82.53 %, respectivamente . Además, estos modelos han puesto en evidencia la importancia de la estructura en la predicción de la patogenicidad de sus variantes proteicas.

5) Se han interpretado biológicamente las predicciones realizadas por los modelos entrenados mediante ensayos experimentales previos y cálculos físicos, mostrando así su potencial clínico y la importancia de seguir desarrollando este tipo de herramientas.

6. Limitaciones del estudio

El trabajo presentado cuenta con varias limitaciones. En primer lugar, se decidió trabajar solo con variantes *missense* debido a su representatividad en ClinVar y a su propia naturaleza. Mientras que otras mutaciones no suponen un reto en la valoración clínica (Ej.: *frameshift*, *nonsense*...etc.), el diagnóstico de las variantes *missense* es todo un desafío [94]. Al tomar esta decisión, se limitó la aplicación de los tres modelos a las mutaciones *missense* de KCNQ2. Por lo tanto, la principal limitación de los algoritmos diseñados es su especificidad al canal de potasio dependiente de voltaje $K_v7.2$ y a sus mutaciones *missense*, lo que les impide aplicarse a otros genes o tipos de mutaciones. No obstante, las ideas aportadas en este trabajo servirían de base para el diseño de nuevos algoritmos específicos en otros genes, como podría ser el resto de miembros de la familia KCNQ.

En segundo lugar, otra de las limitaciones del estudio es el tamaño de la base de datos ($N_{total} < 360$). A pesar de los esfuerzos invertidos en su ampliación, el conjunto de datos de trabajo se considera aún pequeño para el problema biológico a resolver. Este hecho probablemente haya afectado directamente a la calidad de los modelos entrenados desde que el éxito de los

algoritmos de *machine learning* se basa en un conjunto de datos de entrenamiento lo suficiente grande como para aprender de ellos y saber generalizar ante nuevos eventos [95]. Además, el presente estudio es incapaz de distinguir entre BFNE y EE desde que decidieron agruparse las variantes de ambas patologías bajo el término de "*pathogenic_variant*" debido a la dificultad clínica de las epilepsias [82].

Por último, otra de las limitaciones del estudio serían las reclasificaciones de variantes que suceden de manera habitual en ClinVar. Si a medida que suceden estas reclasificaciones no se actualizan los datos de entrenamiento, los algoritmos quedarían obsoletos con el paso del tiempo y perderían su relevancia clínica. Por lo tanto, la actualización del conjunto de datos de trabajo a medida que tienen lugar las reclasificaciones o se descubren nuevas variantes es esencial en los algoritmos diseñados.

7. Perspectivas futuras

Como continuación de este trabajo, y debido a los resultados obtenidos, existen diversas cuestiones abiertas en las que es posible seguir indagando.

En primer lugar, se sugiere seguir trabajando en la base de datos de variantes de KCNQ2. Debido a que se encontraron una serie de mutaciones benignas solapando con variantes patológicas, se recomienda, por un lado, seguir ampliando el número de variantes de $K_v7.2$. Una herramienta que podría emplearse es HuVarBase, una base de datos con variantes humanas que no se empleó durante el estudio y permitiría ampliar el conjunto de datos de trabajo. Por el otro lado, y cada cierto tiempo, se sugiere ir actualizando la clínica de las variantes recopiladas a medida que van sucediendo las reclasificaciones. Recientemente, Larrea-Sebal *et al.* [10] consiguieron desarrollar un modelo de *machine learning* que actualizaba de manera periódica su conjunto de datos de entrenamiento. Aplicando su idea, se conseguiría solucionar ambos contratiempos de manera automática.

En segundo lugar, se sugiere seguir insistiendo en la obtención de un modelo estructural de KCNQ2. A pesar de las limitaciones observadas en el estudio, los descriptores estructurales 3D de las variantes proteicas podrían tener un papel fundamental en su patogenicidad, de igual manera que el descriptor de estructura secundaria diseñado ha sido relevante en su clasificación. Además, hace escasos meses DeepMind puso a disposición de la comunidad científica

a AlphaFold-Multimer, una extensión de AlphaFold2 que permite predecir complejos proteicos [96]. De esta manera, en vez de modelar los monómeros de $K_v7.2$ se podría modelar el canal tetramérico por completo.

Por último, se sugiere entrenar modelos más complejos de *machine learning* pudiéndose obtener así mejores resultados. Asimismo, se podrían aplicar conceptos más novedosos como el de aprendizaje de transferencia, que ofrece una oportunidad para aprovechar la potencia del *deep learning* en situaciones en las que los datos son limitados [5]. Este tipo de aprendizaje consistiría en entrenar un modelo utilizando datos de un gen bien estudiado "X" y luego refinarlo con datos de un gen menos estudiado "Y" (Ej.: KCNQ1 y KCNQ2). El modelo resultante rendiría muy bien con "Y" desde que las "lecciones" aprendidas al modelar "X" se han transferido correctamente a "Y" [97].

Por lo tanto, el presente trabajo ha permitido diseñar tres algoritmos que facilitan el diagnóstico clínico de variantes *missense*en $K_v7.2$ al realizar predicciones precisas de manera rápida y casi sin costes. Asimismo, ha servido para identificar los obstáculos que presenta la predicción de variantes en KCNQ2, pudiéndola mejorar en un futuro según se vayan descubriendo y reclasificando las diferentes variantes *missense*. A pesar de las limitaciones encontradas, los modelos entrenados suponen un gran avance en el diagnóstico clínico de las epilepsias debido a sus dificultades actuales. Finalmente, el presente trabajo ha abierto nuevas oportunidades en el estudio del canal $K_v7.2$ que se espera que se sigan investigando, así como una aplicación clínica que se espera que se siga explotando y desarrollando en un futuro.

8. Material suplementario

El material suplementario del presente estudio, así como el conjunto de datos de trabajo y su código asociado se encuentran disponibles en: https://github.com/albasaezmat/TFM_UAM.git

Si alguna persona estuviera interesada en emplear los algoritmos entrenados para la predicción de la patogenicidad de variantes *missense* en KCNQ2 puede consultar el procedimiento especificado en el Apartado 2 de *ML_models.ipynb*, también depositado en el anterior GitHub.

9. Bibliografía

[1] Cooper GM, Shendure J. Needles in stacks of needles: Finding disease-causal variants in a wealth of genomic data; 2011.

[2] Mayer AN, Dimmock DP, Arca MJ, Bick DP, Verbsky JW, Worthey EA, et al.. A timely arrival for genomic medicine; 2011.

[3] MacArthur DG, Tyler-Smith C. Loss-of-function variants in the genomes of healthy humans. Human Molecular Genetics. 2010;19.

[4] Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. 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. Genetics in Medicine. 2015;17.

[5] McInnes G, Sharo AG, Koleske ML, Brown JEH, Norstad M, Adhikari AN, et al.. Opportunities and challenges for the computational interpretation of rare variation in clinically important genes; 2021.

[6] Roberts NJ, Vogelstein JT, Parmigiani G, Kinzler KW, Vogelstein B, Velculescu VE. The Predictive Capacity of Personal Genome Sequencing. Science Translational Medicine. 2012;4(133).

[7] Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, Chen R, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell. 2012;148.

[8] Katsanis SH, Katsanis N. Molecular genetic testing and the future of clinical genomics; 2013.

[9] Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Research. 2012;40.

[10] Larrea-Sebal A, Benito-Vicente A, Fernandez-Higuero JA, Jebari-Benslaiman S, Galicia-Garcia U, Uribe KB, et al. MLb-LDLr: A Machine Learning Model for Predicting the Pathogenicity of LDLr Missense Variants. JACC: Basic to Translational Science. 2021;6.

[11] Wainberg M, Merico D, Delong A, Frey BJ. Deep learning in biomedicine. Nature Biotechnology. 2018;36.

[12] Ferrer-Costa C, Orozco M, Cruz XDL. Sequence-based prediction of pathological mutations. Proteins: Structure, Function and Genetics. 2004 12;57:811-9.

[13] McCoy AJ, Sammito MD, Read RJ. Possible Implications of AlphaFold2 for Crystallographic Phasing by Molecular Replacement. bioRxiv. 2021.

[14] Izarzugaza JMG, Krallinger M, Valencia A. Interpretation of the consequences of mutations in protein kinases: Combined use of bioinformatics and text mining; 2012.

[15] Illergård K, Ardell DH, Elofsson A. Structure is three to ten times more conserved than sequence-A study of structural response in protein cores. Proteins: Structure, Function, and Bioinformatics. 2009 11;77.

[16] Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Costanzo LD, et al. Protein Data Bank: The single global archive for 3D macromolecular structure data. Nucleic Acids Research. 2019;47.

[17] Ginalski K. Comparative modeling for protein structure prediction; 2006.

[18] Bateman A, Martin MJ, Orchard S, Magrane M, Agivetova R, Ahmad S, et al. UniProt: The universal protein knowledgebase in 2021. Nucleic Acids Research. 2021;49.

[19] Schwede T. Protein modeling: What happened to the "protein structure gap¿; 2013.

[20] David A, Islam S, Tankhilevich E, Sternberg MJE. The AlphaFold Database of Protein Structures: A Biologist's Guide; 2022.

[21] Alquraishi M. AlphaFold at CASP13. Bioinformatics. 2019;35.

[22] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596.

[23] Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021;373.

[24] Golinelli-Pimpaneau B. Prediction of the Iron–Sulfur Binding Sites in Proteins Using the Highly Accurate Three-Dimensional Models Calculated by AlphaFold and RoseTTAFold. Inorganics. 2021;10(1):2.

[25] Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: Public archive of interpretations of clinically relevant variants. Nucleic Acids Research. 2016;44.

[26] Macklin S, Durand N, Atwal P, Hines S. Observed frequency and challenges of variant reclassification in a hereditary cancer clinic. Genetics in Medicine. 2018;20.

[27] Landrum MJ, Kattman BL. ClinVar at five years: Delivering on the promise. Human Mutation. 2018;39.

[28] Slavin TP, Manjarrez S, Pritchard CC, Gray S, Weitzel JN. The effects of genomic germline variant reclassification on clinical cancer care. Oncotarget. 2019;10.

[29] VanDyke RE, Hashimoto S, Morales A, Pyatt RE, Sturm AC. Impact of variant reclassification in the clinical setting of cardiovascular genetics. Journal of Genetic Counseling. 2021;30.

[30] Fiest KM, Sauro KM, Wiebe S, Patten SB, Kwon CS, Dykeman J, et al. Prevalence and incidence of epilepsy. Neurology. 2017 1;88.

[31] Epilepsia; 2019. Available from: https://www.who.int/es/news-room/fact-sheets/detail/epilepsy.

[32] Benign familial neonatal seizures; 2018. Available from: https://globalgenes.org/disease/benign-familial-neonatal-seizures/.

[33] Shorvon SD, Andermann F, Guerrini R. The causes of epilepsy: common and uncommon causes in adults and children. Cambridge University Press; 2011.

[34] Maljevic S, Wuttke TV, Lerche H. Nervous system KV7 disorders: Breakdown of a subthreshold brake. vol. 586; 2008. .

[35] Soldovieri MV, Boutry-Kryza N, Milh M, Doummar D, Heron B, Bourel E, et al. Novel KCNQ2 and KCNQ3 mutations in a large cohort of families with benign neonatal epilepsy: First evidence for an altered channel regulation by syntaxin-1A. Human Mutation. 2014;35.

[36] Miceli F, Soldovieri MV, Joshi N, Weckhuysen S, Cooper E, Taglialatela M. KCNQ2-Related Disorders; 1993.

[37] Ritter DM, Horn PS, Holland KD. In Silico Predictions of KCNQ Variant Pathogenicity in Epilepsy. Pediatric Neurology. 2021 5;118.

[38] Zhang J, Kim EC, Chen C, Procko E, Pant S, Lam K, et al. Identifying mutation hotspots reveals pathogenetic mechanisms of KCNQ2 epileptic encephalopathy. Scientific Reports. 2020 12;10.

[39] Ambrosino P, Alaimo A, Bartollino S, Manocchio L, Maria MD, Mosca I, et al. Epilepsy-causing mutations in Kv7.2 C-terminus affect binding and functional modulation by calmodulin. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease. 2015 9;1852.

[40] Miceli F, Soldovieri MV, Ambrosino P, Barrese V, Migliore M, Cilio MR, et al. Genotype-phenotype correlations in neonatal epilepsies caused by mutations in the voltage sensor of Kv7.2 potassium channel subunits. Proceedings of the National Academy of Sciences of the United States of America. 2013;110.

[41] Schwake M, Athanasiadu D, Beimgraben C, Blanz J, Beck C, Jentsch TJ, et al. Structural determinants of M-type KCNQ (Kv7) K+ channel assembly. The Journal of neuroscience : the official journal of the Society for Neuroscience. 2006;26.

[42] Etxeberria A, Aivar P, Rodriguez-Alfaro JA, Alaimo A, Villace P, Gomez-Posada JC, et al. Calmodulin regulates the trafficking of KCNQ2 potassium channels. The FASEB Journal. 2008;22.

[43] Brown DA. M currents. Ion channels. 1988:55-94.

[44] Marrion NV. Control of M-current; 1997.

[45] DeLano WL. Pymol: An open-source molecular graphics tool. CCP4 Newsletter On Protein Crystallography. 2002.

[46] Li B, Mendenhall JL, Kroncke BM, Taylor KC, Huang H, Smith DK, et al. Predicting the Functional Impact of KCNQ1 Variants of Unknown Significance. Circulation: Cardiovascular Genetics. 2017;10.

[47] Xenakis MN, Kapetis D, Yang Y, Gerrits MM, Heijman J, Waxman SG, et al. Hydropathicity-based prediction of pain-causing NaV1.7 variants. BMC Bioinformatics. 2021;22.

[48] Leman JK, Weitzner BD, Lewis SM, Adolf-Bryfogle J, Alam N, Alford RF, et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks. Nature Methods. 2020 7;17.

[49] Chen K, Chen H, Zhou C, Huang Y, Qi X, Shen R, et al. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. Water Research. 2020;171:115454.

[50] Lupton RC, Allwood JM. Hybrid Sankey diagrams: Visual analysis of multidimensional data for understanding resource use. Resources, Conservation and Recycling. 2017;124.

[51] Montgomerie S, Cruz JA, Shrivastava S, Arndt D, Berjanskii M, Wishart DS. PROTEUS2: a web server for comprehensive protein structure prediction and structure-based annotation. Nucleic acids research. 2008;36.

[52] Van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009.

[53] Wes McKinney. Data Structures for Statistical Computing in Python. In: Stéfan van der Walt, Jarrod Millman, editors. Proceedings of the 9th Python in Science Conference; 2010. p. 56 61.

[54] Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020 Sep;585(7825):357-62.

[55] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825-30.

[56] Inc PT. Collaborative data science. Montreal, QC: Plotly Technologies Inc.; 2015.

[57] Mohammed R, Rawashdeh J, Abdullah M. Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results; 2020. .

[58] Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nature Reviews Molecular Cell Biology. 2021;23(1):40-55.

[59] Muschelli J. ROC and AUC with a Binary Predictor: a Potentially Misleading Metric. Journal of Classification. 2020;37.

[60] Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. Journal of Thoracic Oncology. 2010;5.

[61] Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression.;.

[62] Razquin-Lizarraga A. Computational analysis of 12 mutations in the potassium channel KCNQ2; 2021. TFG.

[63] Fleishman SJ, Leaver-Fay A, Corn JE, Strauch EM, Khare SD, Koga N, et al. Rosettascripts: A scripting language interface to the Rosetta Macromolecular modeling suite. PLoS ONE. 2011;6.

[64] Chaudhury S, Lyskov S, Gray JJ. PyRosetta: A script-based interface for implementing molecular modeling algorithms using Rosetta; 2010.

[65] Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature 2021 596:7873. 2021 7;596:590-6. Available from: https://www.nature.com/articles/s41586-021-03828-1.

[66] Rehmat N, Farooq H, Kumar S, Hussain SU, Naveed H. Predicting the pathogenicity of protein coding mutations using Natural Language Processing. vol. 2020-July; 2020. .

[67] Corcoran RB, Atreya CE, Falchook GS, Kwak EL, Ryan DP, Bendell JC, et al. Combined BRAF and MEK inhibition with dabrafenib and trametinib in BRAF V600-Mutant colorectal cancer. vol. 33; 2015. .

[68] Glusman G. Clinical applications of sequencing take center stage. Genome Biology. 2013;14.

[69] Qi H, Zhang H, Zhao Y, Chen C, Long JJ, Chung WK, et al. MVP predicts the pathogenicity of missense variants by deep learning. Nature Communications. 2021;12.

[70] Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014;515.

[71] Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam HJ, et al. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nature Communications. 2020;11.

[72] Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, et al. The Pfam protein families database. Nucleic Acids Research. 2010 1;38:D211-22.

[73] Rowlands CF, Baralle D, Ellingford JM. Machine Learning Approaches for the Prioritization of Genomic Variants Impacting Pre-mRNA Splicing; 2019.

[74] Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, et al. The Protein Data Bank. urn:issn:0907-4449. 2002 5;58:899-907.

[75] Liu Z, Zheng R, Grushko MJ, Uversky VN, McDonald TV. Functionally Aberrant Mutant KCNQ1 With Intermediate Heterozygous and Homozygous Phenotypes. Canadian Journal of Cardiology. 2018;34.

[76] Sachyani D, Dvir M, Strulovich R, Tria G, Tobelaim W, Peretz A, et al. Structural Basis of a Kv7.1 Potassium Channel Gating Module: Studies of the Intracellular C-Terminal Domain in Complex with Calmodulin. Structure. 2014;22.

[77] Howard RJ, Clark KA, Holton JM, Minor DL. Structural Insight into KCNQ (Kv7) Channel Assembly and Channelopathy. Neuron. 2007;53.

[78] Pinheiro F, Santos J, Ventura S. AlphaFold and the amyloid landscape; 2021.

[79] Perrakis A, Sixma TK. AI revolutions in biology. EMBO reports. 2021;22.

[80] Harrison SM, Rehm HL. Is 'likely pathogenic' really 90Reclassification data in ClinVar; 2019.

[81] Sorelle JA, Thodeson DM, Arnold S, Gotway G, Park JY. Clinical Utility of Reinterpreting Previously Reported Genomic Epilepsy Test Results for Pediatric Patients. JAMA Pediatrics. 2019;173.

[82] Hampel KG, Sánchez MG, Ibáñez AG, Cámara MP, Villanueva V. Desafíos diagnósticos en epilepsia. Revista de Neurología. 2019;68.

[83] Erickson BJ, Korfiatis P, Kline TL, Akkus Z, Philbrick K, Weston AD. Deep Learning in Radiology: Does One Size Fit All? Journal of the American College of Radiology. 2018 3;15:521-6.

[84] Capra JA, Singh M. Predicting functionally important residues from sequence conservation. Bioinformatics. 2007;23.

[85] Smith EL, DeLange RJ, Evans WH, Landon M, Markland FS. Subtilisin Carlsberg: V. THE COMPLETE SEQUENCE; COMPARISON WITH SUBTILISIN BPN; EVOLUTIONARY RELATIONSHIPS. Journal of Biological Chemistry. 1968 5;243:2184-91.

[86] Jepps TA, Barrese V, Miceli F. Editorial: Kv7 Channels: Structure, Physiology, and Pharmacology. Frontiers in Physiology. 2021 4;12:516.

[87] Nakajo K, Kubo Y. KCNQ1 channel modulation by KCNE proteins via the voltage-sensing domain. Journal of Physiology. 2015 6;593:2617-25.

[88] Sun J, MacKinnon R. Cryo-EM Structure of a KCNQ1/CaM Complex Reveals Insights into Congenital Long QT Syndrome. Cell. 2017;169.

[89] Cui J. Voltage-Dependent Gating: Novel Insights from KCNQ1 Channels; 2016.

[90] Orhan G, Bock M, Schepers D, Ilina EI, Reichel SN, Löffler H, et al. Dominant-negative effects of KCNQ2 mutations are associated with epileptic encephalopathy. Annals of Neurology. 2014;75.

[91] Yadav NS, Choudhury D. Conformational perturbation of peptides in presence of polar organic solvents. Journal of Molecular Graphics and Modelling. 2019 6;89:1-12.

[92] Gomis-Perez C, Alaimo A, Fernandez-Orth J, Alberdi A, Aivar-Mateo P, Bernardo-Seisdedos G, et al. An unconventional calmodulin-anchoring site within the AB module of Kv7.2 channels. Journal of Cell Science. 2015;128.

[93] Richards MC, Heron SE, Spendlove HE, Scheffer IE, Grinton B, Berkovic SF, et al. Novel mutations in the KCNQ2 gene link epilepsy to a dysfunction of the KCNQ2-calmodulin interaction. Journal of medical genetics. 2004;41.

[94] Majithia AR, Tsuda B, Agostini M, Gnanapradeepan K, Rice R, Peloso G, et al. Prospective functional classification of all possible missense variants in PPARG. Nature Genetics. 2016;48.

[95] Zhang Y, Ling C. A strategy to apply machine learning to small datasets in materials science. npj Computational Materials. 2018;4.

[96] Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-Multimer. bioRxiv. 2021.

[97] Taroni JN, Grayson PC, Hu Q, Eddy S, Kretzler M, Merkel PA, et al. MultiPLIER: A Transfer Learning Framework for Transcriptomics Reveals Systemic Features of Rare Disease. Cell Systems. 2019;8.