



Predictive Modeling and Structure Analysis of Genetic Variants in Familial Hypercholesterolemia: Implications for Diagnosis and Protein Interaction Studies

Asier Larrea-Sebal^{1,2,3} · Shifa Jebari-Benslaiman^{1,2} · Unai Galicia-Garcia^{1,2} · Ane San Jose-Urteaga¹ · Kepa B. Uribe¹ · Asier Benito-Vicente^{1,2} · César Martín^{1,2}

Accepted: 15 September 2023 / Published online: 17 October 2023
© The Author(s) 2023

Abstract

Purpose of Review Familial hypercholesterolemia (FH) is a hereditary condition characterized by elevated levels of low-density lipoprotein cholesterol (LDL-C), which increases the risk of cardiovascular disease if left untreated. This review aims to discuss the role of bioinformatics tools in evaluating the pathogenicity of missense variants associated with FH. Specifically, it highlights the use of predictive models based on protein sequence, structure, evolutionary conservation, and other relevant features in identifying genetic variants within *LDLR*, *APOB*, and *PCSK9* genes that contribute to FH.

Recent Findings In recent years, various bioinformatics tools have emerged as valuable resources for analyzing missense variants in FH-related genes. Tools such as REVEL, Varsity, and CADD use diverse computational approaches to predict the impact of genetic variants on protein function. These tools consider factors such as sequence conservation, structural alterations, and receptor binding to aid in interpreting the pathogenicity of identified missense variants. While these predictive models offer valuable insights, the accuracy of predictions can vary, especially for proteins with unique characteristics that might not be well represented in the databases used for training.

Summary This review emphasizes the significance of utilizing bioinformatics tools for assessing the pathogenicity of FH-associated missense variants. Despite their contributions, a definitive diagnosis of a genetic variant necessitates functional validation through in vitro characterization or cascade screening. This step ensures the precise identification of FH-related variants, leading to more accurate diagnoses. Integrating genetic data with reliable bioinformatics predictions and functional validation can enhance our understanding of the genetic basis of FH, enabling improved diagnosis, risk stratification, and personalized treatment for affected individuals. The comprehensive approach outlined in this review promises to advance the management of this inherited disorder, potentially leading to better health outcomes for those affected by FH.

Keywords Familial hypercholesterolemia · LDLR · APOB · PCSK9 · Bioinformatics tools · Functional validation

Introduction: Familial Hypercholesterolemia

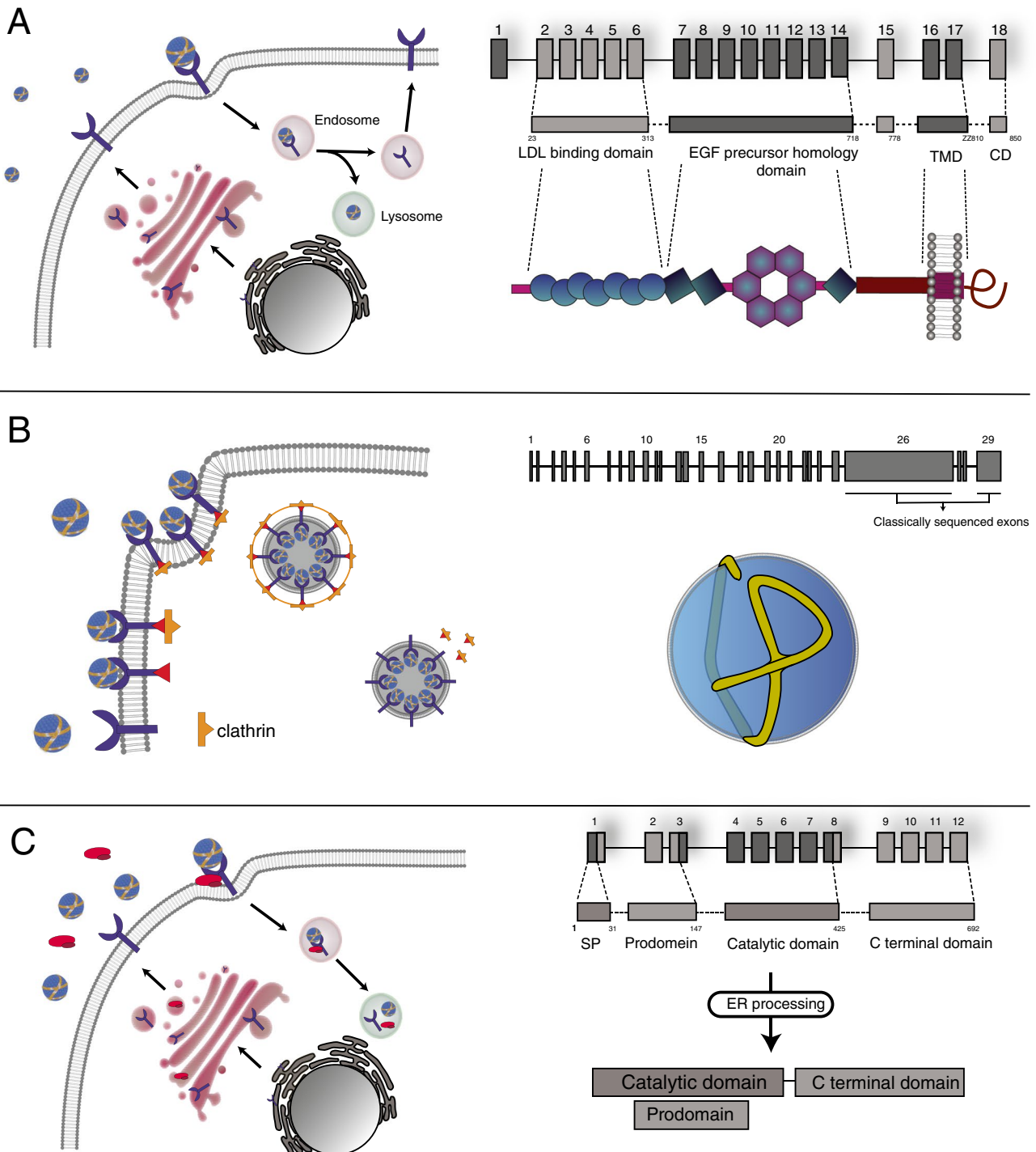
Familial hypercholesterolemia (FH) is a common inherited autosomal semi-dominant disorder primarily characterized by high plasma levels of low-density lipoprotein cholesterol (LDL-C) due to impaired metabolism [1]. If left untreated, persistent elevation of LDL-C throughout a person's lifetime can lead to the development of atherosclerotic plaques and an increased risk of premature cardiovascular disease [2]. The major genetic determinants of FH correspond to pathogenic variants in 3 genes that cover the 99% of FH cases: *LDLR*, *APOB* (apolipoprotein B), and *PCSK9* (Pro-protein Convertase Subtilisin/Kexin Type 9) [3]. The prevalence of FH in its heterozygous

✉ César Martín
cesar.martin@ehu.eus

¹ Department of Biochemistry and Molecular Biology, Universidad del País Vasco UPV/EHU, 48080 Bilbao, Spain

² Department of Molecular Biophysics, Biofisika Institute, University of Basque Country and Consejo Superior de Investigaciones Científicas (UPV/EHU, CSIC), 48940 Leioa, Spain

³ Fundación Biofisika Bizkaia, 48940 Leioa, Spain



form (HeFH) has traditionally been estimated to be around 1 in 500 individuals. However, the frequency can vary between 1 in 200 and 1 in 300 depending on the specific criteria used to define FH (such as genetic variants, LDL-C threshold, clinical score, or a combination of factors) and the populations under study [4]. In the

case of the homozygous form of the disease (HoFH), the prevalence has traditionally been estimated at 1 in 1 million individuals. However, recent studies have revealed a higher prevalence, with estimates reaching as high as 1 in 300,000 individuals [4]. Despite its high prevalence, FH is still underdiagnosed, with less than 1% of the patients

Fig. 1 Cholesterol Homeostasis and Genetic Variants. **(A)** The *LDLR* gene encodes a transmembrane protein that regulates cholesterol homeostasis. LDLR has functional subdomains within its ectodomain and intracellular domain. It interacts with lipoproteins and PCSK9, leading to internalization and degradation. Pathogenic LDLR variants are classified into six subclasses based on their effects. **(B)** The *APOB* gene encodes apoB-100, a key component of lipoproteins and the primary ligand for LDLR. The conformation and binding regions of apoB-100 influence its affinity for LDLR. *APOB* pathogenic variants can occur outside these binding regions, making the identification of pathogenic variants more challenging. This suggests that conformational modifications of apoB-100 may play a significant role in its interaction with LDLR. **(C)** PCSK9 is synthesized as a zymogen and undergoes autocatalysis to release a peptide that inactivates its catalytic activity. Binding of PCSK9 to the LDLR-EGF domain prevents the conformational change required for receptor recycling. Upon endosome acidification, PCSK9's affinity for LDLR increases, leading to the degradation of the LDLR-PCSK9 complex in the lysosome. GOF variants increase LDLR degradation, resulting in elevated LDL-C levels and increased risk of CVD

diagnosed in most countries [5]. Although there is a consensus on the criteria required to diagnose FH there are several clinical scoring systems that evaluate differently the consensus parameters [6, 7]. Among them, Dutch Lipid Clinic Network (DLCN)5 and Simon Broome Register (SBR) [8] are the most used ones. Most FH clinical algorithms consider lipid values (total cholesterol and LDL-C levels), the presence of physical stigmata (tendon xanthoma or corneal arcus), cascade screening and pathogenic DNA variants. Functional validation plays a crucial role in achieving a correct and early diagnosis of FH through genetic testing, which is considered the preferred method for FH diagnosis. However, the majority of FH variants lack functional characterization, requiring additional measures to complement genetic testing for an accurate and definitive diagnosis [9].

LDLR

The *LDLR* gene located on 19p13.2 chromosome encodes a type I transmembrane protein of 839 amino acids, the LDLR, which regulates cholesterol homeostasis in mammalian cells [1] and constitutes the main gene associated with FH. Genetic variants in *LDLR* represent more than 90% of the FH causing variants, with more than 3000 variants annotated in ClinVar database [10]. LDLR is structured into functional subdomains organized within an ectodomain and intracellular domain (Fig. 1). The ectodomain contains the ligand binding domain (LBD), the epidermal growth factor precursor homology domain (EGF) and the O-linked domain. On the other hand, the intracellular domain harbors the transmembrane domain and the cytoplasmic domain, that targets the LDLR to clathrin-coated pits for the internalization of the LDLR-LDL complex [11, 12]. Binding of lipoproteins to the LDLR is mediated

through the interaction of acidic residues in the LBD with basic residues of apoB-100 or ApoE [13]. Additionally, LDLR also interacts with PCSK9, a secreted protein that regulates membrane levels of LDLR through binding to EGF-LDLR domain, leading to degradation of LDLR in the endosome [14].

According to the region of the LDLR affected, the LDLR pathogenic variants can be classified into six sub-classes: class 1: LDLR is not synthesized, known as “null allele”; Class 2: LDLR is retained in the endoplasmic reticulum, completely or partially (2A and 2B, respectively); Class 3: Deficient binding to apoB-100; Class 4: Impaired endocytosis due to a deficient recruitment of the LDLR into clathrin-coated pits; Class 5: Impaired recycling; Class 6: impaired insertion in the membrane [3, 15]. LDLR pathogenic variants have been described along all domains and, depending on their location, they can affect the receptor function differently.

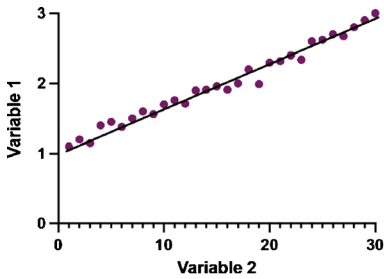
APOB

The *APOB* gene located on the 2p24.1 chromosome is a large and polymorphic gene spanning 43 kb in length, which constitutes the second most common cause of FH. The *APOB* gene comprises 29 exons and 28 introns, and encodes two forms of apolipoprotein B (apoB) in circulating lipoproteins, apoB-48 and apoB-100. ApoB-48 is produced by the small intestine, whereas full-length apoB-100 is produced in the liver. The mature form of apoB-100 is a protein of 4536 amino acids [16], which constitutes both the integral component of several lipoproteins (very low-density lipoprotein (VLDL) and LDL [17]) and the ligand for LDLR [18]. ApoB-100 interacts with lipids in a close manner, and the conformation that adopts the apolipoprotein within the lipid moiety confers the structure and physical properties to the lipoprotein [19] (Fig. 1). The apoB-100 domain that interacts with LDLR was first localized between residues 3386 and 3396 [20]. Although not proven experimentally, an eight-domain model for apoB-100 binding to LDLR was later proposed, in which the LDLR binding regions in apoB-100 expand between residues 2820–3202 and 3243–3498 [21]. The most frequent *APOB* pathogenic variant to date is p.R3527Q [22]. It has been shown that alterations in residue 3527 destabilize the protein, affecting the structure and, thus, the affinity for LDLR [23]. In addition to pathogenic variants within the putative binding domain, a growing number of pathogenic *APOB* variants are being described outside the putative binding regions, which complicates the identification of pathogenic variants [24]. In addition, this heterogeneity of pathogenic variants in *APOB* suggests that conformational modifications of the apolipoprotein could be a key player in the affinity of apoB-100 to LDLR [25].

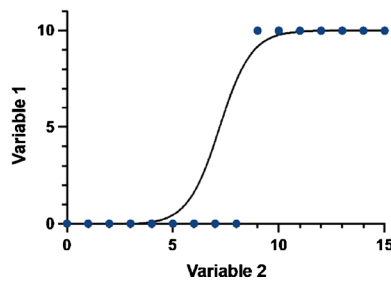
Statistical methodology

Classical methods

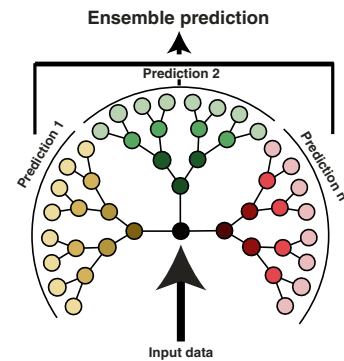
A Linear regression



B Logistic regression

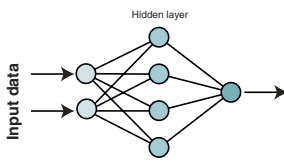


C Random Forest

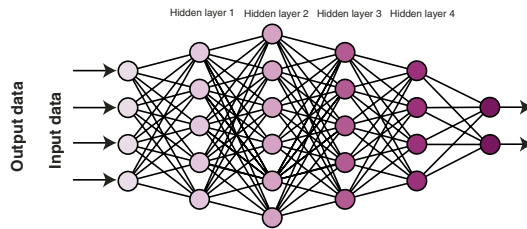


Neural Networks

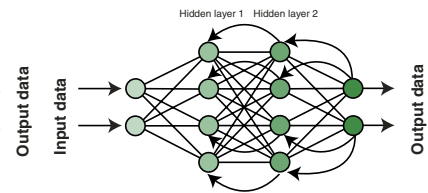
D Single neural network



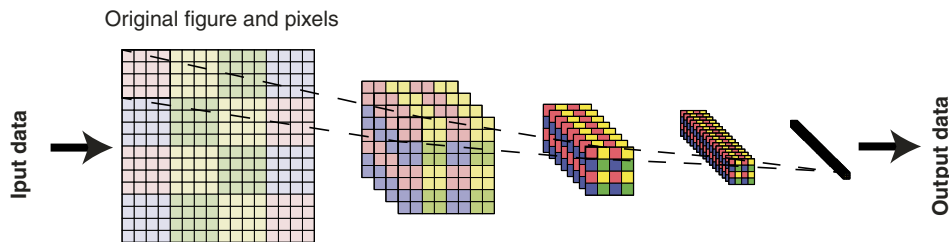
E Deep neural network



F Neural network (feedback)



G Convolutional neural network



Bayesian Regression

- Models relationships between variables with uncertain quantification.
- Enables predictions using probability distributions.
- Integrates prior beliefs and observed data.
- Applies to various techniques (regressions, neural network) as probability based frameworks.
- Enhances understanding and informed decision-making with explicit uncertainty estimates.
- Assumes variable independence for simplicity.
- Requires less data for parameter estimation.
- Computationally faster due to independence.
- Effective with high-dimensional data

Bayes



Naive Bayes



Fig. 2 AI statistical models. Classical methods: **(A)** Linear regression describes the relationship between features with a straight line. **(B)** Logistic regression divides features using sigmoidal curves, a more complex approach. **(C)** In random forest, features are consequently divided to improve the classification accuracy. Each decision “tree” starts from a first split in the database called the “root,” and the division continues creating “branches.” The classification output of each “tree” is combined with other “trees,” creating a “forest,” and the most voted option is the output of the model. **Neural Networks:** Neural networks draw inspiration from biological neural networks where nodes (neurons) communicate via connections (axons and dendrites) with weighted synapses, optimizing information flow. In contrast to classical techniques, neural networks can construct planes or hyperplanes (in multiple dimensions) to effectively separate features. **(D)** A basic neural network includes an input layer, up to three hidden layers, and an output layer. **(E)** Deep neural networks can comprise hundreds of layers, enhancing overall performance but requiring complex development. **(F)** Feed-back neural networks allow signals to traverse layers bidirectionally. **(G)** Convolutional networks are primarily employed for image recognition. **Bayes:** This approach models relationships while quantifying uncertainty. The Bayes approach can be applied to classical models and neural networks, enabling predictions with probability distributions, and incorporating prior information of the problem into the model. Bayesian models enable a deeper understanding of the underlying data and facilitate informed decision-making. In contrast to standard Bayesian models, Naïve Bayes assumes independence between features, simplifying models

PCSK9

PCSK9 belongs to a family of 9 subtilisin-like serine proteases and plays a key function in plasma cholesterol metabolism by regulating LDLR levels through the promotion of LDLR degradation [26]. The *PCSK9* gene is located on the short arm of chromosome 1p32.3. PCSK9 is synthesized as a 72 kDa soluble zymogen (proPCSK9), which undergoes an autocatalytic process at the N-terminal domain. Upon autocatalysis, a 14 kDa peptide is released, which remains attached to the mature protein and inactivates the catalytic activity [27]. Upon binding to the LDLR-EGF domain, PCSK9 prevents the LDLR conformational change required to be recycled. This effect occurs after endosome acidification, which increases the affinity of PCSK9 for LDLR and leads the LDLR-PCSK9 complex to degradation in the lysosome [28] (Fig. 1). Large cohort studies have shown the existence of two major *PCSK9* variants, gain-of-function (GOF) and loss-of-function (LOF) [29, 30]. *PCSK9* pathogenic variants leading to GOF activity have been identified as the third genetic cause of FH [29, 31]. PCSK9 GOF variants increase LDLR degradation, resulting in higher circulating LDL-C levels, which directly increases the risk of developing CVD. On the other hand, LOF variants have a diminished effect on LDLR degradation, thus leading to lower LDL-C levels and decreased CVD risk. Both types of variants are broadly distributed along the three domains of the protein: the prodomain, the catalytic domain, and the C-terminal domain. The mechanisms by which the GOF/LOF PCSK9 variants affect LDLR degradation-rate are

diverse, so predicting the effect of PCSK9 variants is complex [32].

Bioinformatics as a Clinical Tool

Functional characterization is a direct method to assess the activity of a variant by analyzing its effect on the biological processes in which the molecule is involved [15]. However, this is an arduous, time-consuming, and costly process that involves obtaining samples from the patient for subsequent purification or cloning the variant to further characterize their functionality in vitro. Additionally, during the past few years, high-throughput next-generation sequencing (NGS)-based methods have drastically increased the number of FH-related gene variants, opening the floodgates for the development of prediction models [33, 34]. On the other hand, the huge number of genetic variants being described through high-throughput NGS has increased to such an extent that it is almost impossible to functionally characterize all of them. Although functional characterization is essential for the proper analysis of a variant, in silico prediction tools offer a quick, cheap, and increasingly precise alternative.

Bioinformatic tools use the current knowledge about a protein or protein family to create in silico predictive models. These tools mostly rely on artificial intelligence (AI) to develop mathematical models capable of solving complex biological problems by analyzing vast datasets and intricate molecular interactions, ultimately aiding in the prediction of protein structure, function, and interactions [35]. AI involves the development of intelligent systems that act rationally in response to the given inputs. Machine learning (ML), one of the most well-known AI disciplines, applies statistical models and algorithms to analyze data. In contrast to classical programming, where known features (inputs) are used to create the algorithm, ML may use novel or different combinations of inputs and weights [36].

The most important parameters when developing a ML model are the dataset used in the training of the model and the approach used to optimize the results. Depending on the datasets used, an ML model can be supervised or unsupervised [37]. Supervised models learn from labeled training data (pathogenic/benign) and try to fit the algorithm to give accurate predictions [38]. This type of training dataset is applied for regression and classification problems, so it is the most common in the field of pathogenicity prediction. Unsupervised models, on the other hand, use unlabeled data, and are mainly used for clustering and anomaly detection [39].

In terms of the statistical methodology, the primary techniques encompass classical methods, neural networks, and Bayesian regression [35] (Fig. 2). The most basic classical technique is linear regression, where the relationship between one or more numerical features is described using a straight line (Fig. 2A). A more complex classical technique

is logistic regression since the relationship is estimated by a sigmoidal curve (Fig. 2B). Decision trees and random forest are also classical techniques. They are trained by supervised datasets and are mostly used for classification and regression. Each “tree” starts from a “root,” the first split in the dataset that best divides the data into their respective classes. After the split, the process can continue creating “branches.” To create a “forest,” the dataset can be divided into subsamples that are used to create multiple “trees,” and the majority vote among them is used as the final model [36] (Fig. 2C). Neural networks are inspired by biological neural networks, where each node (neuron) communicates with others via connections (axons and dendrites), and these connections are weighted to provide an optimized output. In contrast with classical techniques, neural networks can find planes or hyperplanes (more than three dimensions) to separate the features. The most basic neural network consists of an input layer, up to three hidden layers, and an output layer (Fig. 2D). However, deep neural networks can contain hundreds of layers. This structure gives deep neural network a better overall performance, but they are harder to develop (Fig. 2E). In addition to the number of layers, there are multiple types of neural networks, such as feedforward (standard, a layer communicates with the next one), feed-back (signals can go back in layers), or convolutional (mostly used for image recognition) [40] (Fig. 2F and G). Finally, the Bayesian approach is a statistical method that models relationships between variables while quantifying uncertainty, enabling predictions with probability distributions, informed by prior beliefs, and observed data [41]. The Bayesian approach can be applied to other techniques such as classical regressions or neural networks, treating coefficients as probability distributions, and capturing uncertainty in predictions. Bayesian methods give the possibility to incorporate prior information into the model, a probability distribution that represents biological knowledge, or assumptions about the possible values of a parameter before observing any data. Bayesian models enable a deeper understanding of the underlying data and facilitate informed decision-making, in contrast to black-box models that offer predictions without explicit uncertainty estimates. Naïve Bayes is a classification technique that assumes conditional independence between features, making it computationally efficient but potentially oversimplifying real-world relationships (Fig. 2 lower panel).

Pathogenicity Prediction Software for FH: Analysis of Human Genetic Variants

Single nucleotide variants (SNVs) represent most of the human genetic variants and constitute a major class of genetic risk across common and rare diseases [42]. The

SNVs of special interest are those in which an amino acid is substituted, known as non-synonymous SNVs (nsSNV) since they can affect the biological function of a gene product in several ways. In fact, most of the variants that cause FH described so far are missense, ranging from 46%, 52%, and 83% in *LDLR*, *PCSK9*, and *APOB*, respectively [43]. The effect introduced by a missense variant is difficult to predict; in fact, many of the *LDLR* variants classified as pathogenic by *in silico* predictions were later reclassified after cascade screening and co-segregation studies [44].

The American College of Medical Genetics and Genomics (ACMG) proposed a specific nomenclature and criteria for the classification of pathogenic variants [45]. This classification recommends the use of five distinct subclasses: pathogenic, likely pathogenic, uncertain significance, likely benign, and benign, with an exceptional class for nonsense and frameshift variants, which are almost always pathogenic [46]. Following the ACMG recommendations, 824 out of 2104 *LDLR*, *APOB*, and *PCSK9* variants found in FH patients (655, 77, and 92, respectively) need functional characterization once the nonsense, frameshift, and the characterized ones are discarded [43]. Still, about 40% of all variants need functional evidence to be classified as pathogenic.

Originally, *in silico* tools were designed to give priority to *in vitro* characterization of those variants with higher pathogenic probabilities [47]. In recent years, prediction software has greatly improved in accuracy; however, conclusive diagnosis of a variant is still achieved by *in vitro* characterization or cascade screening. A predictive model considers different characteristics in order to assess the impact of a given variant on protein function, from the evolutionary conservation of an amino acid or nucleotide among homologous sequences to structure analysis [48]. Thus, several prediction software programs have been developed based on the analysis of each particular feature alone or by a combination of several of them.

Prediction software programs can be classified depending on the analyzed parameters (sequence conservation or structural and physicochemical parameters) or the technique used (machine learning and AI) [49]. The latest trend in the development of predictive software involves combining existing models to create innovative frameworks with enhanced accuracy and scope. This approach is revolutionizing the way models are developed, resulting in more robust variant assessment.

Each software has its own weaknesses and strengths, depending on the analyzed criteria, the developed algorithm, and the examined protein. In some cases, a model can be entirely focused on a specific protein, which usually improves the accuracy of the prediction. For these reasons, a prediction is more accurate as there is agreement/consensus among different predictive tools. The ACMG guidelines recommend the use of multiple software packages at once for a more reliable interpretation [45].

There is no guidance/recommendation on which software should be used or how many of them should agree for a prediction to be considered reliable [50•]. In this review, the most commonly used software for predicting the effect of a missense variant in *LDLR*, *APOB*, and *PCSK9* are described, both general and specific ones, to facilitate the selection of the most appropriate for a given variant. Below, the specific features and characteristics of the most commonly used software for predicting *LDLR*, *APOB*, and *PCSK9* missense variants are summarized.

General Predictive Software

Most predictive software has been trained with a wide variety of protein databases, so they are able to predict the pathogenicity of a broad number of proteins. However, they may fail to diagnose proteins with unique characteristics, or their accuracy may be lower.

Early Generation Predictive Software

Several models, including SIFT (Sorting Intolerant From Tolerant), Polyphen-2, and MutationTaster, have paved the way for the advancement of modern predictive tools. These models leverage a variety of features, such as sequence conservation, physicochemical properties of amino acids, and protein structural information, to assess the potential pathogenicity of genetic variants. These software tools were pioneers in the field of pathogenicity prediction and laid the foundation for subsequent developments. Despite their older age, they remain valuable and have been widely used in the scientific community.

SIFT SIFT (<https://sift.bii.a-star.edu.sg/>) is based on protein sequence homology and conservation to predict the pathogenicity of SNPs (single nucleotide polymorphisms) using Bayes [47]. SIFT classifies the queries as tolerated when the change is predicted to not compromise the protein's function, or not tolerated, when an alteration is predicted. This software presumes that well conserved amino acids are important, so its prediction relies solely on amino acidic sequence. Therefore, SIFT can evaluate missense variants only when homologous sequences are available, and it is especially suited for sequences with well-aligned orthologous sequences [51].

PolyPhen-2 Polymorphism Phenotyping v2 (PolyPhen-2) (<http://genetics.bwh.harvard.edu/pph2/>), one of the most widely used pathogenicity predicting software, uses both protein sequence- and structure-based features to evaluate variants, and the effect is predicted by a Naïve Bayesian

classifier [52]. The sequence-based features include position-specific independent counts (PSIC) [53] scores and multiple sequence alignment (MSA) [54] properties, and the position of variants in relation to domain limits. The structure-derived features are solvent accessibility, changes in solvent accessibility for buried residues, and crystallographic B-factor [48].

MutationTaster MutationTaster (<https://www.mutationtaster.org/>) is a software that integrates information from different biomedical databases and analyses evolutionary conservation, splice-site changes, loss of protein features, and changes that might affect the amount of mRNA by a Naïve Bayesian classifier [55]. MutationTaster contains data of SNPs and deletions from 1000 Genomes Project [56] and pathogenic variants found in ClinVar and Human Gene Mutation Database (HGMD) [57]. Common variants in the 1000 Genomes Project are automatically classified as benign, while variants annotated as pathogenic in ClinVar are automatically categorized as pathogenic. Although developed in 2014, a new version of MutationTaster implementing random forest was released in 2021.

Recent Generation Predictive Software

Recognizing the necessity for enhanced accuracy and broader coverage, more recent pathogenicity prediction models have emerged, further building upon the predictions of their predecessors. The latest trend in AI is ensemble models. Ensemble models, an innovative approach in predictive analytics, combine the outputs of multiple individual models to bolster accuracy and robustness in predictions. This technique's advantage lies in minimizing individual model limitations while capitalizing on their strengths, yielding more reliable outcomes. The guiding principle in designing ensemble methods has been “many heads are better than one” [58]. There are many ways of combining and weighting the “base” models, such as bagging (different data samples per model), boosting (sequential training), or stacking (same data samples for each model followed by a meta-model). Several models have been created with this approach in recent years, such as MetaLR and MetaSVM [59], Eigen [60], DANN [61], Condel [62], etc. In this review, we will focus on CADD (Combined Annotation-Dependent Depletion), REVEL (Rare Exome Variant Ensemble Learner), and Varity, some of the most accurate models.

The development and training of these models is much more complex, as they incorporate a larger number of features, and their training is more complex, often involving AI and larger datasets. However, the larger the amount of data available, the more accurate the model will be. The sum of these three factors (larger databases, models to rely on, and

the power of AI) has led to the development of predictive models with an accuracy never seen before.

REVEL REVEL (<https://sites.google.com/site/revelgenomics/>) is an ensemble method for predicting the pathogenicity of missense variants on the basis of combining many individual tools focused on rare variants [63]. REVEL uses bagging random forest technique to incorporate 18 pathogenicity scores from 13 prediction tools (MutPred, FATHMM v2.3, VEST 3.0, PolyPhen-2, SIFT, PROVEAN, MutationAssessor, MutationTaster, LRT, GERP + +, SiPhy, phyloP, and phastCons). The score depends on the preference of the user since the threshold can be modified to improve the sensitivity or the specificity. REVEL was suggested as the optimal in silico predictor by ClinGen [64] FH variant curation expert panel in the guidelines for *LDLR* variant classification [65].

The REVEL method has strengths in several dimensions. It was trained and evaluated on recently discovered disease-causing and neutral variants, similar to possible future variants found in NGS studies. REVEL's integration of a diverse set of predictors enriches its predictive power. In addition, REVEL's careful exclusion of training variants from its predictor components reduces overfitting problems. Demonstrating very high overall performance in independent evaluations, REVEL particularly excels in discerning disease-causing from rare neutral variants. Its value lies in prioritizing relevant variants amid the wealth of rare findings in sequencing analyses, facilitated by pre-calculated scores available for access. The use of the method extends to case–control studies at the genetic level, as evidenced by its adoption in the International Prostate Cancer Genetics Consortium or ClinGen. Its applicability to genes could be studied in the future to interpret variants of unknown significance in various clinical conditions.

CADD CADD (<https://cadd.gs.washington.edu/>) gives pathogenicity scores based on diverse genomic features derived from surrounding sequence context, gene model annotations, evolutionary constraint, epigenetic measurements, and functional predictions [66]. CADD considers the pathogenicity scores of SIFT or PolyPhen-2 but also takes into account models such as mirSVR [67] (ranks microRNA target sites) or Genomic Evolutionary Rate Profiling (GERP) [68] (evolutionary constraints). The model uses logistic regression to fit the data. A differential feature of this software is that CADD is not trained on characterized genomic variants; it uses less biased, much larger training sets. Another unique aspect of CADD is that it does not specify a cut-off value for pathogenicity scores above which a variant is declared pathogenic or benign.

Variety Variety (<http://variety.varianteffect.org/>) is a model optimized to detect rare variants [69]. It is based on four feature types:

conservation, physicochemical properties, protein–protein interaction, and structure-related properties. It considers more than 30 parameters, scores from pathogenicity predictive tools such as SIFT, PROVEAN [70] or Evm [71] among others, and uses random forest and logistic regressions to optimize the output. The model was specifically trained with rare (minor allele frequency [MAF] between 0.5 and 10^{-6}) and extremely rare ($MAF < 10^{-6}$) missense variants from ClinVar, although it is still very accurate with common variants. Being the most recent one, it was tested against CADD and REVEL, outperforming them.

The information provided above is illustrated in Fig. 3.

Predictive Models for LDLR Genetic Variants

In recent years, there has been a rapid development of predictive software for specific proteins, which has provided very specialized models capable of considering specific nuances of each protein. The development of a model of this type requires the existence of extensive databases containing many variants of the protein to be studied. In this way, and by applying machine learning models, the software can be primed with enough information so that it can provide accurate results. Due to the relatively high frequency of nonsense variants of *LDLR*, it has been possible to generate extensive databases (ClinVar, LOVD) that allow the development of highly accurate predictive models.

SFIP-MutID

Structure-based Functional Impact Prediction for Mutation Identification (SFIP-MutID) was developed using structural models of LDLR at both neutral and acidic pH [72]. The structures of LDLR were obtained from the protein data bank (PDB ID, 1F5Y, 1N7D), and to visualize the effect of each variant, homology modeling in Discovery Studio [73] was used. SFIP-MutID considers three aspects of the protein: the affected domain (variants in some domains are more likely to be pathogenic), the structure of the affected area, and, an energy-based score that classifies variants as destabilizing or not. However, this model does not cover certain regions of the LDLR because the structure of the protein is not completely resolved [72].

MLb-LDLr

The most recent LDLR-specific predictive software is Machine Learning-based LDLr (MLb-LDLr) [74]. This model is trained on the ClinVar database, where more than 1300 *LDLR* missense characterized variants are annotated (last update: November 2022). It considers seven features of the altered amino acid to give a prediction: conservation of the substituted residue, original and substituting amino





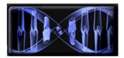
Early generation predictive software	Recent generation predictive software
<p>SIFT Sorting Intolerant From Tolerant</p>  <ul style="list-style-type: none"> -- Based mostly on amino acid conservation -- Uses amino acid substitution matrix to estimate the effect of a mutation -- Accurate when well-aligned orthologous sequences are available -- Not useful if no homologous sequence is available -- Naïve Bayes 	<p>Variety</p>  <ul style="list-style-type: none"> -- 4 features: Conservation, physicochemical properties, protein-protein interaction and structure-related properties -- More than 30 parameters -- Trained with rare pathogenic variants -- Random forest and logistic regression
<p>MutationTaster</p>  <ul style="list-style-type: none"> -- Based on conservation analysis, splice site analysis and protein features such as secondary structure, transmembrane domain, post-translational modifications -- If present, gives clinical information about the mutation -- Predicts intronic, single amino acid and complex alterations -- Naïve Bayes (2014), Random forest (2021) 	<p>CADD Combined annotation-Dependent Depletion</p>  <ul style="list-style-type: none"> -- Based on evolutionary conservation: if a variant is well preserved, it is probably neutral or benign, meanwhile a de novo mutation could be damaging -- Trained on less biased, much larger datasets instead of well characterized variants -- Two outputs: raw scores (relative values, more accurate, not comparable across models) and scaled scores (normalized values, less resolution) -- Logistic regression
<p>PolyPhen-2 Polymorphism Phenotyping v2</p>  <ul style="list-style-type: none"> -- Sequence-based features: amino acid properties (polarity, charge...), conservation and functional domains -- Structure-based features: Tertiary structures, stability, binding interfaces -- The prediction score ranges from benign to possibly damaging or probably damaging, with higher scores suggesting a higher impact -- Naïve Bayes 	<p>REVEL Rare Exome Variant Ensemble Learner</p> <ul style="list-style-type: none"> -- Ensemble method combining 13 individual tools and 18 pathogenicity scores -- Multiple threshold scores to increase sensitivity or specificity -- Trained with rare pathogenic variants -- Random forest

Fig. 3 Pathogenicity Prediction Software. Early generation: Based mostly on conservation, physicochemical properties, and structure. SIFT predicts pathogenicity of SNPs based on protein sequence homology based on the Bayes approach. It focuses on well-conserved amino acids and requires homologous sequences for accurate evaluation of missense variants. PolyPhen-2 utilizes both protein sequence- and structure-based features to evaluate variants using Naïve Bayes. It incorporates position-specific independent counts (PSIC) scores, multiple sequence alignment (MSA) properties, and structural characteristics to predict the effect of variants. MutationTaster integrates information from biomedical databases to analyze evolutionary conservation, splice-site changes, loss of protein features, and mRNA-related changes for predicting pathogenicity. It uses a Naïve Bayesian classifier. **Recent generation:** These modern models use ensemble

techniques that combine the output of multiple individual models to get a more precise result. REVEL uses random forest to integrate 13 prediction tools, focused on rare missense variants. It was trained with recently discovered variants to mimic possible variants found in the future. CADD provides pathogenicity scores based on diverse genomic features and considers pathogenicity scores of SIFT and Polyphen-2 among others. It employs large training sets and logistic regression to fit the data. CADD does not rely on a specific cut-off value for pathogenicity determination. Variety is based on 4 feature types: conservation, physicochemical properties, protein-protein interactions, and structure-related properties. It considers more than 30 parameters and uses random forest and logistic curve to optimize the output. Variety also focuses on rare missense variants

acids, charge, hydrophobicity, size change, and affected domain. MLb-LDLr is an open-access predictive software provided to the scientific community (<https://www.ehu.es/en/web/hypercholesterolemia-mechanisms/mlb-ldlr>). The

introduction of a machine learning algorithm provides a predictive model with a specificity of 92.5% and a sensitivity of 91.6%, which shows high accuracy in predicting both pathogenic and benign variants.

Predictive Models for APOB Genetic Variants

Compared to the large study carried out in *LDLR*, *APOB* genetic variants are not as well characterized, and pathogenicity predictions for this protein are much less reliable. While most proteins pathogenic variants tend to occur in highly conserved regions, *APOB* variants are distributed all over the protein [75]. In fact, the most used in silico tools have failed to correctly predict some of the most common *APOB* variants [76], except for genetic variants with a major involvement in receptor binding [20–22]. These results indicate that computer-based predictions of functional effects cannot yet reliably predict the effect of *APOB* SNVs. Several factors underlie the inaccuracy in the prediction of the effects of a missense apoB-100 variant, among them: the huge size of the protein, the lack of a crystallographic structure of the native protein, and the nature of protein folding within a lipid moiety, which cannot be addressed by in silico tools to study the protein–lipid interactions.

In addition to the pathogenic receptor binding variants of apoB-100, there exist other pathogenic variants that exert an impact on the structural integrity of the apoB molecule, leading to the impairment of very low-density lipoprotein (VLDL) and low-density lipoprotein (LDL) assembly processes. These genetic variants underlie the etiology of hereditary familial hypobetalipoproteinemia (FHBL), a clinical condition distinguished by compromised hepatic lipid secretion and limited transport to peripheral tissues. The spectrum of *APOB* pathogenic variants contributing to both biallelic *APOB*-FHBL and heterozygous *APOB*-FHBL is largely represented by frameshift, nonsense, and splice variants. These variants result in the production of a truncated apoB protein, characterized by an incomplete sequence, thereby perturbing its functional properties. Consequently, this perturbation results in marked reduction in levels of total cholesterol, LDL, VLDL, and serum triglycerides [77].

In this context, precise discrimination between variants that elevate and those that lower LDL cholesterol levels is crucial for clinical accuracy. Prediction software must effectively distinguish such pathogenic variants, holding paramount significance in guiding appropriate treatment strategies and understanding patient well-being implications. Given the multifaceted role of apoB-100, these considerations are pivotal for ensuring clinical management and genetic interpretation precision.

Predictive models for PCSK9 genetic variants

Predicting the functional consequences of missense variants in *PCSK9* is a challenging task due to the varied outcomes they can produce. While some variants result in reduced *PCSK9* function and are deemed benign, others lead to increased activity, thereby causing autosomal dominant hypercholesterolemia. Consequently, the prediction of

PCSK9 pathogenicity introduces additional complexity as software tools must not only anticipate deleterious effects on protein structure or conformation but also determine the variant's pathogenic, atheroprotective, or neutral nature.

Assessment of the performance of software tools such as SIFT and PolyPhen-2 in analyzing GOF and LOF variants in multiple genes has been conducted. The findings indicate that both software tools exhibit heightened sensitivity and specificity for LOF variants compared to GOF variants [78]. This behavior can be attributed to the fact that LOF variants often involve substantial amino acid substitutions with significant physicochemical changes, thereby instilling greater confidence in the predictions. Moreover, GOF variants are less prevalent, resulting in less extensive algorithm training. When applying this understanding to *PCSK9*, one can anticipate higher accuracy in predicting LOF variants, despite the existence of a greater number of reported *PCSK9* GOF variants. Nonetheless, certain regions of the *PCSK9* protein, such as the *LDLR* binding site and furin cleavage site [28], are well characterized, allowing for more accurate in silico predictions for variants occurring in these areas.

This highlights the intricate nature of prediction within the complex landscape of *PCSK9*'s functionality. Consequently, critical importance lies in the capacity of prediction software to accurately distinguish between GOF pathogenic variants and LOF cardioprotective variants. It becomes increasingly evident that the mere labeling of a *PCSK9* variant as “pathogenic” by software does not guarantee its attribution as a cause of FH. This is particularly relevant as software might interpret a LOF variant as pathogenic due to its failure to meet the criteria for benign classification, even though it may possess beneficial physiological outcomes.

Despite the complexities involved, the performance of commonly used pathogenicity predictive software for *LDLR*, *APOB*, and *PCSK9* can be assessed empirically. Table 1 summarizes the expected performance of these software tools based on the available data.

Structure-Modeling Software

Since the structure of a protein determines its function, protein structure prediction (PSP) is a major challenge in biochemistry. A protein's ability to fold into different conformations is essential for the viability of many biological processes. Therefore, knowing the 3D conformation of a protein is decisive for being able to predict the effect of a variant on its biological function [80, 81••].

The determination of a protein's structure has traditionally relied on high-resolution experimental techniques such as X-ray crystallography [63], NMR spectroscopy [64], and cryo-electron microscopy [65]. While these methods yield

Table 1 Pathogenicity predictive software expected performance on *LDLR*, *APOB*, and *PCSK9*

Pathogenicity predictive models	<i>LDLR74</i>	<i>APOB</i>	<i>PCSK9</i>
SIFT	High accuracy -Sn: 86% -Sp: 88%	Low accuracy -Lack of structure -Low conservation	Regular -Unknown mechanism -LOF > GOF79
PolyPhen-2	High accuracy -Sn: 93% -Sp: 88%	Low accuracy -Lack of structure	Regular -Unknown mechanism -LOF > GOF79
MutationTaster	High accuracy -Sn: 95% -Sp: 78%	Low accuracy -Low conservation	Regular -Unknown mechanism
REVEL	High accuracy* -Sn: 95% -Sp: 90%	Low accuracy -Low conservation -Low performance base models	Regular -Unknown mechanism -Low performance base models
CADD	High accuracy -Sn: 89% -Sp: 94%	Low accuracy -Low conservation	Regular -Unknown mechanism
Variety	High accuracy* -Sn: 94% -Sp: 91%	Low accuracy -Low conservation	Regular -Unknown mechanism -Low performance base models
SFIP-MutID	Regular -Sn: 90% -Sp: 22%	-	-
MLb-LDLr	High accuracy -Sn: 91% -Sp: 91%	-	-
Overall analysis	Reliable predictions	Unreliable predictions	Moderate predictions

Sn Sensitivity, *Sp* Specificity.

* Unpublished data currently under review in another study by Larrea-Sebal.

the most precise protein structures, the process of crystallization poses a significant bottleneck, particularly given the vast number of protein sequences to be solved [66]. This situation highlights the necessity of generating *in silico* models to provide accurate structure predictions. In this sense, two main PSP approaches have been used over the years: template-based modeling (TBM) and template-free modeling (FM) methods. TBM methods use the structural framework of existing proteins obtained from the PDB, while FM methods predict the structure without any template. The accuracy of TBM relies on the existence of evolutionary similar proteins, obtaining very precise predictions for proteins with high sequence identity (SI) templates. However, when SI drops below 30%, the accuracy of the model decreases [82]. In those cases, FM methods are more useful because they are based on physics- and knowledge-based energy functions. As they construct protein structures from scratch, they are often referred to as *ab initio* or *de novo* modeling approaches [79, 83]. In general, FM does not achieve the same accuracy as TBM, but the gap between both methods is narrowing thanks to the use of deep learning approaches [84, 85].

PSP is generally composed of four main components: the input, a trunk, an output, and a refinement module [86]. The primary input is typically the protein's primary sequence,

although modern models have incorporated additional information such as homologous multiple sequence alignment (MSA). The trunk component analyzes the input data and utilizes folding engines, empirical knowledge of sequence-structure properties [87, 88], or more recently, contact maps (binary matrices encoding residues likely to be in contact) [89, 90] to predict the protein's structure. The structures generated by the trunk component are subsequently transformed into 3D structures by the output module, which determines the atomic coordinates of the protein. Finally, the resulting 3D structure undergoes refinement, which includes the addition of side-chain atoms and optimization of the overall conformation. Unlike prediction software, the purpose of these models is to forecast the effect on the structure of the protein and not the pathogenicity of a variant. With that information, the effect of the variant can be inferred, either at the structural level or at the intermolecular level, if it affects the binding site with another molecule [91] (Fig. 4).

Although PSP algorithms have been used for a long time, very recently, a breakthrough has occurred with the development of AlphaFold, a new software that predicts the protein structure of an amino acid sequence, and allows visualizing and analyzing the results. AlphaFold is a structure prediction software developed by DeepMind

Protein Structure Prediction (PSP)

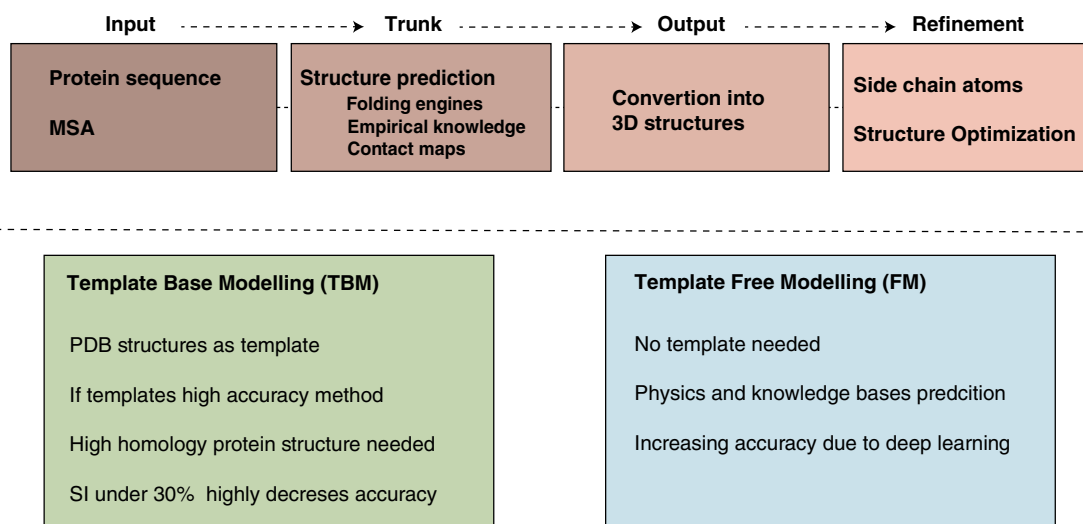


Fig. 4 Protein Structure Prediction Workflow. PSP models, comprising input, trunk, output, and refinement modules, enable protein structure prediction. The input, usually the protein's primary sequence, is analyzed by the trunk component using folding engines, sequence-structure knowledge, or contact maps to predict the structure. The output module converts the trunk-generated structures

into 3D structures by determining atomic coordinates. Refinement involves optimizing the conformation and adding side-chain atoms. PSP models focus on predicting structural effects rather than variant pathogenicity. Nevertheless, they provide valuable insights into the impact of variants on protein structure, including intermolecular interactions at binding sites

[85] that first appeared in the 13th edition of *Critical Assessment of Structure Prediction (CASP)* [92]. CASP is a community-wide assessment that tests the ability of various software to model protein structure from amino acid sequence [93]. In CASP13, AlphaFold received exceptional ratings, outperforming all other models by a wide margin. Because of this, AlphaFold was considered an anomalous leap in the field of protein structure prediction [94]. AlphaFold is based on co-evolution-dependent methods. These models work by detecting residues that co-evolved, i.e., that have mutated together over time, suggesting that both residues interact and are close to each other in 3D space [95]. Using this approach, it can be inferred whether two residues are in contact or not (binary contact matrices), which allows the acquisition of 3D coordinates. However, AlphaFold uses a more sophisticated co-evolution-dependent method based on RaptorX [96]. Instead of predicting binary contact, Raptor X predicts the distance between residues using discretized spatial ranges and calculates the mean and variance of the predicted distribution to localize each atom. Furthermore, AlphaFold uses a hundred-layer deep learning neural network to solve the structure, a technique never before applied to PSP [97]. All in all, AlphaFold, and its newer version AlphaFold 2 released for CASP14 (2020), are the most accurate PSP models available [98].

In addition to structure prediction, recent PSP algorithms also include the ability to predict the effect of missense variants using structure analysis [91, 99–109]. These algorithms predict 3D protein structure from amino acid sequences by searching for the most stable 3D state under native conditions. If a pathogenic variant destabilizes the molecule, then it can be expected that the algorithm will predict a more unfolded state [110]. However, these models are still in their very early stages of development and are not very accurate. For example, the predictive capability of AlphaFold-2 was tested on 2648 single-point missense variants over 121 proteins, with accurate results when compared to experimentally obtained structures [91]. Nevertheless, other studies reported that while AlphaFold is currently unable to predict structural effects of missense variants, it is conceivable that incorporation of experimental data and a database for storing structure-disrupting pathogenic variants will enable this feature in future versions of protein structure prediction programs [110, 111].

PSP software can be especially useful to study the effect of a missense variant on FH-related gene variants since many of the pathogenic variants occur at intermolecular interaction sites. ApoB-100 interacts with LDLR to be internalized, and the interaction of PCSK9 with LDLR regulates receptor expression. The residues involved in

both processes are well characterized, and due to the lack of the complete structure of LDLR and apoB-100, PSP software can be decisive.

Structure-Modeling of LDLR

LDLR is a well characterized protein, with tens of X-ray crystallographic and NMR solved structure-files available on Uniprot. However, none of them covers the entire protein, with most of them covering a few hundred residues with X-ray resolution ranging from 1.2 to 7 Å. The most complete solved structure covers the ecto-domain, amino acids 22–720 [112] (PDB ID 1N7D). There are crystals of both LDLR alone and LDLR with stabilizers such as PCSK9. The prediction made by AlphaFold is also available, but there are some discrepancies between the AlphaFold predicted structure and the crystal structure. LDLR requires calcium ions to enable correct folding of the protein's ligand binding domain and ligand binding [113]. Without calcium, LDLR is not functional, so crystallographic experiments have been carried out in the presence of calcium. However, AlphaFold does not consider this parameter, so the obtained prediction is not accurate. Despite this, AlphaFold can be used for the structural prediction of other domains not affected by calcium ions, such as the β -propeller or the EGF domains; however, most of the structures obtained from X-ray crystallography do not cover the LDLR C-terminal sequence (400–860 residues).

Structure-Modeling of apoB-100

Being part of a lipoprotein, apoB-100 interacts closely with lipids and remains with the lipoprotein throughout its metabolism [114]. ApoB-100 is highly insoluble in aqueous solution, and due to its nature, it has not been possible to obtain its structure by crystallography [115] or electron cryo-microscopy [116]. Solving the apoB-100 structure using PSP has not been successful either, and nowadays, the only predicted structure available on Uniprot (P04114) was obtained by SWISS-MODEL [117]. However, this PSP is not accurate, and does not resemble the belt-like structure that surrounds LDL particles.

Structure-Modeling of PCSK9

The structure of PCSK9 is well known, with plenty of PCSK9 structures solved by X-ray crystallography on Uniprot [118–120]. Most of the structures are accurate, covering the entire protein with a resolution of 2 to 3 Å. PCSK9 crystals have been obtained with and without adjuvants, sometimes even in complex with LDLR, which helps understanding how PCSK9 interacts with the receptor [121]. Structure prediction made by AlphaFold (PDB ID AF-Q8NBP7-F1)

is very accurate, resembling other PCSK9 crystallographic structures (PDB ID 6U3X, 6U2P). Accordingly, this makes it possible to use AlphaFold for predicting the structure of PCSK9 variants, allowing a more in-depth analysis of the variant's effect. In addition, by combining the extensive knowledge of the binding sites of LDLR and PCSK9, the crystallographic data allows obtaining accurate models to study the effect of a missense variant on the LDLR-PCSK9 complex formation.

Docking

Protein docking is the process of calculating the 3D structure of a protein complex starting from the individual structures of each protein [122]. It can be considered a further step in obtaining the structure of a protein complex since it also predicts how proteins interact. Molecular docking is a highly used technique in drug design, where it helps predict how ligands may bind to target proteins in 3D. However, protein–protein docking is not widely used [123] and is very helpful to delineate the interaction characteristics, the effect of a variant in the interaction, or affinity predictions [124••]. Intermolecular interactions are pivotal for many biological processes. These interactions are very specific, and any variant affecting the interaction interfaces is more harmful than others [125]. As it happens with the structure of a protein, early docking analysis relies on X-ray crystallography, NMR, and cryo-EM to study the interactions.

Since docking is based on structure, similarly to PSP models, the same two approaches can be taken to obtain the structures: “free” docking and template-based docking [126]. Template-based methods usually yield more accurate predictions when good templates are available, while “free” docking is advised when no template is available. However, docking techniques are also classified depending on the flexibility of the interacting proteins. Rigid-body techniques assume that proteins maintain their structure when interacting with other molecules, while flexible docking considers atoms' vibration, giving generally more accurate results [127].

The protein–protein docking process consists of two major steps: pose generation and scoring. Pose generation constitutes the first phase and serves to discard near-native structures, which is crucial for an accurate prediction. The native stage of both proteins is obtained by translating and rotating them until a few fitting poses are obtained. The scoring phase involves assigning scores to each possible conformation based on up to five characteristics, depending on the model. Force-field based scores consider non-bonded terms (van der Waals potential) and bonded terms (bond angle) [128, 129]. Empirical scores utilize intermolecular interactions and changes in accessible surface area

[130]. The knowledge-based score takes into account existing knowledge on protein interaction [131]. The last two scores are consensus-based [132] and machine learning-based [133].

The performance of docking methods is tested in the blind prediction challenge known as the Critical Assessment of Prediction of Interactions (CAPRI) [134], and it was designed on the model of CASP [135]. Both are community-wide experiments where different predictive models are tested, but whereas in CASP, protein folding is predicted from amino acid sequences, in CAPRI, protein assemblies are modeled by docking component structure [123]. Regarding CAPRI, both free and template-based docking models are tested, and the best scoring models are considered the most accurate ones. In one of the latest calls (2016) [136], the best models were SwarmDock [137], followed by ZDOCK [138], pyDock [139], HADDOCK [140], and Cluspro [141]. These models cover most docking methods; therefore, they are suitable for various types of docking problems.

Docking Software

ZDOCK

ZDOCK is a template-free rigid-body protein–protein docking program that uses a Fast Fourier Transform (FFT) algorithm that considers shape complementarity, electrostatics, and statistical potential for scoring. ZDOCK has been improved by adding the rescoring scheme called IRaPPa (Integrative Ranking of Protein–Protein Assemblies) that, briefly, uses physicochemical descriptors and combines a large selection of metrics to improve the selection of near-native solutions [133, 142]. The portal for running the program M-ZDOCK is available (<https://zlab.umassmed.edu/m-zdock/>) [143].

pyDOCK

Similarly to ZDOCK, pyDOCK is a template-free rigid-body docking program that uses FFT, but its main scoring functions are desolvation and electrostatics. PyDOCK has also been improved by IRaPPa, and it uses ZDOCK for the pose generation, meaning that both software results are similar. The web server for pyDOCKWEB is also freely available (<https://life.bsc.es/pid/pydockweb>) [144].

SwarmDock

SwarmDock, unlike ZDOCK and pyDOCK, is a flexible docking method that utilizes a particle swarm optimization

of 350 parameters to optimize the posing [145]. SwarmDock considers 17 structural parameters to optimize the conformation and relative position of each particle. IRaPPa has also been applied to this method.

HADDOCK

HADDOCK (High Ambiguity Driven DOCKing) is a semi-flexible docking protocol that uses empirical and bioinformatic scores to drive docking, especially van der Waals and Coulomb electrostatic energies. HADDOCK 2.4 is the software developed with this model, and it is among the most used ones [146].

Cluspro

Cluspro is a rigid-body docking method that relies on PIPER for pose generation, a docking program based on FFT [147]. Cluspro differs from other rigid body-based methods because, instead of using the lowest energy structure, it analyzes a cluster of the 1000 lowest ones, assuming that the real docked conformation is not necessarily the one with the lowest energy, but it will probably be in that cluster. Cluspro Web is also widely used software (<https://cluspro.bu.edu/login.php>) [141].

Use of Docking to Study LDLR-apoB-100 and LDLR-PCSK9 Interactions

Docking studies can be very helpful when predicting the effect of FH-related variants due to the receptor–ligand nature of LDLR with apoB-100 and PCSK9. A large number of these variants occur on the binding sites of these proteins, which could affect the affinity of the binding and cause FH. Regarding the most appropriate docking software to be used in each case, for LDLR-apoB-100 interaction, as no reliable apoB-100 structure is available, free docking software is advised. On the other hand, for assessing LDLR-PCSK9 interaction, as both PCSK9 and LDLR have well-characterized crystals, template-based software can be used.

LDLR-apoB-100 Interaction

The most LDLR affected domains by missense variants are the LBD and EGF-like domains, both key players in binding to apoB-100. Although these regions are highly conserved, not all the pathogenic variants located within these domains are correctly classified by predictive software. Subtle changes in the structure of the protein may lead to a failed prediction by the PSP software. In these cases, docking studies give an in-depth analysis of the effect of the variant on the binding of LDLR and apoB-100, considering not only the LDLR structure but also the interaction with apoB-100.

Docking software has previously been used as an approach to predict the effect of missense variants on the efficiency of binding to apoB-100 and describe the mechanism by which these LDLR variants cause FH [148]. Barbosa et al. characterized six LDLR variants using, among other techniques, docking models to assess the effect of the variant on the binding of LDLR to apoB-100 [149]. LDLR and apoB-100 sequences were obtained from Uniprot, the structure was generated by AlphaFold2, and the protein–protein docking was analyzed following a previously described protocol [148]. Briefly, a cluster of Cluspro, FireDock [150], Haddock, and Patchdock [151] software was used, and the results were ranked according to FFT score values. Overall, docking and molecular interactions analyses showed that p.(Cys184Tyr) and p.(Gly373Asp) LDLR variants alter the interaction of the receptor with apoB-100 [149].

Docking assays could also be performed the other way around, by analyzing the effect of apoB-100 variants on its binding to LDLR.

LDLR-PCSK9 Interaction

In the past few years, reliability of PCSK9-LDLR docking-based predictions has largely improved due to the increased number of described GOF- and LOF-PCSK9 variants [152–160] and the availability of several high-resolution crystallographically resolved PCSK9 structures. The structure of the PCSK9-LDLR complex has also been resolved, thus making the use of docking highly recommended to study the nature of their interaction. As shown in recent studies, docking can help to understand the mechanisms leading to pathological or beneficial effects. For example, docking has given new insights regarding the mechanisms of action of Ser127Arg [161] and Asp374Tyr [162] GOF PCSK9 variants [163]. These docking studies have indicated that the two pathogenic variants confer significantly higher binding affinity for LDLR as well as different binding modes, which impair LDLR from adapting its closed conformation.

AI-Driven Enhancement of Predictive Models in Bioinformatics

The integration of AI in bioinformatics has emerged as a pivotal solution to address the challenges posed by the overwhelming amount of biological data. AI techniques, such as machine learning and deep learning, have shown significant potential for handling and analyzing vast datasets, thereby enhancing the accuracy of predictive models. These AI-driven approaches enable researchers to extract meaningful patterns and relationships from diverse biological data sources, including genomics, proteomics, and transcriptomics. By leveraging AI algorithms, bioinformatics

researchers are able to develop more sophisticated prediction models that account for intricate interactions within biological systems.

The use of AI has drastically increased in clinical genomics. It has been applied in a wide range of conditions and approaches, such as patient photography analysis (facial analysis for disease identification, radiologic studies, microscopy data) [164], cardiology predictions (hypertension incident, atrial fibrillation, aortic stenosis) [165] blood biomarkers (mantle cell lymphoma [166], anemia [167]), interpretation of copy number variants [168], or classification of non-coding variants [169]. Regarding variant pathogenicity predictions, AI has revolutionized the field, providing advanced tools for accurate assessment. Starting from linear regressions to present-day deep learning models, the performance of these tools has increased exponentially. NGS technologies, platforms such as ClinVar, and advanced high-resolution crystallography techniques have led to the creation of extensive databases, enabling the development of highly precise predictive models. Models often have access to the position where the genetic variant occurs, evolutionary conservation of the position, prevalence of the variant in the population, probable effect of the variant on the mRNA or protein, affected domains, clinical phenotypes of the individuals, previously seen effects in other patients, family history, etc. AI can process and combine all this information, and the interpretation and weighing of each variable is the key to success.

Future Perspectives and Conclusions

The rapid development of bioinformatics tools has dramatically increased the accuracy of prediction software, both in assigning pathogenicity values to variants and in predicting their effect on the structure of the protein. Even so, there is still room for improvement, especially in obtaining protein structures and predicting protein–protein interactions. Most prediction algorithms are trained on empirical results, such as structures obtained by X-ray crystallography or functional in vitro characterization of variants, so the sensitivity and accuracy of in silico predictive tools are limited by the amount of available experimental data.

Algorithms can also be improved from within by applying deep learning techniques [170]. Deep learning has already been implemented in some PSP processes such as MSA, contact map prediction, or convolutional neural networks, and AlphaFold and RaptorX are based on these advanced machine learning tools. Regarding pathogenicity prediction models, the combination of ensemble techniques and deep learning has emerged in recent years. Until recently, ensemble and deep learning models

were regarded as separate methods in bioinformatics. Nowadays, the blending of these two popular techniques is causing a new wave of progress and the use of next-gen machine learning methods, known as ensemble deep learning [171, 172]. The next big breakthrough in the predictive software field will arise from ensemble deep learning implementation in every aspect of the algorithms.

Considering the bioinformatic tools mentioned in this review and their performance on the three most frequently FH-involved genes, different approaches should be taken to predict the pathogenicity of a variant. Pathogenicity prediction software is very accurate for *LDLR* variants, showing a very high hit rate when analyzing variants submitted in ClinVar, so this software should be the first approach to inferring the effect of an

LDLR variant. Then, depending on the location of the genetic variant, PSP and docking software can be used to analyze the possible effect on the biological function of the receptor. The 3D structure of the ectodomain is well characterized, but there is only one submission of the intracellular domain (778–860 residues) in Uniprot, and it does not cover the entire region. Finally, docking software should be considered for genetic variants occurring in the ectodomain, especially those affecting LBD or EGF-like domains, which interact with ligands, apoB-100 and PCSK9.

As for *APOB*, bioinformatic predictions seem to not be very accurate yet. It is discussed that this may be due to its interaction with lipids and its polymorphism, which affect both the conservation of residues and the quaternary

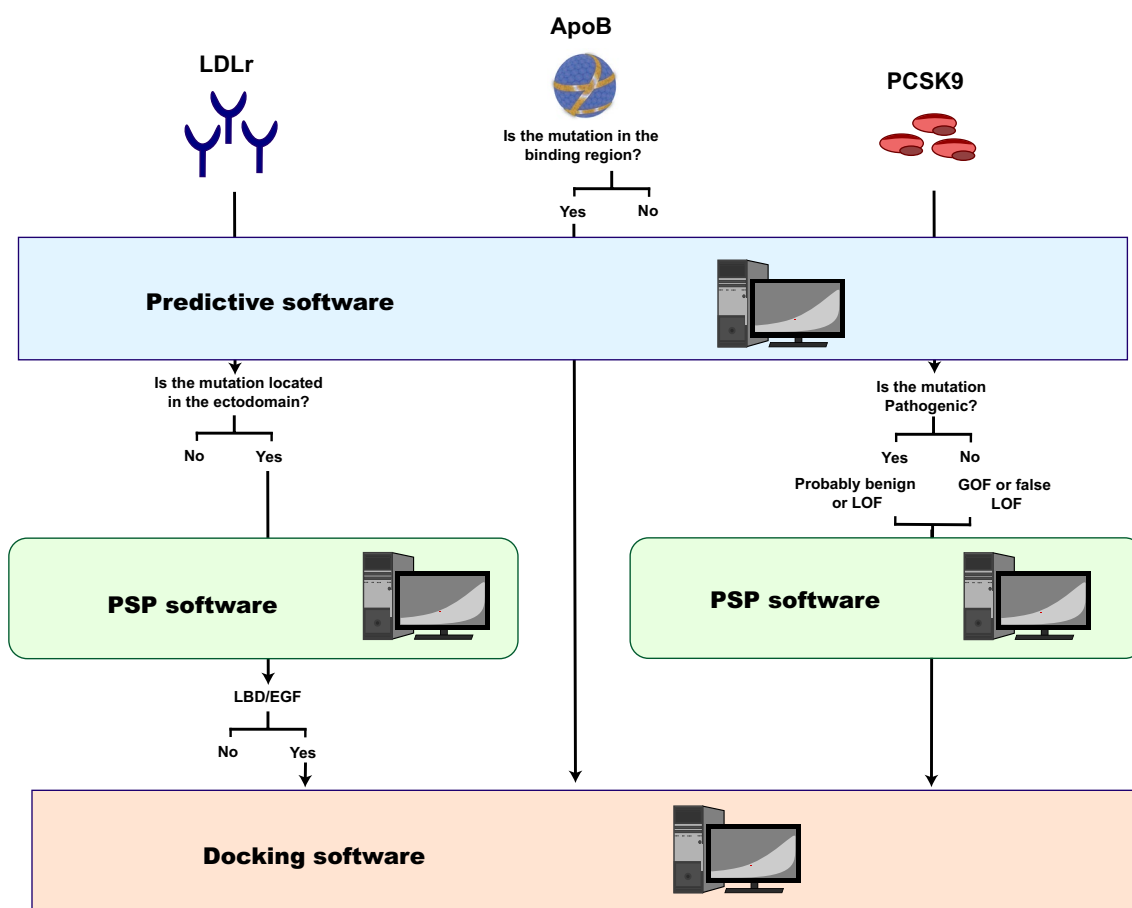


Fig. 5 Workflow illustrating the selection of bioinformatics tools based on the mutated gene and affected domain for accurate pathogenicity prediction. Predictive software demonstrates high reliability in assessing genetic variants within the *LDLR* gene. Specifically, when the variant occurs in the ectodomain, it is recommended to utilize PSP software. Conversely, if the variant affects the ligand-binding domain (LBD) or epidermal growth factor (EGF) domains, the analysis should be complemented with docking software for comprehensive evaluation. In the case of *APOB* variants, pathogenicity predictions can be considered reliable solely when the variants impact

the binding domain. In such instances, the integration of docking software can further enhance the analysis. However, it is important to note that pathogenicity predictions for *APOB* may be limited in other regions, requiring additional approaches for accurate assessment. Regarding *PCSK9* pathogenicity predictions, they provide valuable insights but may not offer a comprehensive diagnostic assessment of the variant. To obtain a more in-depth analysis, the combined utilization of predictive software and docking software is recommended. This integrated approach allows for a thorough investigation of *PCSK9* variants and their potential implications

structure of the protein. These two parameters, conservation and structure, are key factors for both pathogenicity prediction and PSP software, which hampers the usefulness of predictive software. Even with these drawbacks, bioinformatic tools have been used for *APOB* variant pathogenicity prediction and protein–protein docking assays with LDLR [149]; therefore, they should not be dismissed. Accurate discrimination between LDL cholesterol-raising and lowering variants is essential for future clinical precision. Effective differentiation in prediction software is crucial for guiding treatments and understanding patient health implications, especially given the complex role of apoB-100.

Regarding PCSK9, the structure of the protein is well described, and crystals covering the entire protein are available in Uniprot. However, due to the GOF or LOF variants, software prediction is not accurate. PSP software can accurately predict the effect of a variant on the 3D structure, but the interpretation of the results remains in the hands of the researcher. In any case, the use of this software reveals whether the variant affects the protein structure in any way, and if it does, this variant should be a candidate for in vitro characterization. The importance of future directions becomes evident in the vital task of software accurately distinguishing between GOF pathogenic variants and LOF variants. On the other hand, docking studies are more reliable, especially if the variant occurs in a binding region, since the direct interaction is analyzed. A full workflow of the software that should be used for each gene variant is shown in Fig. 5.

In conclusion, the integration of bioinformatics tools, protein structure modeling, and docking methodologies offers great promise for advancing our understanding of FH. These approaches provide valuable insights into the pathogenicity of genetic variants, protein structure–function relationships, and intermolecular interactions. Future advancements in AI and these methodologies hold the potential to enhance FH diagnosis, risk assessment, and the development of personalized treatment strategies for affected individuals.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This research was funded by Grupos Consolidados Gobierno Vasco 2021, grant number IT1720-22. A.L.-S. was supported by a grant PIF (2019–2020), Gobierno Vasco and partially supported by Fundación Biofísica Bizkaia. S.J.-B. was supported by a Margarita Salas Grant 2022 from the University of the Basque Country.

Declarations

Conflict of Interest All authors have nothing to declare. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could constitute a potential conflict of interest.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

Papers of particular interest, published recently, have been highlighted as:

- Of importance
- Of major importance

1. Brown MS, Goldstein JL. A Receptor-Mediated Pathway for Cholesterol Homeostasis. *Science*. 1986;232:34–47.
2. Ference BA, et al. Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European Atherosclerosis Society Consensus Panel. *Eur Heart J*. 2017;38:2459–2472.
3. Berberich AJ, Hegele RA. The complex molecular genetics of familial hypercholesterolaemia. *Nat Rev Cardiol*. 2019;16:9–20.
4. Vallejo-Vaz AJ, et al. Pooling and expanding registries of familial hypercholesterolaemia to assess gaps in care and improve disease management and outcomes: Rationale and design of the global EAS Familial Hypercholesterolaemia Studies Collaboration. *Atheroscler Suppl*. 2016;22:1–32.
5. Nordestgaard BG, et al. Familial hypercholesterolaemia is underdiagnosed and undertreated in the general population: guidance for clinicians to prevent coronary heart disease: Consensus Statement of the European Atherosclerosis Society. *Eur Heart J*. 2013;34:3478–90.
6. Austin MA. Genetic Causes of Monogenic Heterozygous Familial Hypercholesterolemia: A HuGE Prevalence Review. *Am J Epidemiol*. 2004;160:407–20.
7. Ahmad ZS, et al. US physician practices for diagnosing familial hypercholesterolemia: data from the CASCADE-FH registry. *J Clin Lipidol*. 2016;10:1223–9.
8. Risk of fatal coronary heart disease in familial hypercholesterolaemia. Scientific Steering Committee on behalf of the Simon Broome Register Group. *BMJ*. 1991;303:893–896.
9. Benito-Vicente A, et al. The importance of an integrated analysis of clinical, molecular, and functional data for the genetic diagnosis of familial hypercholesterolemia. *Genet Med*. 2015;17:980–8.
10. Jialal I, Barton Duell P. Diagnosis of Familial Hypercholesterolemia: Table 1. *Am J Clin Pathol*. 2016;145:437–439.
11. Lehrman MA, Goldstein JL, Brown MS, Russell DW, Schneider WJ. Internalization-defective LDL receptors produced by genes with nonsense and frameshift mutations that truncate the cytoplasmic domain. *Cell*. 1985;41:735–43.

12. Yokode M, et al. Cytoplasmic sequence required for basolateral targeting of LDL receptor in livers of transgenic mice. *J Cell Biol.* 1992;117:39–46.
13. Esser V, Limbird LE, Brown MS, Goldstein JL, Russell DW. Mutational analysis of the ligand binding domain of the low density lipoprotein receptor. *J Biol Chem.* 1988;263:13282–90.
14. Zhang D-W, et al. Binding of Proprotein Convertase Subtilisin/Kexin Type 9 to Epidermal Growth Factor-like Repeat A of Low Density Lipoprotein Receptor Decreases Receptor Recycling and Increases Degradation. *J Biol Chem.* 2007;282:18602–12.
15. Benito-Vicente A, et al. Validation of LDLr Activity as a Tool to Improve Genetic Diagnosis of Familial Hypercholesterolemia: A Retrospective on Functional Characterization of LDLr Variants. *IJMS.* 2018;19:1676.
16. Knott TJ, et al. Complete protein sequence and identification of structural domains of human apolipoprotein B. *Nature.* 1986;323:734–8.
17. Mahley RW, Innerarity TL, Rall SC, Weisgraber KH. Plasma lipoproteins: apolipoprotein structure and function. *J Lipid Res.* 1984;25:1277–94.
18. Schumaker VN, Phillips ML, Chatterton JE. Apolipoprotein B and Low-Density Lipoprotein Structure: Implications for Biosynthesis of Triglyceride-Rich Lipoproteins. In: *Advances in Protein Chemistry* vol. 45, Elsevier; 1994. 205–248.
19. Hevonoja T, Pentikäinen MO, Hyvönen MT, Kovanen PT, Ala-Korpela M. Structure of low density lipoprotein (LDL) particles: basis for understanding molecular changes in modified LDL. *Biochim Biophys Acta.* 2000;1488(3):189–210. [https://doi.org/10.1016/s1388-1981\(00\)00123-2](https://doi.org/10.1016/s1388-1981(00)00123-2).
20. März W, et al. Accumulation of ‘Small Dense’ Low Density Lipoproteins (LDL) in a Homozygous Patient with Familial Defective Apolipoprotein B-100 Results from Heterogenous Interaction of LDL Subfractions with the LDL Receptor. *J Clin Invest.* 1993;92:2922–33.
21. Kriško A, Etchebest C. Theoretical model of human apolipoprotein B100 tertiary structure. *Proteins.* 2006;66:342–58.
22. Innerarity TL, et al. Familial defective apolipoprotein B-100: low density lipoproteins with abnormal receptor binding. *Proc Natl Acad Sci USA.* 1987;84:6919–23.
23. Borén J, Ekström U, Ågren B, Nilsson-Ehle P, Innerarity TL. The Molecular Mechanism for the Genetic Disorder Familial Defective Apolipoprotein B100. *J Biol Chem.* 2001;276:9214–8.
24. Thomas ERA, et al. Identification and biochemical analysis of a novel APOB mutation that causes autosomal dominant hypercholesterolemia. *Mol Genet Genomic Med.* 2013;1:155–61.
25. Alves AC, Etchebarria A, Soutar AK, Martin C, Bourbon M. Novel functional APOB mutations outside LDL-binding region causing familial hypercholesterolaemia. *Hum Mol Genet.* 2014;23:1817–28.
26. Seidah NG. The Proprotein Convertases, 20 Years Later. In: Mbikay M, Seidah, NG, editors. *Proprotein Convertases* vol. 768. Humana Press; 2011 23–57.
27. Benjannet S, et al. NARC-1/PCSK9 and Its Natural Mutants. *J Biol Chem.* 2004;279:48865–75.
28. Lopez D. PCSK9: An enigmatic protease. *Biochim Biophys Acta (BBA) - Mol Cell Biol Lipids.* 2008;1781:184–191.
29. Maxwell KN, Breslow JL. Proprotein convertase subtilisin kexin 9: the third locus implicated in autosomal dominant hypercholesterolemia. *Curr Opin Lipidol.* 2005;16:167–72.
30. Cohen J, et al. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat Genet.* 2005;37:161–5.
31. Abifadel M, et al. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nat Genet.* 2003;34:154–6.
32. Uribe KB, et al. A Systematic Approach to Assess the Activity and Classification of PCSK9 Variants. *IJMS.* 2021;22:13602.
33. Iacocca MA, Dron JS, Hegele RA. Progress in finding pathogenic DNA copy number variations in dyslipidemia. *Curr Opin Lipidol.* 2019;30:63–70.
34. Khafizov K, Madrid-Aliste C, Almo SC, Fiser A. Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. *Proc Natl Acad Sci USA.* 2014;111:3733–8.
35. Aradhya S, et al. Applications of artificial intelligence in clinical laboratory genomics. *Am J Med Genet Pt C ajmg.c.32057 (2023).* <https://doi.org/10.1002/ajmg.c.32057>.
36. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to Machine Learning, Neural Networks, and Deep Learning. *Neural Netw Transl Vis Sci Technol.* 2020;27:14.
37. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol Adv.* 2021;49:107739.
38. Jiang T, Gradus JL, Rosellini AJ. Supervised Machine Learning: A Brief Primer. *Behav Ther.* 2020;51:675–87.
39. Badillo S, et al. An Introduction to Machine Learning. *Clin Pharmacol Ther.* 2020;107:871–85.
40. Zhang Z, et al. Deep learning in omics: a survey and guideline. *Brief Funct Genomics.* 2019;18:41–57.
41. Ickstadt K, Schäfer M, Zucknick M. Toward Integrative Bayesian Analysis in Molecular Biology. *Annu Rev Stat Appl.* 2018;5:141–67.
42. Qi H, et al. MVP predicts the pathogenicity of missense variants by deep learning. *Nat Commun.* 2021;12:510.
43. Chora JR, Medeiros AM, Alves AC, Bourbon M. Analysis of publicly available LDLR, APOB, and PCSK9 variants associated with familial hypercholesterolemia: application of ACMG guidelines and implications for familial hypercholesterolemia diagnosis. *Genet Med.* 2018;20:591–8.
44. Bourbon M, Alves AC, Medeiros AM, Silva S, Soutar AK. Familial hypercholesterolaemia in Portugal. *Atherosclerosis.* 2008;196:633–42.
45. Richards S, 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. *Genet Med.* 2015;17:405–24.
46. Bourbon M, Alves AC, Sijbrands EJ. Low-density lipoprotein receptor mutational analysis in diagnosis of familial hypercholesterolemia. *Curr Opin Lipidol.* 2017;28:120–9.
47. Sim N-L, et al. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* 2012;40:W452–7.
48. Thusberg J, Olatubosun A, Vihinen M. Performance of mutation pathogenicity prediction methods on missense variants. *Hum Mutat.* 2011;32:358–68.
49. Garcia FADO, Andrade ESD, Palmero EI. Insights on variant analysis in silico tools for pathogenicity prediction. *Front Genet.* 2022;13:1010327.
50. Gunning, A. C. et al. Assessing performance of pathogenicity predictors using clinically relevant variant datasets. *J Med Genet.* 2021;58:547–555. **This study holds importance as it presents an independent validation of pathogenicity predictors, utilizing both “open” and “clinically representative” datasets, to assess the performance of recent meta-predictors and commonly used in silico tools, revealing superior performance of meta-predictors, particularly REVEL, in a clinically relevant context and discouraging the use of a consensus-based approach in current practice.**

51. Ng PC, Henikoff S. Predicting Deleterious Amino Acid Substitutions. *Genome Res.* 2001;11:863–74.
52. Adzhubei IA, et al. A method and server for predicting damaging missense mutations. *Nat Methods.* 2010;7:248–9.
53. Sunyaev SR, et al. PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng Des Sel.* 1999;12:387–94.
54. Chatzou M, et al. Multiple sequence alignment modeling: methods and applications. *Brief Bioinform.* 2016;17:1009–23.
55. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. Mutation-Taster evaluates disease-causing potential of sequence alterations. *Nat Methods.* 2010;7:575–6.
56. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature.* 2012;491:56–65.
57. Stenson PD, et al. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet.* 2014;133:1–9.
58. Altman N, Krzywinski M. Ensemble methods: bagging and random forests. *Nat Methods.* 2017;14:933–4.
59. Dong C, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet.* 2015;24:2125–37.
60. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet.* 2016;48:214–20.
61. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics.* 2015;31:761–3.
62. González-Pérez A, López-Bigas N. Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. *Am J Hum Genet.* 2011;88:440–9.
63. Ioannidis NM, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet.* 2016;99:877–85.
64. Rehm HL, et al. ClinGen — The Clinical Genome Resource. *N Engl J Med.* 2015;372:2235–42.
65. Chora JR, et al. The Clinical Genome Resource (ClinGen) Familial Hypercholesterolemia Variant Curation Expert Panel consensus guidelines for LDLR variant classification. *Genet Med.* 2022;24:293–306.
66. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 2019;47:D886–94.
67. Betel D, Koppal A, Agius P, Sander C, Leslie C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.* 2010;11:R90.
68. Davydov EV, et al. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Comput Biol.* 2010;6:e1001025.
69. Wu Y, et al. Improved pathogenicity prediction for rare human missense variants. *Am J Hum Genet.* 2021;108:1891–906.
70. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics.* 2015;31:2745–7.
71. Hopf TA, et al. Mutation effects predicted from sequence covariation. *Nat Biotechnol.* 2017;35:128–35.
72. Guo J, et al. Systematic prediction of familial hypercholesterolemia caused by low-density lipoprotein receptor missense mutations. *Atherosclerosis.* 2019;281:1–8.
73. BIOVIA Discovery Studio 2017 R2: A comprehensive predictive science application for the Life Sciences. 2017.
74. Larrea-Sebal A, et al. MLB-LDLr. JACC: Basic to Translational Science. 2021;6:815–27.
75. Benn M, et al. Common and Rare Alleles in Apolipoprotein B Contribute to Plasma Levels of Low-Density Lipoprotein Cholesterol in the General Population. *J Clin Endocrinol Metab.* 2008;93:1038–45.
76. Benn M. Apolipoprotein B levels, APOB alleles, and risk of ischemic cardiovascular disease in the general population, a review. *Atherosclerosis.* 2009;206:17–30.
77. Burnett JR, ChB M, Hooper AJ, Hegele A. APOB-Related Familial Hypobetalipoproteinemia.
78. Flanagan SE, Patch A-M, Ellard S. Using SIFT and PolyPhen to Predict Loss-of-Function and Gain-of-Function Mutations. *Genet Test Mol Biomarkers.* 2010;14:533–7.
79. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol.* 1997;268:209–25.
80. Pre and post-AlphaFold. Ismi, D. P., Pulungan, R., & Afiahayati. Deep learning for protein secondary structure prediction. *Comput Struct Biotechnol J.* 2022;20:6271–86.
81. ●● Jisna VA, Jayaraj PB. Protein Structure Prediction: Conventional and Deep Learning Perspectives. *Protein J* 2021;40:522–544. **This work is of outstanding importance as it provides a comprehensive review of the transformative impact of deep neural networks on protein secondary structure prediction, highlighting recent advancements and potential future directions to enhance accuracy and expand the scope of the field.**
82. Kryshtafovych A, et al. Evaluation of the template-based modeling in CASP12. *Proteins.* 2018;86:321–34.
83. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field: QUARK Ab initio Prediction Method. *Proteins.* 2012;80:1715–35.
84. Zheng W, et al. Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins.* 2019;87:1149–64.
85. Senior AW, et al. Improved protein structure prediction using potentials from deep learning. *Nature.* 2020;577:706–10.
86. AlQuraishi M. Machine learning in protein structure prediction. *Curr Opin Chem Biol.* 2021;65:1–8.
87. Leman JK, et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat Methods.* 2020;17:665–80.
88. Yang J, et al. The I-TASSER Suite: protein structure and function prediction. *Nat Methods.* 2015;12:7–8.
89. Golkov V, et al. Protein contact prediction from amino acid co-evolution using convolutional networks for graph-valued images. In: Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R, editors. *Advances in Neural Information Processing Systems* 29 (NIPS 2016). 2016.
90. Jones DT, Singh T, Kosciolk T, Tetchner S. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics.* 2015;31:999–1006.
91. Akdel M, et al. A structural biology community assessment of AlphaFold2 applications. *Nat Struct Mol Biol.* 2022;29:1056–67.
92. Cheng J, et al. Estimation of model accuracy in CASP13. *Proteins.* 2019;87:1361–77.
93. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins.* 2019;87:1011–20.
94. Marcu Ş-B, Tăbîrcă S, Tangney M. An Overview of AlphaFold’s Breakthrough. *Front Artif Intell.* 2022;5:875587.
95. Lapedes AS, Giraud B, Liu L, Stormo GD. Correlated mutations in models of protein sequences: phylogenetic and structural effects. In: *Institute of Mathematical Statistics Lecture Notes - Monograph Series.* Institute of Mathematical Statistics: 1999. 236–256. <https://doi.org/10.1214/lnms/1215455556>.
96. Xu J. Distance-based protein folding powered by deep learning. *Proc Natl Acad Sci USA.* 2019;116:16856–65.

97. AlQuraishi M. AlphaFold at CASP13. *Bioinformatics*. 2019;35:4862–5.
98. Ozden B, Kryshchak A, Karaca E. Assessment of the CASP14 assembly predictions. *Proteins*. 2021;89:1787–99.
99. Delgado J, Radosky LG, Cianferoni D, Serrano L. FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics*. 2019;35:4168–4169.
100. Park H, et al. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *J Chem Theory Comput*. 2016;12:6201–12.
101. Rodrigues CHM, Pires DEV, Ascher DB. DYNAMUT2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Sci*. 2021;30:60–9.
102. Porta-Pardo E, Hrabe T, Godzik A. Cancer3D: understanding cancer mutations through protein structures. *Nucleic Acids Res*. 2015;43:D968–73.
103. Pires DEV, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*. 2014;30:335–42.
104. Porta-Pardo E, Godzik A. Mutation Drivers of Immunological Responses to Cancer. *Cancer Immunol Res*. 2016;4:789–98.
105. Sundaram L, et al. Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet*. 2018;50:1161–70.
106. Woodard J, Zhang C, Zhang Y. ADDRESS: A Database of Disease-associated Human Variants Incorporating Protein Structure and Folding Stabilities. *J Mol Biol*. 2021;433:166840.
107. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res*. 2005;33:W306–W310.
108. Tokuriki N, Tawfik DS. Stability effects of mutations and protein evolvability. *Curr Opin Struct Biol*. 2009;19:596–604.
109. Quan L, Lv Q, Zhang Y. STRUM: structure-based prediction of protein stability changes upon single-point mutation. *Bioinformatics*. 2016;32:2936–46.
110. Pak MA, et al. Using AlphaFold to predict the impact of single mutations on protein stability and function. 2021. <http://biorxiv.org/lookup/doi/10.1101/2021.09.19.460937>.
111. Buel GR, Walters KJ. Can AlphaFold2 predict the impact of missense mutations on structure? *Nat Struct Mol Biol*. 2022;29:1–2.
112. Rudenko G, et al. Structure of the LDL Receptor Extracellular Domain at Endosomal pH. *Science, New Series*. 2002;298:2353–8.
113. Simmons T, Newhouse YM, Arnold KS, Innerarity TL, Weisgraber KH. Human Low Density Lipoprotein Receptor Fragment. *J Biol Chem*. 1997;272:25531–6.
114. Kane JP. Apolipoprotein B: Structural and Metabolic Heterogeneity. *Annu Rev Physiol*. 1983;45:637–50.
115. Segrest JP, Jones MK, De Loof H, Dashti N. Structure of apolipoprotein B-100 in low density lipoproteins. *J Lipid Res*. 2001;42:1346–67.
116. Kumar V, et al. Three-Dimensional cryoEM Reconstruction of Native LDL Particles to 16Å Resolution at Physiological Body Temperature. *PLoS One*. 2011;6:e18841.
117. Waterhouse A, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*. 2018;46:W296–303.
118. Tibolla G, Norata GD, Artali R, Meneghetti F, Catapano AL. Proprotein convertase subtilisin/kexin type 9 (PCSK9): From structure–function relation to therapeutic inhibition. *Nutr Metab Cardiovasc Dis*. 2011;21:835–43.
119. Pedro-Botet J, Badimón L. PCSK9: estructura y función. PCSK9 y receptor de lipoproteínas de baja densidad. Mutaciones y cambios derivados de estas. *Clín Investig Arterioscler*. 2016;28:3–8.
120. Piper DE, et al. The Crystal Structure of PCSK9: A Regulator of Plasma LDL-Cholesterol. *Structure*. 2007;15:545–52.
121. Bottomley MJ, et al. Structural and Biochemical Characterization of the Wild Type PCSK9-EGF(AB) Complex and Natural Familial Hypercholesterolemia Mutants. *J Biol Chem*. 2009;284:1313–23.
122. Ritchie D. Recent Progress and Future Directions in Protein-Protein Docking. *CPPS*. 2008;9:1–15.
123. Janin J. Protein–protein docking tested in blind predictions: the CAPRI experiment. *Mol Biosyst*. 2010;6:2351.
- 124.●● Sunny S, Jayaraj PB. Protein–Protein Docking: Past, Present, and Future. *Protein J*. 2022;41:1–26. **This article holds outstanding importance as it provides a comprehensive assessment of existing computational docking algorithms, their challenges, and future prospects in the critical area of protein interaction prediction, emphasizing the potential role of artificial intelligence to address current limitations and lead the field towards more accurate and reliable results.**
125. David A, Sternberg MJE. The Contribution of Missense Mutations in Core and Rim Residues of Protein-Protein Interfaces to Human Disease. *J Mol Biol*. 2015;427:2886–98.
126. Porter KA, Desta I, Kozakov D, Vajda S. What method to use for protein–protein docking? *Curr Opin Struct Biol*. 2019;55:1–7.
127. Fan J, Fu A, Zhang L. Progress in molecular docking. *Quant Biol*. 2019;7:83–89.
128. Feng T, et al. HawkRank: a new scoring function for protein–protein docking based on weighted energy terms. *J Cheminform*. 2017;9:66.
129. Kynast P, Derreumaux P, Strodel B. Evaluation of the coarse-grained OPEP force field for protein-protein docking. *BMC Biophys*. 2016;9:4.
130. Roy AA, Dhawanjewar AS, Sharma P, Singh G, Madhusudhan MS. Protein Interaction Z Score Assessment (PIZSA): an empirical scoring scheme for evaluation of protein–protein interactions. *Nucleic Acids Res*. 2019;47:W331–7.
131. Andreani J, Faure G, Guerois R. InterEvScore: a novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution. *Bioinformatics*. 2013;29:1742–9.
132. Chermak E, et al. CONSRANK: a server for the analysis, comparison and ranking of docking models based on inter-residue contacts. *Bioinformatics*. 2015;31:1481–3.
133. Moal IH, et al. IRAPPA: information retrieval based integration of biophysical models for protein assembly selection. *Bioinformatics*. 2017;33:1806–13.
134. Janin J, et al. CAPRI: A Critical Assessment of PRredicted Interactions. *Proteins*. 2003;52:2–9.
135. Mintseris J, et al. Protein-protein docking benchmark 2.0: An update. *Proteins* 2005;60:214–216.
136. Lensink MF, Velankar S, Wodak SJ. Modeling protein-protein and protein-peptide complexes: CAPRI 6th edition: Modeling Protein-Protein and Protein-Peptide Complexes. *Proteins*. 2017;85:359–77.
137. Moal, I. H., Chaleil, R. A. G. & Bates, P. A. Flexible Protein-Protein Docking with SwarmDock. In: Marsh JA, editor. *Protein Complex Assembly* vol. 1764. Springer: New York, 2018. 413–428.
138. Pierce BG, et al. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics*. 2014;30:1771–3.
139. Pons C, Solernou A, Perez-Cano L, Grosdidier S, Fernandez-Recio J. Optimization of pyDock for the new CAPRI challenges: Docking of homology-based models, domain-domain assembly and protein-RNA binding. *Proteins*. 2010;78:3182–8.
140. Vangone A, et al. Sense and simplicity in HADDOCK scoring: Lessons from CASP-CAPRI round 1: HADDOCK in CASP-CAPRI Round 1. *Proteins*. 2017;85:417–23.
141. Kozakov D, et al. The ClusPro web server for protein–protein docking. *Nat Protoc*. 2017;12:255–78.

142. Vreven T, et al. Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J Mol Biol.* 2015;427:3031–41.
143. Pierce B, Tong W, Weng Z. M-ZDOCK: a grid-based approach for Cn symmetric multimer docking. *Bioinformatics.* 2005;21:1472–8.
144. Jiménez-García B, Pons C, Fernández-Recio J. pyDockWEB: a web server for rigid-body protein–protein docking using electrostatics and desolvation scoring. *Bioinformatics.* 2013;29:1698–9.
145. Poli R, Kennedy J, Blackwell T. Particle swarm optimization: An overview. *Swarm Intell.* 2007;1:33–57.
146. van Zundert GCP, et al. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J Mol Biol.* 2016;428:720–725.
147. Kozakov D, Brenke R, Comeau SR, Vajda S. PIPER: An FFT-based protein docking program with pairwise potentials. *Proteins.* 2006;65:392–406.
148. Borges JB, et al. Genomics, epigenomics and pharmacogenomics of familial hypercholesterolemia (FHBGEP): A study protocol. *Res Soc Adm Pharm.* 2021;17:1347–55.
149. Barbosa TKA, et al. LDLR missense variants disturb structural conformation and LDLR activity in T-lymphocytes of Familial hypercholesterolemia patients. *Gene.* 2023;853:147084.
150. Mashiach E, Schneidman-Duhovny D, Andrusier N, Nussinov R, Wolfson HJ. FireDock: a web server for fast interaction refinement in molecular docking. *Nucleic Acids Res.* 2008;36:W229–32.
151. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res.* 2005;33:W363–7.
152. Kwon GP, Schroeder JL, Amar MJ, Remaley AT, Balaban RS. Contribution of Macromolecular Structure to the Retention of Low-Density Lipoprotein at Arterial Branch Points. *Circulation.* 2008;117:2919–27.
153. Bergeron N, Phan BAP, Ding Y, Fong A, Krauss RM. Proprotein Convertase Subtilisin/Kexin Type 9 Inhibition. *Circulation.* 2015;132:1648–66.
154. Glerup S, Schulz R, Laufs U, Schlüter K-D. Physiological and therapeutic regulation of PCSK9 activity in cardiovascular disease. *Basic Res Cardiol.* 2017;112:32.
155. Homer VM, et al. Identification and characterization of two non-secreted PCSK9 mutants associated with familial hypercholesterolemia in cohorts from New Zealand and South Africa. *Atherosclerosis.* 2008;196:659–66.
156. Abifadel M, et al. Identification and characterization of new gain-of-function mutations in the PCSK9 gene responsible for autosomal dominant hypercholesterolemia. *Atherosclerosis.* 2012;223:394–400.
157. Fasano T, Sun X-M, Patel DD, Soutar AK. Degradation of LDLR protein mediated by ‘gain of function’ PCSK9 mutants in normal and ARH cells. *Atherosclerosis.* 2009;203:166–71.
158. Lagace TA, et al. Secreted PCSK9 decreases the number of LDL receptors in hepatocytes and in livers of parabiotic mice. *J Clin Invest.* 2006;116:2995–3005.
159. Nassoury N, et al. The Cellular Trafficking of the Secretory Proprotein Convertase PCSK9 and Its Dependence on the LDLR. *Traffic.* 2007;8:718–32.
160. Fisher TS, et al. Effects of pH and Low Density Lipoprotein (LDL) on PCSK9-dependent LDL Receptor Regulation. *J Biol Chem.* 2007;282:20502–12.
161. Timms KM, et al. A mutation in PCSK9 causing autosomal-dominant hypercholesterolemia in a Utah pedigree. *Hum Genet.* 2004;114:349–53.
162. Leren T. Mutations in the PCSK9 gene in Norwegian subjects with autosomal dominant hypercholesterolemia: Mutations in the PCSK9 gene. *Clin Genet.* 2004;65:419–22.
163. Martin WR, Lightstone FC, Cheng F. In Silico Insights into Protein-Protein Interaction Disruptive Mutations in the PCSK9-LDLR Complex. *IJMS.* 2020;21:1550.
164. Ledgister Hanchard SE, et al. Scoping review and classification of deep learning in medical genetics. *Genet Med.* 2022;24:1593–1603.
165. Busnatu Ștefan, et al. Clinical Applications of Artificial Intelligence—An Updated Overview. *JCM.* 2022;11:2265.
166. Carreras J, Nakamura N, Hamoudi R. Artificial Intelligence Analysis of Gene Expression Predicted the Overall Survival of Mantle Cell Lymphoma and a Large Pan-Cancer Series. *Healthcare.* 2022;10:155.
167. Dutra VDF, Biassi TP, Figueiredo MS. Sick cell anemia: hierarchical cluster analysis and clinical profile in a cohort in Brazil. *Hematol Transfus Cell Ther.* 2023;45:45–51.
168. Sládeček T, et al. Combination of expert guidelines-based and machine learning-based approaches leads to superior accuracy of automated prediction of clinical effect of copy number variations. *Sci Rep.* 2023;13:10531.
169. Moyon L, Berthelot C, Louis A, Nguyen NTT, RoestCrollius H. Classification of non-coding variants with high pathogenic impact. *PLoS Genet.* 2022;18:e1010191.
170. Pakhrin SC, Shrestha B, Adhikari B, Kc DB. Deep Learning-Based Advances in Protein Structure Prediction. *IJMS.* 2021;22:5553.
171. Cao Y, Geddes TA, Yang JYH, Yang P. Ensemble deep learning in bioinformatics. *Nat Mach Intell.* 2020;2:500–8.
172. Mohammed A, Kora R. A comprehensive review on ensemble deep learning: Opportunities and challenges. *J King Saud Univ - Comput Inf Sci.* 2023;35:757–74.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.