

**MACHINE-LEARNING-BASED INTERPRETATION OF RARE DISEASE VARIANTS  
LEVERAGING GENOMICS AND COMPUTATIONAL STRUCTURAL BIOLOGY**

By

**Souhrid Mukherjee**

Dissertation

Submitted to the Faculty of the  
Graduate School of Vanderbilt University  
in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY  
in  
BIOLOGICAL SCIENCES

JUNE 30th, 2022

Nashville, Tennessee

Approved:

Antonis Rokas Ph.D. (chair)

Nicole Creanza Ph.D.

Carlos Lopez Ph.D.

Rizwan Hamid M.D. Ph.D.

Jens Meiler Ph.D. (co-advisor)

John Anthony (Tony) Capra Ph.D. (co-advisor)

To Maa and Baba and Rick,  
for constantly supporting, encouraging and inspiring me towards higher education.

## ACKNOWLEDGEMENTS

I want to begin by humbly expressing my utmost gratitude to my mentors Dr. Tony Capra and Dr. Jens Meiler for accepting to mentor me together, even though I had very little coding experience when I first joined graduate school. Working with both of them has been a very rewarding experience, and I have learned a lot about proteins, personalized medicine and pilsners. This dissertation would not be possible without their encouragement to try new things, enthusiasm to discuss new results and the erudition to comprehend the limitations of the research. I interviewed with them before I came to Vanderbilt, and I'm very fortunate to have got the opportunity to work with them.

I express my deepest appreciation for my committee chair Dr. Antonis Rokas, and my thesis committee members Dr. Nicole Creanza, Dr. Rizwan Hamid and Dr. Carlos Lopez for guiding me on my PhD journey and helping with my research trajectory. They have always been available when I needed them to discuss research updates and big picture questions. I want to thank the entire UDN clinical and research team, especially Dr. Phillips, Dr. Cogan and Dr. Hamid for working with me patiently over the past five years while I figured out the best format to present the data. I have learnt a great deal from the weekly UDN meetings, and my motivation to keep working on developing tools to help diagnose rare diseases can be traced to those meetings. I want to thank all the patients that we have analyzed and reported on over the years and their families.

I have had a wonderful time in Vanderbilt during my PhD, and I would like to thank Beth and Carolyn for helping navigate graduate school, lab rotations, coursework and generally making this process more fun. I would like to thank all my friends in the Capra and Meiler labs for being helpful and supportive, especially Bian, David, Mary Lauren, Laura, Ling, Greg and

Abin for letting me bounce ideas off them and getting endless coffees together. I am grateful to Heather and Carie for making my PhD more convenient for always having an answer to my questions and setting up solutions to problems I didn't know I was about to run into. I thank the National Institutes of Health (NIH) for funding Dr. Capra, Dr. Meiler and the UDN, allowing me to conduct this research. This dissertation would not be possible without using the computational resources of the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, TN.

Graduate school has definitely been a learning experience, with its ups and downs, and sometimes I needed a little distraction to help me continue. I have to specially thank Pishi for all her help with getting used to living in Nashville. Over the pandemic, the Friday night libations served as a stress buster like no other and I am eternally thankful to Deepayan, Avisek, Purbodoy, Anabil, Rudra Da, Soumita Di and Gourab for ensuring I did not lose my mind, though I may have lost a few hepatocytes. I am extremely thankful for Elleansar (Ellie), Oanh and Jessica for just being phone call away whenever I craved for a pho with friends or just a causal conversation over hotpot.

It has taken a village and a half to get me to a PhD program in the USA, and there are too many people who have contributed to this directly. Dipanjan and Sayan (Ganguly) found me in school, and accepted me as a part of their tribe. I thank them for being my partners in theater and many misadventures over the years. I appreciate Soumil, Samrat Da, Zubin, Pip, Maitreyo, Arumoy Da, Kunla Da, Raja Da and SKG for helping and guiding me on my way towards a PhD. I would like to recognize the role of Farah Di, John, Colleen, Jim and Binnu in inspiring me to apply for a PhD and supporting me fervently over the last half decade. I have to dedicate this dissertation to Parijat, Anirudhya (Gadai), Sankalpa, Souparno and Budhaditya, without



whom I cannot imagine a career as a researcher and amidst whom I began my research journey, as well as co-wrote and performed allegorical, biological satirical plays. We have not been in the same city for almost ten years and I am glad to know that they are just a phone call away.

I want to incessantly thank Maa and Baba for inspiring me towards studying biology, thinking critically and giving me the courage and motivation to undertake the challenges of research and graduate school. Their words of guidance and encouragement have given me the energy to carry on when times are tough. I have been inspired by my grandfathers from a very young age, and they played a huge role in me choosing a PhD in Biology. I express my deepest gratitude for my brother Rick, whose timely interventions surrounding KKRicket and words of wisdom regarding career options have contributed a great deal. I am very grateful to Thanh for being there for me when I most needed it and being my best friend and constant partner when I was new to this country. Thanh motivated me to keep going with lab rotations, course work, qualification exam and committee meetings and taught me everything from League to cooking Cá Kho Tộ, and the secret to closing Ruffles bags. This dissertation would not exist without the people mentioned above. Thank you everyone for contributing to this journey.

## TABLE OF CONTENTS

<b>ACKNOWLEDGEMENTS</b>	<b>iii</b>
<b>SUMMARY</b>	<b>1</b>
<b>CHAPTER 1: Introduction</b>	<b>3</b>
Rare diseases	3
UDN	4
Digenic diseases	5
Machine learning	6
Random Forests	7
Protein structure	8
MD simulations	9
Roadmap	10
<b>CHAPTER 2: Designing DiGePred – a machine-learning digenic gene pair classifier to help diagnose rare disease</b>	<b>12</b>
Summary	12
Introduction	13
Results	13
Digenic disease gene pairs have different attributes than non-digenic disease gene pairs.....	13
Random forest classifiers accurately identify digenic pairs using network and functional features .....	19
Including additional features improves ability to identify digenic disease genes .....	22
Digenic disease genes can be distinguished from many non-digenic gene sets.....	23
Feature importance varies for classifiers trained on different non-digenic sets.....	23
DiGePred accurately identifies held-out digenic pairs.....	26
DiGePred identifies novel digenic pairs from the recent literature .....	28
Discussion	33
<b>CHAPTER 3: Application of DiGePred to real world scenarios – predict digenic pairs in individuals suffering from rare diseases</b>	<b>36</b>
Summary	36
Introduction	36

<b>Results</b>	<b>38</b>
DiGePred has a low false positive rate in real-world applications.....	38
Prediction of digenic pairs among patients with undiagnosed disease .....	49
Prediction of digenic pairs among all human gene pairs at various confidence thresholds .....	52
<b>Discussion</b>	<b>60</b>
 <b>CHAPTER 4: A Personalized Structural Biology (PSB) approach reveals the molecular mechanisms underlying heterogeneous epileptic phenotypes caused by de novo KCNC2 variants</b>	 <b>62</b>
<b>Summary</b>	<b>62</b>
<b>Introduction</b>	<b>63</b>
<b>Results</b>	<b>66</b>
Undiagnosed Diseases Network patient with DEE-like symptoms.....	66
.....	69
Structural modeling suggests distinct functional effects for candidate KCNC2 variants .....	69
.....	73
V469L causes loss of channel function, while V471L causes gain of function.....	73
V469L expression is lower while V471L expression is higher than wild type.....	75
V469L constricts the channel pore in molecular dynamics simulations increasing the energetic barrier for K <sup>+</sup> ion permeation.....	76
<b>Discussion</b>	<b>83</b>
 <b>CHAPTER 5: Conclusion</b>	 <b>87</b>
 <b>CHAPTER 6: Methods</b>	 <b>91</b>
<b>Digenic gene pairs</b>	<b>91</b>
<b>Non-digenic gene pairs</b>	<b>91</b>
<b>Six Network and Functional Features</b>	<b>92</b>
Pathway similarity .....	92
Phenotype similarity .....	92
Co-expression.....	93
PPI distance.....	93
Pathway distance .....	93
Literature distance .....	93
<b>Five Evolutionary Features</b>	<b>94</b>
Evolutionary age.....	94
Gene essentiality.....	94
Loss of function intolerance (pLI).....	94
Selection pressure (dN/dS).....	95

Haploinsufficiency.....	95
<b>Gene-focused network and functional features _____</b>	<b>95</b>
Number of pathways.....	95
Number of phenotypes.....	95
Network neighbors.....	96
Number co-expressed.....	96
<b>Encoding gene level features _____</b>	<b>96</b>
<b>Performance Quantification _____</b>	<b>96</b>
<b>Training and Testing the DiGePred Random Forest Models _____</b>	<b>97</b>
<b>Evaluation using additional digenic pairs not in DIDA _____</b>	<b>97</b>
<b>Feature Importance _____</b>	<b>99</b>
<b>Prediction Score Thresholds _____</b>	<b>99</b>
<b>Estimating the False Positive Rate at various score thresholds _____</b>	<b>99</b>
<b>Comparison with ORVAL _____</b>	<b>100</b>
<b>Gene ontology (GO) enrichment _____</b>	<b>101</b>
<b>Structural modeling of Kv3.2 _____</b>	<b>101</b>
<b>MD system setup _____</b>	<b>102</b>
<b>Conventional MD simulations _____</b>	<b>103</b>
<b>Umbrella MD simulations _____</b>	<b>104</b>
<b>Heterologous expression of Kv3.2 ion channel and whole-cell voltage clamp electrophysiology</b>	<b>105</b>
<b>DNA constructs for WT and variants of KCNC2 _____</b>	<b>106</b>
<b>Western Blot _____</b>	<b>107</b>
<b><i>DATA AND CODE AVAILABILITY _____</i></b>	<b><i>108</i></b>
<b><i>UNDIAGNOSED DISEASES NETWORK CONSORTIUM _____</i></b>	<b><i>110</i></b>
<b><i>REFERENCES _____</i></b>	<b><i>113</i></b>
<b><i>LIST OF FIGURES _____</i></b>	<b><i>145</i></b>
<b><i>LIST OF TABLES _____</i></b>	<b><i>149</i></b>

*LIST OF VIDEOS* \_\_\_\_\_ **150**

*LIST OF DATASETS* \_\_\_\_\_ **151**

## **SUMMARY**

I have worked on developing tools and techniques to interpret rare genetic variants, uncover rare disease mechanisms and help diagnose individuals suffering from rare undiagnosed diseases, in collaboration with the Undiagnosed Diseases Network (UDN). The UDN was established to help provide clinical intervention roadmaps for individuals with rare diseases. Rare genetic diseases affect more than 300 million people around the world; however, the causative genes and variants have not been identified for most.

Rare disease phenotypes can be caused as a consequence of variants in more than one gene, and digenic diseases result from variants in two genes. Experimentally evaluating digenic combinations cannot be done for all possible candidates in a timely manner to help diagnose individuals with undiagnosed diseases. A computational prediction and prioritization of digenic pairs can reduce the number of candidates for experimental validation by several orders of magnitude. I have developed a machine-learning classifier (DiGePred) to predict human gene pairs with the potential to cause digenic diseases, based on a database of known digenic diseases and features derived from biological networks, genomics and evolutionary biology. The classifier could accurately identify known digenic pairs in the held-out testing dataset, as well as recently discovered digenic pairs from recent literature, not used for training. I also demonstrated the low false positive rate of DiGePred on unaffected relatives of individuals with rare diseases, being studied by the UDN.

Our group has collaborated with the UDN to help interpret rare variants using computational structural biology and molecular dynamics (MD) simulations, and predict digenic disease causing candidate gene pairs. The digenic classifier predicts gene pairs with the potential

to cause digenic disease when carrying rare deleterious variants simultaneously. However, the variants do not factor into the prediction. There are several tools available to interpret pathogenicity of genetic variants from a genomic or amino acid sequence paradigm.

However, protein functions are mediated in the 3D conformational space, where amino acid residues distant in sequence could be proximal in 3D space, after the protein folds to adopt a functionally active conformation. There are increasingly more protein structural models available for genes, especially with the advent of AlphaFold. Rare variants in the coding region of genes can result in changes to the 3D protein structure. There are currently several variant interpretation tools available that consider the 3D protein structural context, however, often the context is not used to explore disease mechanisms.

I have contributed to the establishment of a “Personalized Structural Biology” approach, based on computational structural biology and Molecular Dynamics (MD) simulations, in collaboration with biochemists and electrophysiologists, to interpret rare disease variants. I led a comprehensive analysis and illustration of this approach on the effects of rare *de novo* missense variants in *KCNC2*, a gene coding for the Kv3.2 potassium ion channel. I was able to postulate a mechanism using structural biology insights, that were validated using biochemical and electrophysiology experiments, and the rationale was provided using Molecular Dynamic (MD) simulations.

During my PhD, I have worked extensively in collaboration with the UDN to develop computational tools and techniques to help resolve the mechanisms underlying rare diseases, and interpret rare variants using computational structural biology and machine learning.

## **CHAPTER 1: Introduction**

Since the completion of the human genome project<sup>1,2</sup>, the genetic origin of many severe diseases has been determined. Genetic variants in almost 3000 human proteins have been identified as causing altered physiology, observable as disease phenotypes<sup>3,4</sup>. Many of these diseases identified are monogenic diseases which usually follow Mendelian patterns of inheritance, with variants in only a single gene being identified as causative<sup>5-7</sup>. Often a single rare deleterious variant forms the basis of these Mendelian phenotypes, and diagnosis of these diseases has been focused on analyzing the effect of rare variants<sup>8-12</sup>. At the other end of the spectrum, there are polygenic diseases that result from the combined impact of common variants in multiple genes and non-coding loci, as well. The genetic loci linked to polygenic diseases such as hypertension, coronary heart disease and diabetes have been identified using Genome wide Association studies (GWAS)<sup>13</sup>. Polygenic risk scores<sup>14</sup> are currently used to assess the genetic risk of polygenic diseases for individuals during their lifetime. Polygenic diseases usually affect thousands of people, while Mendelian diseases affect very few people in comparison, and are often characterized as “rare” diseases.

### *Rare diseases*

A rare disease is defined as an affliction that affects fewer than 200,000 individuals in the USA, according to the Orphan Drug Act of 1983<sup>15,16</sup>. In the European Union, a condition affecting fewer than 1 in 2000 individuals is termed as a rare disease. Over 7,000 rare diseases have been identified so far, with an estimated 300 million people suffering from rare diseases worldwide. Roughly one in 10 people in the world are afflicted by rare disease. Approximately 80% of all rare diseases are genetic in origin, and more than half of the known rare diseases do not have



causative genes and variants discovered.<sup>4,5,7</sup> Several rare diseases are tracked via new born screening and routine medical screening in adults; however, a precise estimate of number of individuals suffering from rare diseases is difficult to obtain. As a result, several research cohorts have been established to help prognose and design treatment plans for individuals with rare diseases. The Undiagnosed Diseases Network (UDN)<sup>17</sup> was established by the NIH in 2014 to help address the challenge of rare diseases.

### *UDN*

The Undiagnosed Diseases Network (UDN)<sup>17-19</sup>, funded by the National Institutes of Health Common Fund, comprises a team of researchers and clinicians across the country to address medical mysteries using an interdisciplinary approach including genomics, bioinformatics and other computational techniques. The purpose is to ameliorate the health of individual patients and families afflicted by rare and undiagnosed diseases, and lead to mechanistic understanding of rare diseases. There are 12 sites across the country that serve as clinical sites, where medical practitioners and healthcare providers, geneticists and bioinformaticians work together to help resolve the rare disease phenotypes. There is a sequencing core, model organisms screening center, and metabolomics core.

The UDN has received more than 5000 applications from individuals suffering from rare undiagnosed disease, and roughly 2000 applications were analyzed after preliminary reviews. The UDN team has successfully diagnosed over 500 individuals, using whole genome or exome sequencing, model organism screening, metabolomic analyses. The UDN has contributed to rare disease research and mechanistic knowledge with over 150 manuscripts and over 500 additions

of rare variants to ClinVar. Although this approach has yielded much success,<sup>20–32</sup> more than half of all UDN cases remain undiagnosed.

Our team, at Vanderbilt University, provides personalized computational structural biology analysis for candidate rare variants identified in an individual suffering from rare disease and being analyzed by the UDN. I hypothesized that in many of these unsolved, rare cases might involve variants in multiple genes that only when combined result in a disease phenotype complicating diagnosis. I developed a machine learning classifier to predict gene pairs that could lead to digenic diseases, rare diseases arising from variants in two genes, and I communicate digenic dual molecular hypotheses for the clinical phenotypes on the basis of the variants in the UDN patient, as well.

### *Digenic diseases*

Variants in more than one gene, usually ranging between two and four genes, can synergistically lead to disease via different mechanisms, such as direct molecular interactions or multiple genes in the same pathway<sup>33–36</sup>. Digenic inheritance was first demonstrated in 1994, when concurrent mutations in two genes, Retinal outer membrane protein 1 (ROM1) and Peripherin 2 (PRPH), were found to cause retinitis pigmentosa.<sup>37</sup> Digenic inheritance is the simplest form of oligogenic inheritance in which the combination of a small number of variants leads to disease.<sup>38–40</sup> Digenic diseases have been previously classified into two categories: true digenic, where variants in both genes are essential for development of clinical phenotypes, and composite, where one variant is responsible for causing a clinical phenotype, and the second variant severely exacerbates the disease.<sup>41,42</sup> However, in all cases of digenic inheritance the phenotype results from the combined effect of two variants. In isolation, the individual variants that form a digenic pair are benign or

lead to a less extreme phenotype. However, upon simultaneous mutation, the variants either interact to produce disease or combine to produce a more complex, and usually more severe, phenotype that cannot be explained by variants in one gene alone.

Since the discovery of a digenic cause for retinitis pigmentosa (RP) in 1994, many additional digenic diseases have been identified. The Digenic Diseases Database (DIDA)<sup>39</sup> has chronicled several hundred cases of digenic disease in 2017. Analyses of DIDA have revealed that digenic disease causing gene pairs are more likely to functionally and/or physically interact with one another than expected by chance<sup>39</sup>. Machine learning approaches have been developed to distinguish between different types of digenic disease pairs<sup>42</sup> and to identify disease causing variant combinations,<sup>43,44</sup> including oligogenic combinations of greater than two genes<sup>45</sup>.

### *Machine learning*

Machine learning methods are versatile approaches to derive and visualize functional interactions in large-scale data, without the need explicitly define them beforehand.<sup>46-48</sup> The merit of application of machine learning in computational biology is the scope to develop predictive models without complete comprehension of underlying physiological mechanisms.<sup>49</sup> Incorporation of disparate biological data from different sources such as genomics, proteomics and electronic medical health records data can lead to better performing machine learning models, with an improved ability to grasp and explain complex biological mechanisms.<sup>50,51</sup> However, biological data can be very sparse and not well defined, with the number of samples often fewer than the number of variables available. This disparity can be explained by the cost associated with generating data from individuals, for example genome sequencing or cancer

studies. The disparity is sometimes referred to as the “curse of dimensionality”, and has the potential to lead to inaccurate models owing to overfitting or missing data.<sup>52</sup>

Machine learning methods have been previously used to study rare disease mechanisms<sup>53</sup>, help provide prognosis for individuals with rare diseases<sup>54</sup> and develop better clinical and pharmaceutical intervention strategies<sup>55</sup>. Investigations into rare disease biology using machine learning methods have been undertaken in many countries<sup>56</sup>. A common impediment to high-throughput machine learning methods is unstructured data in the form of freeform text records or non-standardized medical health records<sup>57</sup>. The incorporation of medical data in canonical vocabulary such as the Orphanet rare disease nomenclature<sup>58</sup> or the Human Phenotype Ontology (HPO) terms<sup>59,60</sup> has been demonstrated to improve machine learning based analysis of rare diseases. Machine learning has been previously used in conjunction with the UDN, to predict the realistic estimates to the probability of patients to be accepted into the UDN cohort for analysis, based on the clinical phenotypes that manifested in the individual.<sup>61</sup> Ensemble methods such as Random Forests, support vector machines (SVMs) and artificial neural networks have been most frequently used to analyze rare diseases.

### *Random Forests*

An ensemble classification model relies on accumulating predictions from many different classifiers to get a final prediction value or label.<sup>62</sup> Combining outputs from different classifiers has the advantage of improving the robustness of the ensemble, leading to more accurate predictions for a wider range of data. A Random Forest (RF) model averages along an arbitrary distribution of decision trees to derive the final predictive value. Each tree provides one class or label continuum as an output, and each individual sample is classified as belonging to the class

based on the consensus from all the trees part of the ensemble.<sup>63</sup> RFs are less susceptible to overfitting and shown to perform better in the cases of high missing data and high dimensional data.<sup>64</sup> RFs have been previously used in bioinformatics, proteomics, and genetics paradigms.<sup>65–67</sup> A strength of RFs is the ability to conveniently rank and prioritize features for the prediction task by calculating feature importance using a non-parametric Gini impurity reduction metric.<sup>65,68,69</sup> Machine learning methods have been used to predict the impact of rare variants on human physiology, as well on protein structure and function.<sup>55</sup>

### *Protein structure*

The vast diversity of physiological functions of human proteins is mediated by the precise three-dimensional structural conformation.<sup>70,71</sup> According to Anfinsen's thermodynamic hypothesis,<sup>72–74</sup> the amino acid sequence of the protein contains the information needed for the protein to adopt its native 3D conformation, and the information is encoded in the energetic landscape of the protein, with native conformation considered to have the lowest energetic profile. The Levinthal's paradox<sup>75,76</sup> postulates that the entire conformational landscape does not need to be sampled to attain the native conformation. The energetic landscape is often considered to be shaped as a “funnel” with the energetic state of current conformations driving the protein to be folded towards the native conformation.<sup>70,77,78</sup>

The methods of resolving native protein structure experimentally such as NMR spectroscopy, X-ray crystallography and Cryo electron microscopy (cryo-EM) have become more accurate and robust more recently<sup>79–84</sup>. However, NMR spectroscopy for large proteins remains challenging<sup>85</sup>, X-ray crystallography requires large amounts of proteins to optimize crystallization and structure determination<sup>86</sup>, while cryo-EM samples are difficult to prepare and

resolving the structure of smaller proteins and flexible regions can be cumbersome<sup>84</sup>. The prediction of protein structure from genomic sequence has become more accurate and advanced over the past decade.<sup>70</sup> Machine learning methods have also been employed to help determine the native conformation of proteins, as well as predict the effect of variants on protein structure and folding thermodynamics. Molecular modeling software such as Rosetta<sup>87,88</sup> and FoldX<sup>89,90</sup> have been used to predict the structural conformation of amino acid sequences, and estimate the energy changes of the system associated with mutations in the sequence. The advent of AlphaFold<sup>91,92</sup> has led to structural models becoming available for a >90% of all human proteins on UniProt allowing modeling of thermodynamic changes involving genetic variants. This has further enabled the modeling the effect of genetic variants on altering protein structure and function to become a major component of modern genomic medicine.

A notable limitation of conventional modeling of protein structure is that the analysis usually only samples the native conformation or a functional stable conformation of the protein. However, protein structures are highly dynamic and often different conformations have different functional roles in physiology. It is possible that a mutation could thermodynamically favor a particular conformation over another or render the protein unable to adopt a certain conformation completely, leading to impaired function. There are methods available to simulate the dynamics of protein structures and study protein folding on a longer time scale.

### *MD simulations*

Molecular dynamics (MD) simulations for proteins predict the movement of every atom in the system over a certain period of time, based on physical forces determining atomic interactions.<sup>93,94</sup> The first MD simulation of a protein was performed in 1977<sup>95</sup>, and it has

become far more common in recent years with exponentially more protein structural models available<sup>70</sup> and computational analysis becoming more affordable and accessible with graphics processing units (GPUs) now permitting sophisticated simulations in a cost effective manner.<sup>96–</sup>  
<sup>101</sup> The MD simulation software packages<sup>96,102</sup> have become easier to setup and interpret and the physical approximations underlying modeling atomic interactions have become more accurate. MD simulations yield a trajectory file which is essentially a three-dimensional movie chronicling the atomic-level configuration of the protein system at every time point over the simulated time interval. MD simulations are usually coupled with experimental methods to further elucidate protein folding and thermodynamic changes arising from genetic variants mechanisms more clearly.<sup>93,94,103</sup>

Although MD simulations have become comparatively more convenient to perform over the past few years, there is still the need for considerable iteration to get concordant data. The results can still be difficult to interpret in the absence of experimental data, and may require downstream experiments to validate findings.<sup>94</sup> Identifying residues and interactions vital for protein folding and thermodynamic stability from MD simulations can be inaccurate in some cases. Functional methods based on mutagenesis of protein sequences can elucidate the impact of point mutations on organism physiology. I have used MD simulations to study the impact of rare *de novo* missense variants in a potassium ion channel on protein dynamics and ion transport through channel pore.

### *Roadmap*

In Chapter 2, I have discussed the development of DiGePred, the digenic disease Random Forest machine-learning classifier. I have explained the features used for training, the performance of

the classifier during training, and testing the performance and establishing the prediction threshold using a held-out testing data set and novel digenic pairs from recent literature. In Chapter 3, I have chronicled the application of DiGePred on rare disease cohorts in the UDN, and external cohorts as well. There was a novel digenic disease pair discovered in a cohort of individuals suffering from MRKH, a connective tissue disease affecting the female reproductive system. In Chapter 4, I have written about the “personalized structural biology” approach our team has conceptualized. I have postulated a computational structural biology-based hypothesis regarding the mechanism by which a *de novo* missense variant in a potassium ion channel; can lead to developmental epileptic encephalopathy (DEE) like phenotypes in a child. I collaborated with experimental biochemists and electrophysiologists to validate our hypothesis, and I used MD simulations to provide rationale for the experimental observations, and uncover the mechanism of action.



## **CHAPTER 2: Designing DiGePred – a machine-learning digenic gene pair classifier to help diagnose rare disease**

### *Summary*

The central hypothesis of this work is that many of the rare genetic disorders that remain unresolved after analysis by the UDN can be caused by multiple variants in more than one gene. I developed DiGePred, a random forest classifier for identifying candidate digenic disease gene pairs using features derived from biological networks, genomics, evolutionary history, and functional annotations. I trained the DiGePred classifier using DIDA, the largest available database of known digenic disease causing gene pairs, and several sets of non-digenic gene pairs, including variant pairs derived from unaffected relatives of UDN patients. DiGePred achieved high precision and recall in cross-validation and on a held-out test set (PR area under the curve >77%), and I further demonstrated its utility using digenic pairs from the recent literature. This work enables the discovery of genetic causes for rare non-monogenic diseases by providing a means to rapidly evaluate variant gene pairs for the potential to cause digenic disease.

This work has been published in the American Journal of Human Genetics (AJHG) in October 2021. (**Mukherjee S**, Cogan JD, Newman JH, Phillips JA, Hamid R, Undiagnosed Diseases Network, Meiler J, Capra JA. Identifying digenic disease genes using machine learning in the undiagnosed diseases network. *Am J Hum Genet*, 2021 Oct 7;108(10):1946-1963. doi: 10.1016/j.ajhg.2021.08.010.)



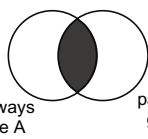

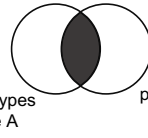

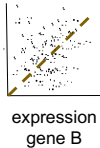

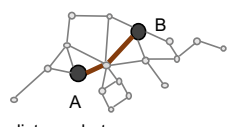
## *Introduction*

I hypothesized that the disease phenotype in some unresolved rare disease patients is likely a result of digenic inheritance and develop DiGePred, a high-throughput machine learning tool for evaluating the likelihood that dysfunction of gene pairs leads to digenic disease. I focused on the specific challenge of identifying gene pairs that have functional or phenotypic potential to cause a digenic disease when both are disrupted in a patient. I considered all cases in DIDA, which includes cases where both variants are required for disease and cases in which having the variants simultaneously modifies disease presentation or severity. My approach is based on supervised machine learning using a random forest classifier trained on diverse functional, network, and evolutionary properties of known digenic gene pairs versus realistic sets of non-digenic gene pairs, including variant pairs from healthy individuals. This work has already been published in the American Journal of Human Genetics (AJHG) in September 2021.

## *Results*

### *Digenic disease gene pairs have different attributes than non-digenic disease gene pairs*

My goal in this study is to develop a machine learning classifier for identifying gene pairs that cause disease when both are disrupted simultaneously, but produce no or less severe phenotypes when disrupted in isolation. To this end, I considered all unique known digenic disease pairs curated by the DIDA database and contrast them with several informative sets of non-digenic disease pairs. Pairs of genes harboring mutations known to cause digenic disease have distinct biological properties when compared with random gene pairs<sup>39</sup>. Previous work has shown that digenic disease pairs have high protein interaction network connectivity and proximity. More than 35% of known digenic disease pairs directly interact on a protein-protein interaction (PPI)

Feature	Source	Logic
pathway similarity	 	
phenotype similarity		
co-expression rank		
PPI distance pathway distance literature distance	 UCSC GENE and PATHWAYS INTERACTIONS	

**FIGURE 1: Network and Functional Features (NFFs) used for machine-learning-based identification of digenic disease gene pairs**

I considered six network and functional features (NFFs) for training the digenic disease classifiers: i) *pathway similarity*: Jaccard similarity of pathway annotations from KEGG and Reactome for both genes; ii) *phenotype similarity*: Jaccard similarity of phenotype annotations from HPO for both genes, iii) *co-expression rank*: co-expression rank of gene pair compared to all other gene pairs across multiple tissues from COXPRESdb; iv-vi) *network distances* between the genes on protein-protein, pathway, and literature mined interaction networks from UCSC gene and pathway interaction browser database.

network, and ~60% of digenic gene pairs are one gene away on the interaction network. Similarly, ~20% of digenic pairs are in the same biochemical pathway, and ~40% are expressed in the same tissues<sup>39</sup>.


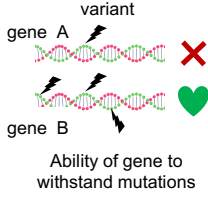
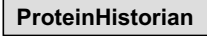
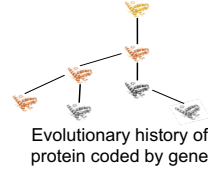

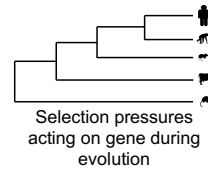

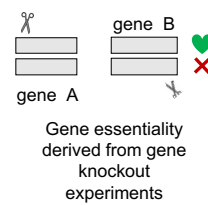
Based on this prior knowledge I devised a list of

six “network and functional features” (NFFs) to use as attributes for distinguishing between digenic and non-digenic gene pairs (**Figure 1**):

1) *Pathway Similarity*, defined as the Jaccard similarity<sup>116</sup>

between the genes’

membership in ~1800 pathways from KEGG<sup>117</sup> and Reactome<sup>118,119</sup>; 2) *Phenotype Similarity*, the Jaccard similarity between the ~6000 phenotypes from Human Phenotype Ontology (HPO)<sup>120</sup> associated with the genes; 3) *Co-expression Rank*, defined as the rank of the co-expression of the genes across 23 co-expression platforms from 11 species compared to other gene pairs from

Evolutionary features		
Feature	Source	Logic
loss of function intolerance haploinsufficiency		
protein age		
selection pressure		
gene essentiality		

**FIGURE 2:** Additional feature sets used for machine learning classification of digenic diseases

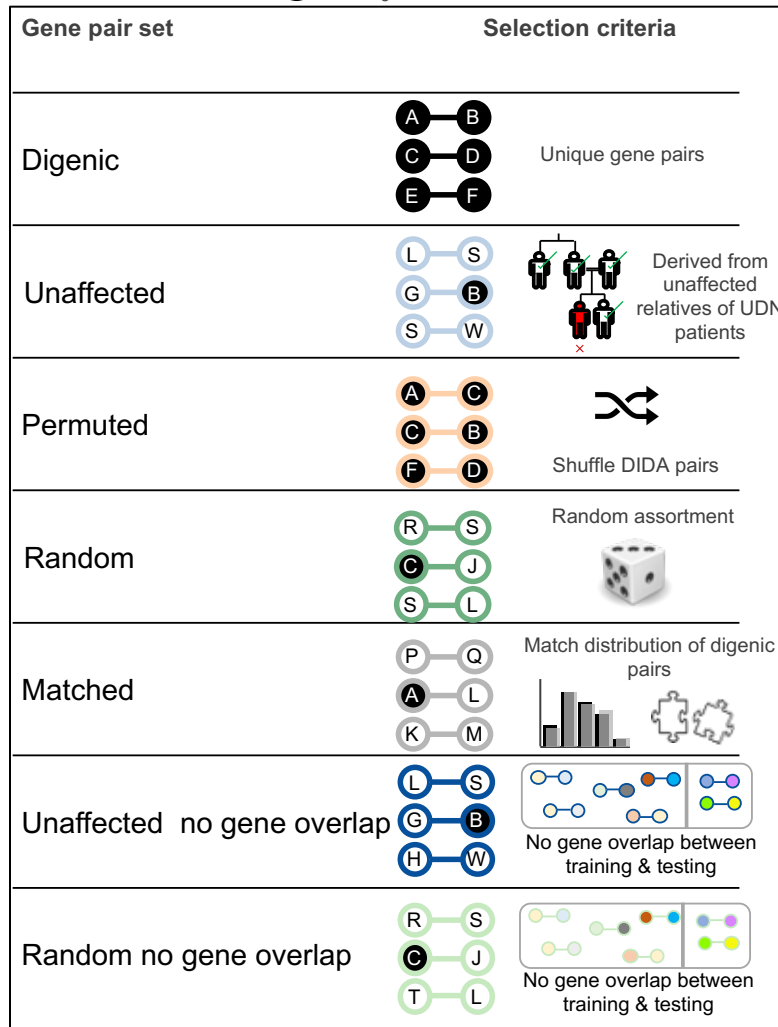
Evolutionary biology and genomics features: i) loss of function intolerance; ii) haploinsufficiency, measures of mutational load on gene; iii) protein age, measure of evolutionary age of protein coded by gene; iv) dN/dS score, measures the constraints on selection during mammalian evolution of gene; v) essentiality score, derived from gene KO experiments, measures how vital a gene is to organism survival.

COXPRESdb<sup>121</sup>; 4) *PPI Distance*, the distance on a global PPI network; 5) *Pathway Distance*, the distance on an annotated biochemical pathway network; and 6) *Literature Distance*, the distance on a literature-mined interaction network, derived from the UCSC gene and pathway interaction database<sup>122</sup>.

Since the ultimate application is the detection of potential digenic diseases in patients, most of the results focus on comparisons of known digenic gene pairs and gene pairs with variants in “unaffected” parents, siblings,

and other relatives of 25 UDN patients (**Figure 3**). However, as I have shown below, the results are similar using other strategies for defining non-digenic disease gene pairs.

(X) Not digenic disease gene (Y) Digenic disease gene



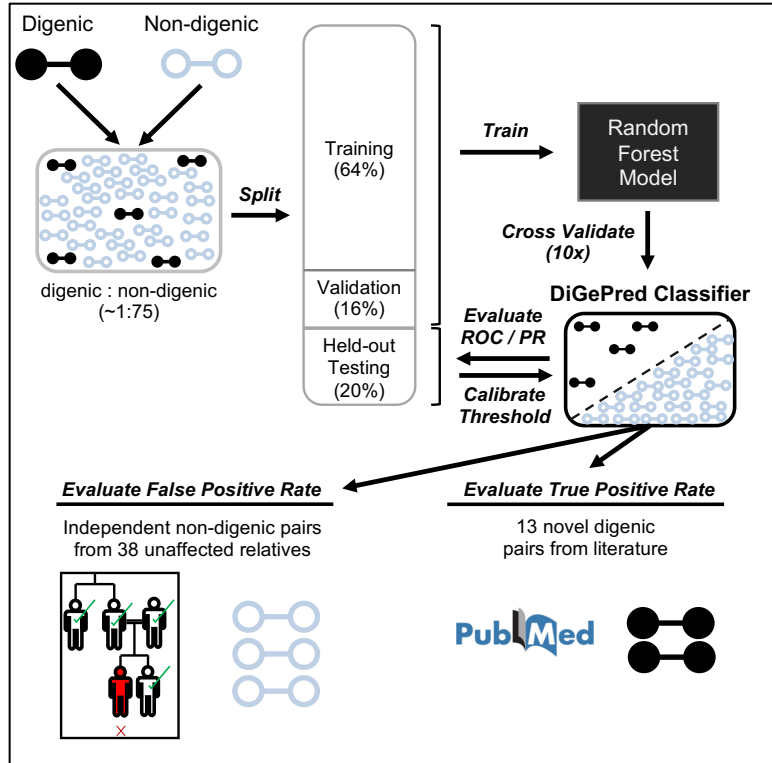
**FIGURE 3:** Positive and negative training sets used for classification of digenic disease genes

Digenic gene pairs were derived from DIDA. Unique digenic gene pair combinations (n=140) were used for training and evaluation. *Unaffected* gene pairs were derived from genes with variants in unaffected members of UDN patient families. *Permuted* negative gene pairs were generated by computing all possible permutations of genes in digenic pairs, excluding the known digenic combinations. *Random* gene pairs were generated by selecting random pairs of all human genes, excluding any known to be digenic. *Matched* gene pairs were selected from random gene pairs so that the set matched the NFF distribution of digenic pairs (**Figure 4**). *Unaffected no gene overlap* and *Random no gene overlap* pairs were selected subsets from the unaffected and random pairs respectively, such that there was no overlap between the training and held-out testing pairs.

I compared the distribution of the NFFs for known digenic pairs and for non-digenic gene pairs from unaffected relatives of UDN patients. As expected from previous work, the distribution of each NFF was significantly different between digenic and non-digenic pairs (**Figure 5**;  $P < 10^{-20}$  for each, Mann-Whitney U (MWU) test). This suggests that a machine learning approach may enable distinguishing digenic from non-digenic disease pairs.

To further explore the properties of digenic disease genes and the ability of a classification approach to recognize them, I defined three additional sets of

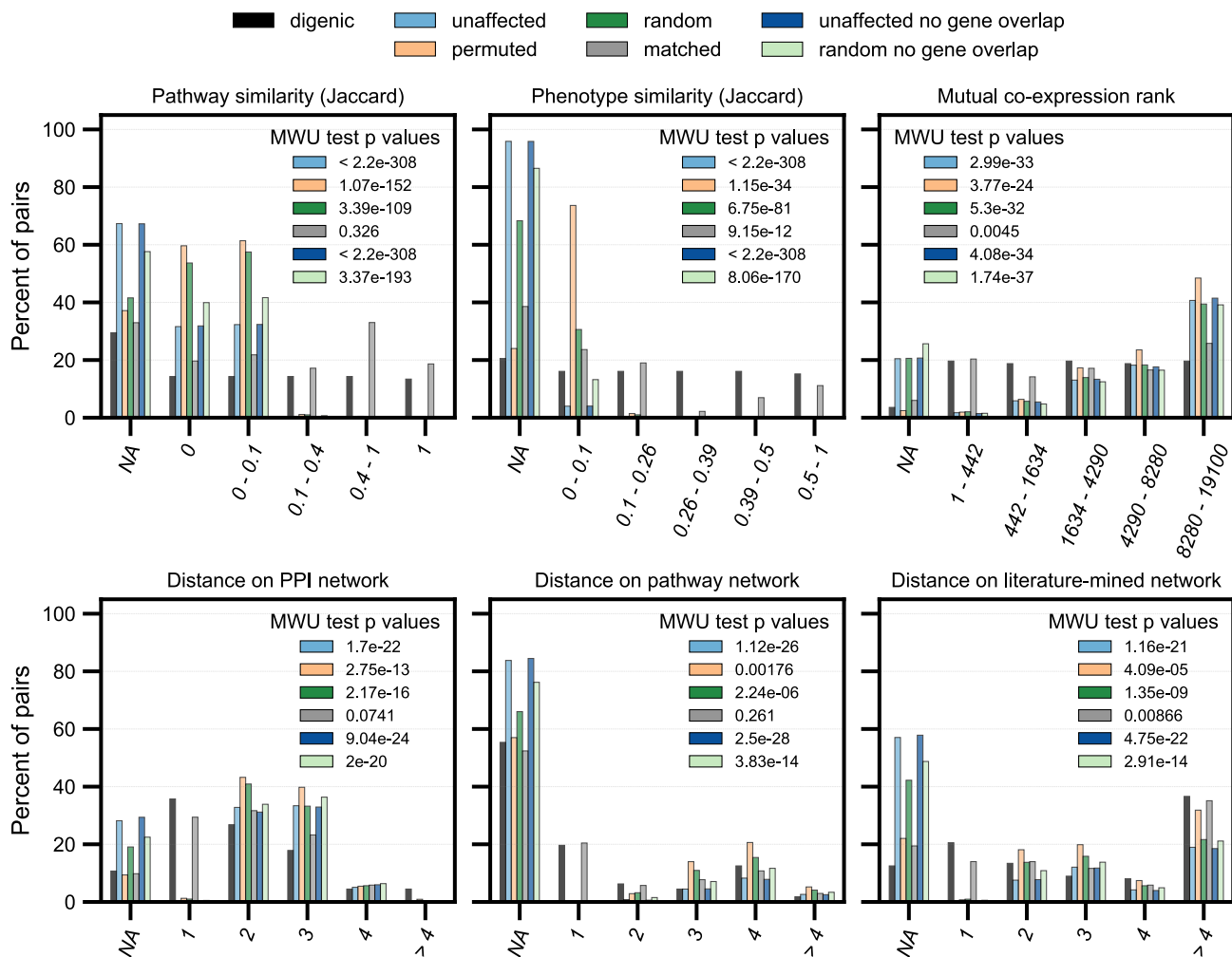
non-digenic disease gene pairs (Figure 3). First, I created a “permuted” non-digenic set by generating all possible gene pairs from genes known to be involved in a digenic gene pair, and removing the pairs known to be digenic. Second, I constructed a “random” set of non-digenic gene pairs by randomly selecting gene pairs from all possible human genes (excluding known digenic pairs). Third, I created a “matched” non-digenic gene pair set that closely matched the NFF distributions of the digenic gene pairs; however, I was not able to match the distribution of all NFFs perfectly given the



**FIGURE 4:** Schematic of the protocol for training and evaluating the DiGePred digenic disease pair classifier

Known digenic pairs (positives) and variant gene pairs from healthy individuals (negatives) were combined at ~1:75 ratio. The combined pairs were divided into training (64%), validation (16%) and held-out test datasets (20%). The DiGePred random forest classifier was trained and cross-validated using the training and validation sets. The final performance estimate for the trained DiGePred classifier was quantified by the area under the Receiver Operator Characteristic (ROC) and Precision-Recall (PR) curves (AUCs) on the held-out test set. This set was also used to establish suggested thresholds on the continuous DiGePred score. DiGePred’s potential clinical utility was further demonstrated by applying it to an additional positive set of 13 novel digenic pairs from the recent literature, one novel gene pair in a resolved UDN patient and an external set of non-digenic gene pairs from 38 unaffected relatives of UDN patients.

skewed distribution of the digenic disease pairs (Figure 5). Nonetheless, the matched set enables exploration of how well the classification approach can identify digenic pairs among non-digenic



**FIGURE 5:** Distribution of network features is different for digenic and non-digenic pairs; similar for matched gene pairs

Feature distributions of network and functional features (NFFs) for *digenic* gene pairs (black), *unaffected* (light blue), *permuted* (orange), *random* (dark green), *matched* (grey), *unaffected no gene overlap* (dark blue) and *random no gene overlap* (light green) gene pairs. The feature value bins shown along X axis and proportion of gene pairs along Y axis. Distributions compared using MWU test, P values shown. **A)** Jaccard similarity of KEGG and Reactome pathways associated with both genes; **B)** Jaccard similarity of HPO phenotypes associated with both genes; **C)** Mutual co-expression rank, comparison of co-expression of gene pair to all other gene pairs across multiple co-expression platforms; **D-F)** Distance on experimental PPI, biochemical pathways and literature-mined interaction networks, obtained from UCSC gene and pathway interaction browser database. MWU P consistently higher for matched pairs for all features. unaffected relatives of UDN patients.

pairs with similar NFF distributions. To be conservative, I also constructed non-digenic gene pair sets with no overlap between the individual genes present in the training and the held-out test datasets<sup>123</sup>. These are subsets of the unaffected and random sets and will be referred to as “unaffected no gene overlap” and “random no gene overlap,” respectively (**Figure 3**).

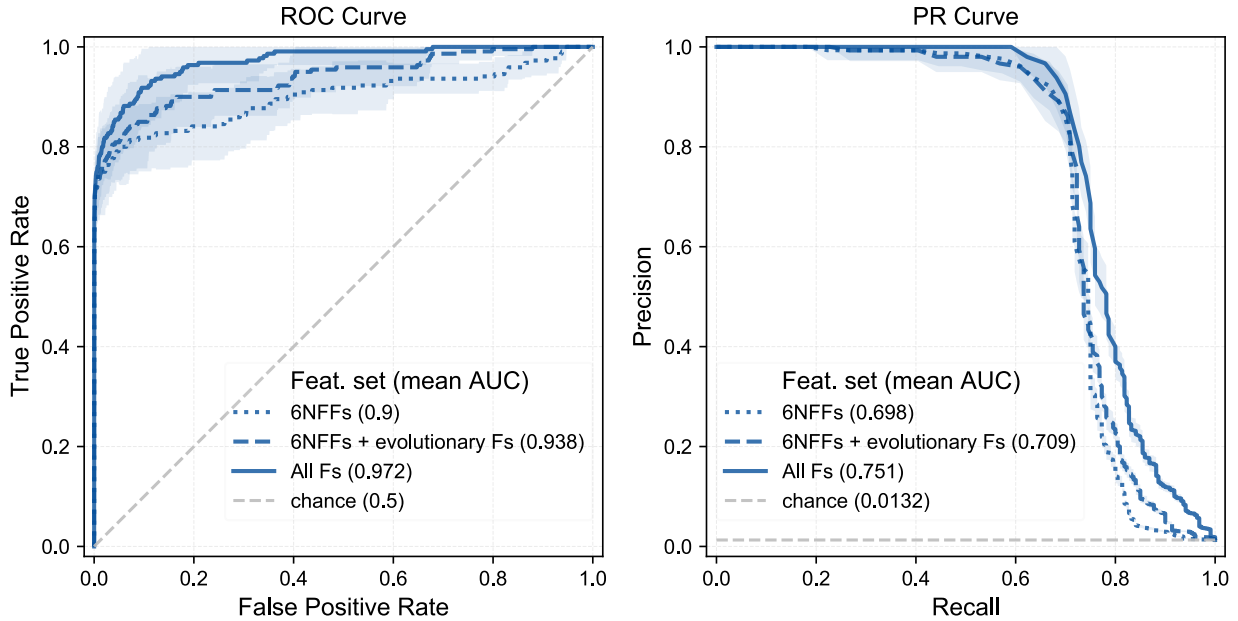
*Random forest classifiers accurately identify digenic pairs using network and functional features*

I divided the available gene pairs into training (64%), validation (16%), and held-out test sets (20%). I trained, evaluated, and compared different models using 10-fold cross-validation within the training and validation sets (**Figure 4**). The test set was only analyzed after models had been finalized. Comprehensive studies of genetic interactions have found that one in approximately 40 gene pairs interact.<sup>124</sup> This suggests that digenic interactions are likely rare; only a very small fraction of all possible gene pairs are likely to produce digenic disease. I trained the random forest machine learning classifier using the six NFFs to distinguish 140 digenic disease gene pairs (positives) from ~8,400 negative gene pairs. Unless otherwise specified, I focused in the main text on the “unaffected no gene overlap” negative set and present others in Supplementary Material. The large class imbalance (~1:75) reflects the expectation few digenic gene pairs among all possible pairs to be evaluated; this exact ratio was selected due to data availability. I evaluated performance using Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves.

The random forest classifier distinguished digenic and non-digenic gene pairs very accurately using the six NFFs. It achieved an average ROC area under the curve (AUC) of 0.90 and a PR AUC of 0.698 on average over 10 folds of cross-validation on the training and validation sets (**Figure 6**). The algorithm retains near perfect precision at recall above 60%

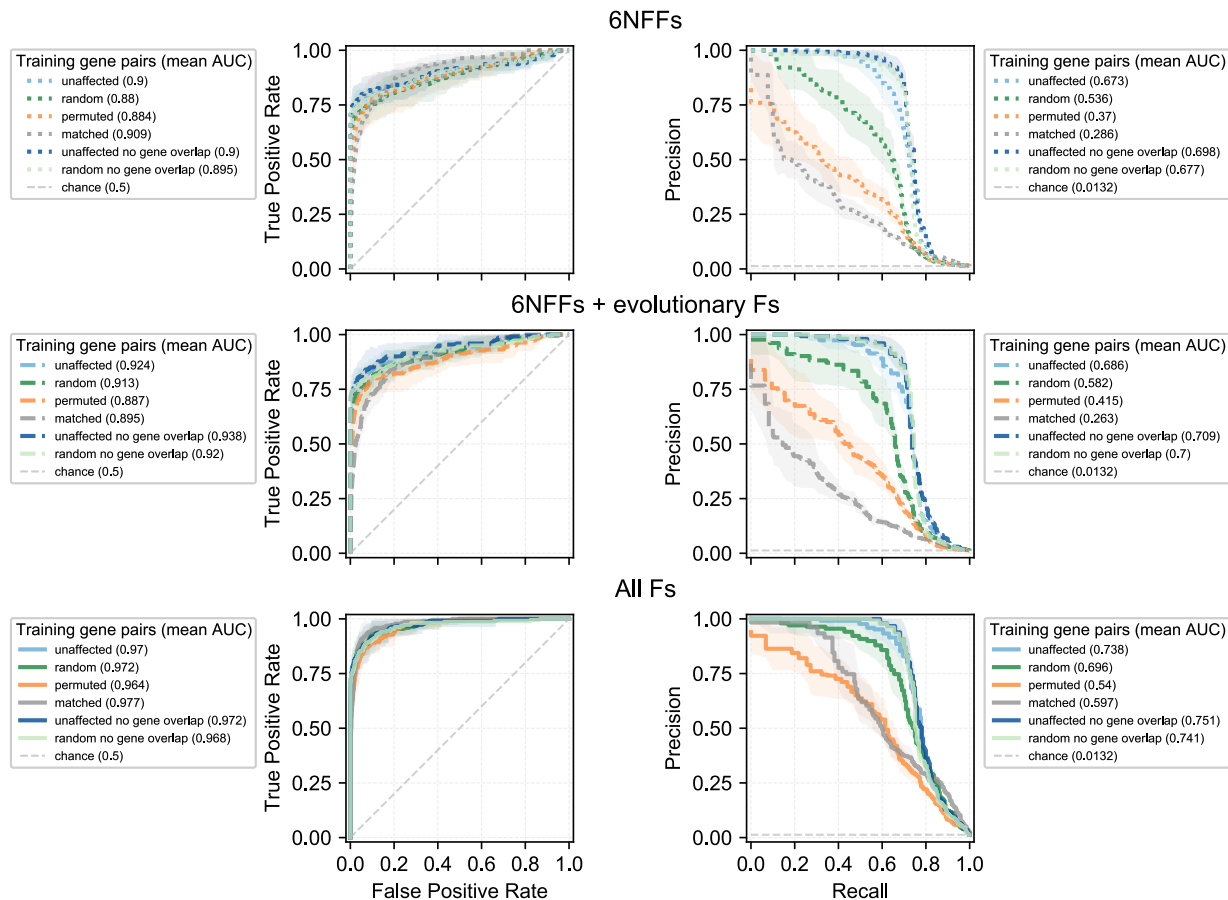


(Figure 6B). Since I was evaluating multiple classification approaches, the held-out test set was not considered in this analysis.



**FIGURE 6:** Classifier accurately identified digenic pairs from all non-digenic gene pairs using various feature sets; addition of features improved performance.

Performance of classifiers at distinguishing between known digenic pairs from DIDA (positives) and gene pairs from 25 healthy individuals (negatives) trained using different feature sets as evaluated by: (A) Receiver Operating Characteristic (ROC) curves and (B) Precision-Recall (PR) curves. Classifiers trained on three sets of features are compared: i) six network and functional features (NFFs) (dotted line); ii) the six NFFs and evolutionary genomics features; and iii) the six NFFs, evolutionary genomics features, and gene-level network and functional features. The mean curves across 10-fold cross-validation on the training and validation sets are plotted with shaded areas representing the standard deviation. Since this analysis is developing and evaluating multiple possible classifiers, I held-out the test set for final evaluation (Figure 10).  
 Positives: digenic pairs from DIDA; training  
 Negatives: “unaffected” non-digenic pairs; training



**FIGURE 7:** Classifier accurately identified digenic pairs from all non-digenic gene pairs using various feature sets; addition of features improved performance.

Performance of the classifier on distinguishing between digenic pairs and non-digenic sets of gene pairs: **Unaffected** (light blue), **Random** (dark green), **Permuted** (orange), **Matched** (grey), **Unaffected no gene overlap** (dark blue) and **Random no gene overlap** (light green) training data, measured by area under the Receiver Operating Characteristic (ROC) (**A**, **C**, **E**) and Precision-Recall (PR) curves (**B**, **D**, **F**) (AUCs), using different feature sets: **A-B**) six network features (NFFs) (dotted line); **C-D**) NFFs + Evolutionary biology and genomics features (EBGFs) (dashed line); **E-F**) NFFs + EBGFs + NFF related features (solid line). The ROC and PR AUCs for each non-digenic set increased with added features, and the unaffected no gene overlap was the best performing set.

Positives: digenic pairs from DIDA; training

Negatives: non-digenic pairs from various sources; training

*Including additional features improves ability to identify digenic disease genes*

The performance of the classifier based on the six NFFs alone was strong; however, there are many other sources of biological information beyond the NFFs that could potentially inform either the nature of the relationship between genes or the relative likelihood and risk of a gene being mutated and causing disease. I tested if including additional features in training the classifier would increase performance of the classifier.

First, I trained classifiers using the six NFFs and five additional evolutionary features that reflect the evolutionary history and constraint on the genes (**Figure 6- 7**). These features were: 1) the evolutionary ages of the genes; 2) their essentiality; 3) their intolerance to loss of function mutations, 4) the selection pressure acting on them through mammalian evolution (dN/dS) and 5) their haploinsufficiency scores. The addition of evolutionary features, as the quadratic mean of the values associated with both genes, substantially improved classifier performance: average ROC AUC of 0.938 and PR AUC of 0.709 (**Figure 6-7**).

Next, I considered additional features derived from network and functional annotations of the gene pairs. These features were designed to add additional gene-focused (rather than gene-pair-focused) context and explore the sufficiency of the six NFFs. These features were: 1) the number of pathways, 2) phenotypes, 3) network neighbors, and 4) genes co-expressed for each individual gene in a candidate pair. As above, I used the quadratic mean to combine these gene-level features. Considering these features also further improved classifier performance, with an average ROC AUC of 0.972 and PR AUC of 0.751 for all features (**Figure 6-7**).

### *Digenic disease genes can be distinguished from many non-digenic gene sets*

I used the same training and evaluation approach as described in the previous section for the unaffected no gene overlap negative set to train random forest classifiers to distinguish digenic disease gene pairs from each of the additional negative sets (random, permuted, and matched) using all the network, functional, and evolutionary features. In each case, the classifiers performed very well (**Figure 7**). The classifiers trained to distinguish digenic pairs from random and random no gene overlap pairs performed the best (mean ROC AUC of 0.972, 0.968 and PR AUC of 0.696, 0.741), with all features included for training. As expected, given their similar attributes to the digenic pairs, the permuted and matched negative sets are more challenging for the classifiers, but they still achieved very strong performance with average ROC AUCs of 0.964 and 0.977 and PR AUCs of 0.54 and 0.597, respectively.

### *Feature importance varies for classifiers trained on different non-digenic sets*

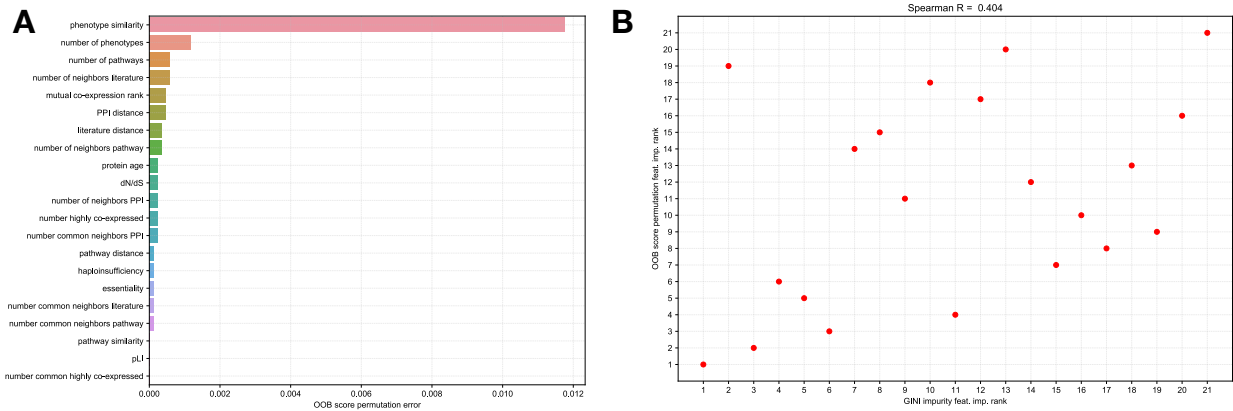
I estimated the importance of the features to the classifiers using the mean decrease in node impurity approach (**Figure 8**). For the classifier trained using variant gene pairs from unaffected relatives, the Jaccard similarity of phenotypes associated with each gene for a gene pair was the

unaffected	0.09	0.33	0.04	0.05	0.04	0.03	0.02	0.02	0.02	0.02	0.03	0.04	0.06	0.03	0.02	0.03	0.03	0.01	0.02	0.03	0.03
permuted	0.07	0.16	0.06	0.04	0.04	0.03	0.03	0.02	0.05	0.05	0.05	0.06	0.05	0.05	0.03	0.04	0.04	0.01	0.05	0.03	0.05
random	0.08	0.28	0.05	0.04	0.02	0.02	0.03	0.02	0.04	0.03	0.03	0.08	0.05	0.04	0.03	0.03	0.03	0.01	0.03	0.03	0.03
matched	0.02	0.07	0.06	0.02	0.02	0.03	0.04	0.03	0.06	0.05	0.06	0.10	0.07	0.05	0.06	0.08	0.06	0.01	0.04	0.02	0.05
unaffected no gene overlap	0.10	0.37	0.04	0.06	0.03	0.02	0.02	0.01	0.02	0.02	0.03	0.04	0.07	0.02	0.02	0.02	0.02	0.01	0.02	0.02	0.02
random no gene overlap	0.10	0.37	0.05	0.05	0.03	0.02	0.02	0.01	0.03	0.02	0.02	0.05	0.07	0.03	0.02	0.03	0.02	0.01	0.02	0.02	0.02
	pathway similarity	phenotype similarity	mutual co-expression rank	PPI distance	pathway distance	literature distance	protein age	essentiality	pLI	dN/dS	haploinsufficiency	number of pathways	number of phenotypes	number of neighbors PPI	number of neighbors pathway	number of neighbors literature	number highly co-expressed	number common highly co-expressed	number common neighbors PPI	number common neighbors pathway	number common neighbors literature

**FIGURE 8:** *Phenotype features were most important for classifier to identify digenic pairs; evolutionary features more important for matched pairs*

GINI feature importance values for classifier to identify digenic gene pairs from unaffected, random, matched and permuted non-digenic gene pairs. Phenotype similarity (37%) had highest importance, followed by pathway similarity (10%), number of phenotypes (7%) and PPI distance (6%). Evolutionary and genomics features more important when distinguishing digenic pairs from matched gene pairs.

highest weighted feature (37%). The pathway similarity and the mean number of phenotypes for the gene pair were among the other important features (10% and 7% of the weight, respectively). The feature importance values were similar for the classifiers trained using random gene pairs and permuted digenic gene pairs (**Figure 8**).



**FIGURE 9: Comparison of feature importance ranks measured using GINI and permutation OOB error approaches**

**(A)** The feature importance values were calculated using the out-of-bag (OOB) error score after permutation of feature values. The values denoted the mean error in classification of out-of-bag samples after scrambling of feature values. Phenotype similarity and number of phenotypes had the highest OOB error scores. **(B)** The feature importance ranks calculated using the GINI mean decrease in impurity approach (X-axis) and the out-of-bag (OOB) error score after permutation of feature values (Y-axis). The most and least important features using both approaches were the same (phenotype similarity and number common highly co-expressed, respectively). The high importance and low importance features generally overlapped. The Spearman rho correlation is 0.404. The one major discrepancy between the methods was the importance of pathway similarity, which had high GINI importance (rank=2), but low permutation OOB importance (rank=19). The reason for this was the ability of the model to rely on phenotype similarity and pathway distance to compensate, which does not lead to misclassification in the permutation analysis. In contrast, the high GINI importance for phenotype similarity suggests that it was often selected as representative of these features in the trained models.

The feature importance values were most different for the matched classifier; it placed significantly lower feature importance on the NFFs. This was expected, because by design the differences between the positive and negative training examples in individual NFFs were minimal for this classifier. Instead, a range of evolutionary and individual gene-level functional features took on similar levels of importance (**Figure 8**). This indicates that information in gene-level features related to evolution, gene importance, and relevance to physiology contain useful information about the likelihood of gene pairs interacting to produce digenic disease.

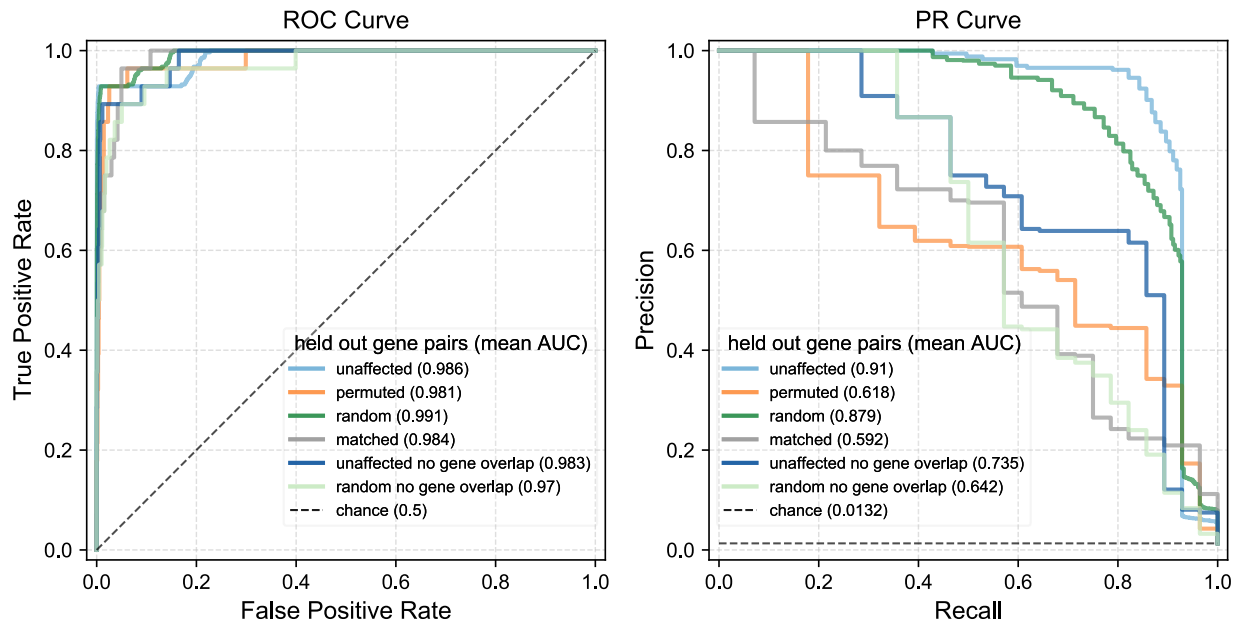
The impurity approach for feature importance calculation can be biased, especially when the classification task includes features with both continuous and discrete values<sup>125</sup>. Therefore, I

also used a permutation approach to calculate the importance for each feature based on the error in classification after the feature values were scrambled. Phenotype similarity was still the most important feature (**Figure 9A**), and the feature importance values calculated based on the impurity and the permutation approach generally agreed (Spearman rho = 0.404, **Figure 9B**).

#### *DiGePred accurately identifies held-out digenic pairs*

To obtain an unbiased estimate of the best classifiers' performance, I evaluated them using held-out test sets of digenic and non-digenic pairs. These sets were not used for training or validating the classifier and maintained the 1:75 ratio used during training. The classifiers trained using gene pairs observed in unaffected relatives of UDN patients as negatives most closely reflect the distribution of gene mutations likely to be seen in real clinical applications. Based on the previous results, the best balance between performance and stringency in selecting the negatives was achieved for the unaffected no gene overlap model, with all the features used during training. I focused on this model going forward, but report results for all classifiers.

The ROC AUC for the unaffected no overlap classifier on the held-out sets was 0.983, while the mean PR AUC was 0.735 (**Figure 10**). The classifiers trained on the other non-digenic gene pair sets also performed well on their corresponding held-out sets: the ROC AUCs were better than 0.97 and PR AUCs were better than 0.59 in all cases (**Figure 10**).



**FIGURE 10:** *Classifiers accurately distinguish digenic pairs from non-digenic pairs on held-out test sets.*

(A) ROC and (B) PR curves for random forest classifiers trained using all features on digenic gene pairs and various negative sets (indicated in the legend) and evaluated on the appropriate held-out test sets. These test sets consisted of DIDA held-out pairs as positives and six different held-out negative sets: i) *Unaffected*, derived from healthy relatives of UDN patients (light blue); ii) *Permuted*, derived by generating permutations of known digenic pairs (orange); iii) *Random*, derived by randomly selecting pairs of genes (dark green); iv) *Matched*, derived by matching the distribution of network and functional features observed among the digenic pairs (grey); v) *Unaffected no gene overlap*, derived from healthy relatives of UDN patients and no genes in common between the training and test datasets (dark blue); vi) *Random no gene overlap*, derived by randomly selecting pairs of genes with no genes in common between the training and test datasets (light green). The area under the ROC curves (AUROCs) were  $>0.97$  in all cases, while the area under the PR curves (AUPRs) were  $>0.6$  in all cases. In all subsequent analyses, the *Unaffected no gene overlap* classifier will be referred to as “DiGePred”.

To establish thresholds for predicting potential digenic gene pairs based on the output of the unaffected no gene overlap classifier, I computed thresholds that maximize the  $F_1$  and  $F_{0.5}$  scores. The  $F_1$  is maximized at a digenic score of 0.156, and the  $F_{0.5}$  is maximized at a digenic score of 0.496. Since I anticipate that precision is more important than recall in most applications, I suggest use of the  $F_{0.5}$ -based threshold. At this threshold, the classifier correctly



identified 13 of 28 digenic gene pairs in the held-out test set, with a false positive rate of 0.14% (**Figure 10, Dataset D1**). I refer to this model as the DiGenic Predictor (DiGePred).

*DiGePred identifies novel digenic pairs from the recent literature*

While the test set was not seen by the classifier prior to evaluation, it was still obtained from DIDA, the source of digenic pairs for training and testing. Thus, I further applied DiGePred to 13 digenic pairs obtained from recent literature, not included in DIDA (**Table T1**). I derived three digenic pairs ((*CEP290, RPE65*), (*AH11, CEP290*), (*CEP290, CRB1*)) from the validation set used by a recently published digenic classifier<sup>43</sup>. The other digenic gene pairs ((*CLCNKA, CLCNKB*), (*TCF3, TNFRSF13B*), (*IFNAR1, IFNGR2*), (*PCDH15, USH1G*), (*LAMA4, MYH7*), (*KCNE2, KCNH2*), (*CLCNKB, SLC12A3*), (*CACNA1C, SCN5A*), (*FGFR1, KLB*), (*CLCN7, TCIRG1*)) were derived from recently reported cases of digenic disease, respectively: (Abdallah et al., 2019; Ameratunga et al., 2017; Heida et al., 2019; Hoyos-Bachiloglu et al., 2017; Kong et al., 2019; Nieto-Marín et al., 2019; Nozu et al., 2008; Schrauwen et al., 2018; Stone et al., 2019; Yang et al., 2018). I noted that these pairs include some similar phenotypes and overlapping genes, and so should not be viewed as 13 independent tests.

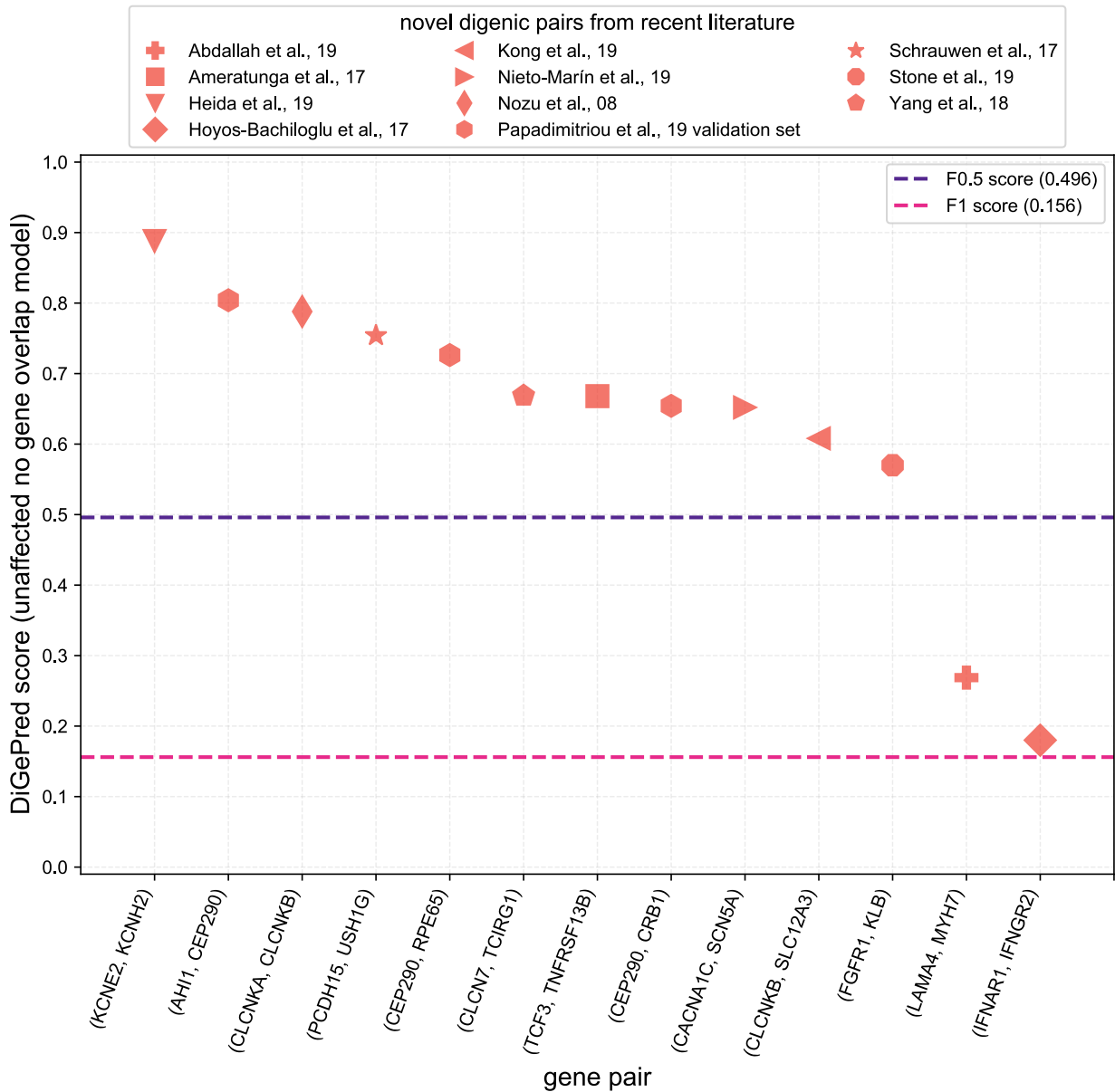
DiGePred correctly identified 11 of the 13 novel digenic pairs at the  $F_{0.5}$  threshold. (**Figure 11-12**). Two of the gene pairs missed at the  $F_{0.5}$  threshold, *IFNAR1* and *IFNGR2* (Hoyos-Bachiloglu et al. 2017) and *LAMA4* and *MYH7* (Abdallah et al. 2019) were identified as digenic at the  $F_1$  threshold (expected FPR of 0.5%) (**Figure 11-12**).

TABLE T1: Novel digenic pairs from recent literature, not in Digenic Database (DIDA) and not used for training.

#	Gene A	Gene B	Paper	Phenotypes	DiGePred score	
1	AHI1	CEP290	Coppieters et al., 10,	Papadimitriou et al., 19 (validation set)	Leber Congenital Amaurosis, Joubert syndrome	0.804
			Coppieters et al., 10			
			Coppieters et al., 10			
2	CEP290	RPE65			Leber Congenital Amaurosis	0.726
3	CEP290	CRB1			Leber Congenital Amaurosis	0.654
4	TCF3	TNFRSF13B	Ameratunga et al., 17		Primary immunodeficiency disorder and systemic lupus erythematosus	0.668
5	IFNAR1	IFNGR2	Hoyos-Bachiloglu et al., 17		Primary immunodeficiency	0.18
6	PCDH15	USH1G	Schrauwen et al., 17		Profound non-syndromic hearing impairment	0.754
7	LAMA4	MYH7	Abdallah et al., 19		Infantile Dilated Cardiomyopathy	0.269
8	KCNE2	KCNH2	Heida et al., 19		Long QT Syndrome Type 2 and Type 6	0.888
9	CLCNKB	SLC12A3	Kong et al., 19		Gitelman syndrome	0.608
10	CACNA1C	SCN5A	Nieto-Marín et al., 19		Long QT phenotype	0.652

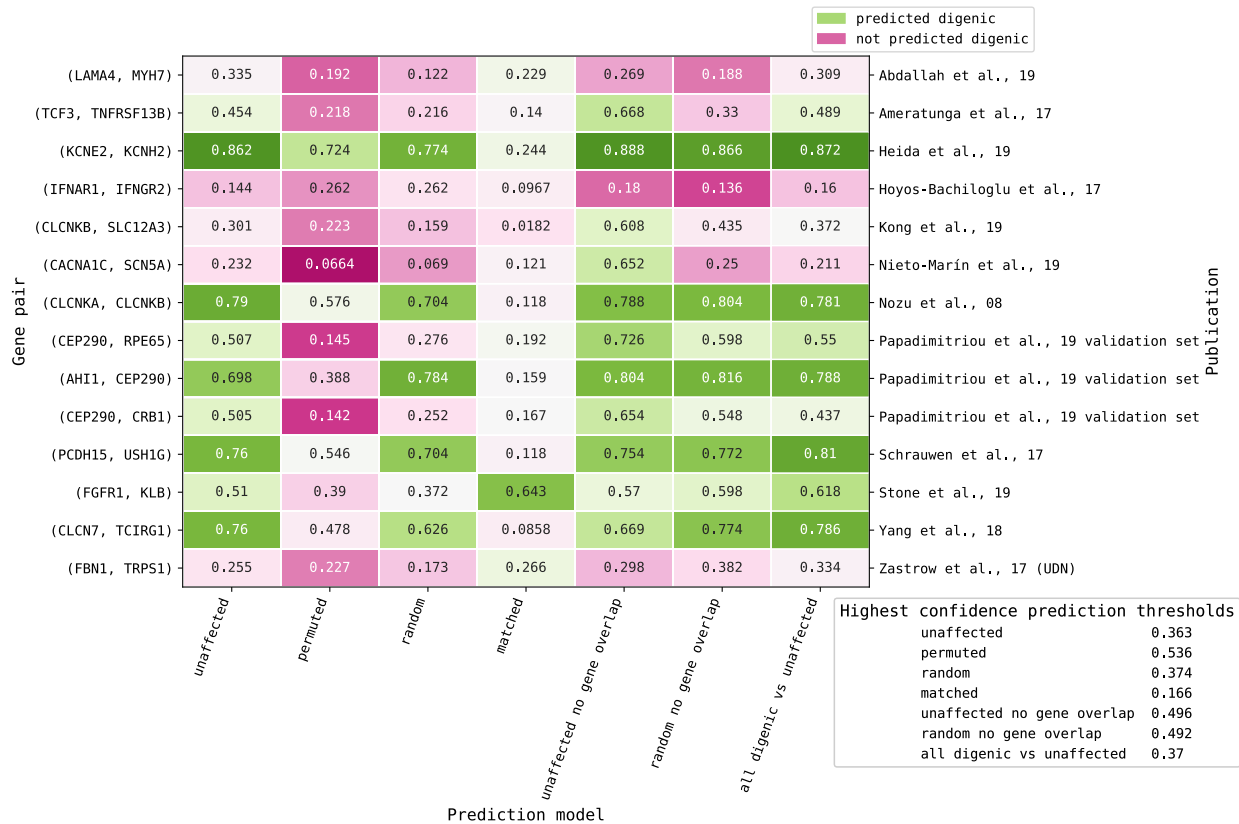
<b>11</b>	FGFR1	KLB	Stone et al., 19	Endocrine Specific FGF-21 Signaling Defects and Extreme Insulin Resistance	0.57
<b>12</b>	CLCN7	TCIRG1	Yang et al., 18	Osteopetrosis	0.669
<b>13</b>	CLCNKA	CLCNKB	Nozu et al., 2008	Bartter syndrome, sensorineural deafness	0.788
<b>14</b>	<i>FBNI</i>	<i>TRPS1</i>	<i>Zastrow et al., 17 (UDN)</i>	<i>Marfan syndrome and TRPS1</i>	<i>0.239</i>

Genes, publication, and patient phenotypes for 13 recently identified digenic disease pairs. Red indicates a DiGePred score beneath the F0.5 threshold. The 14<sup>th</sup> pair (italics) is not strictly digenic, but underlies disease in a UDN patient.



**FIGURE 11:** *DiGePred* accurately identifies novel digenic pairs from the recent literature.

Geometric shapes in red indicate the *DiGePred* scores assigned to 13 novel digenic pairs reported in the recent literature. The dashed pink and purple lines represent the *DiGePred* score thresholds that maximize the  $F_1$  (0.156) and the  $F_{0.5}$  (0.496) metrics (**Figure 13**). Given the importance of precision in clinical applications, I propose the score maximizing the  $F_{0.5}$  metric or higher as a threshold for calling a gene pair digenic. At this threshold 11 of the 13 novel digenic pairs are predicted to be digenic with a low expected false positive rate ( $\leq 0.14\%$ ). All digenic pairs score above the  $F_1$  threshold. The *DiGePred* classifier was trained using all features and the unaffected no gene overlap set as negatives.



**FIGURE 12:** Validation of other models of classifier using novel digenic pairs from recent literature.

The novel digenic pairs from recent literature and name of first author of the publication are along the Y axis. The predicted scores by the different models of DiGePred, trained on different negative sets, are along the X axis. Green indicates prediction as digenic based on  $F_{0.5}$  prediction threshold, pink indicates no digenic prediction. 11/13 novel digenic pairs are identified as digenic by the unaffected no gene overlap model. Solved UDN case with overlapping phenotypes not predicted as digenic at highest threshold.

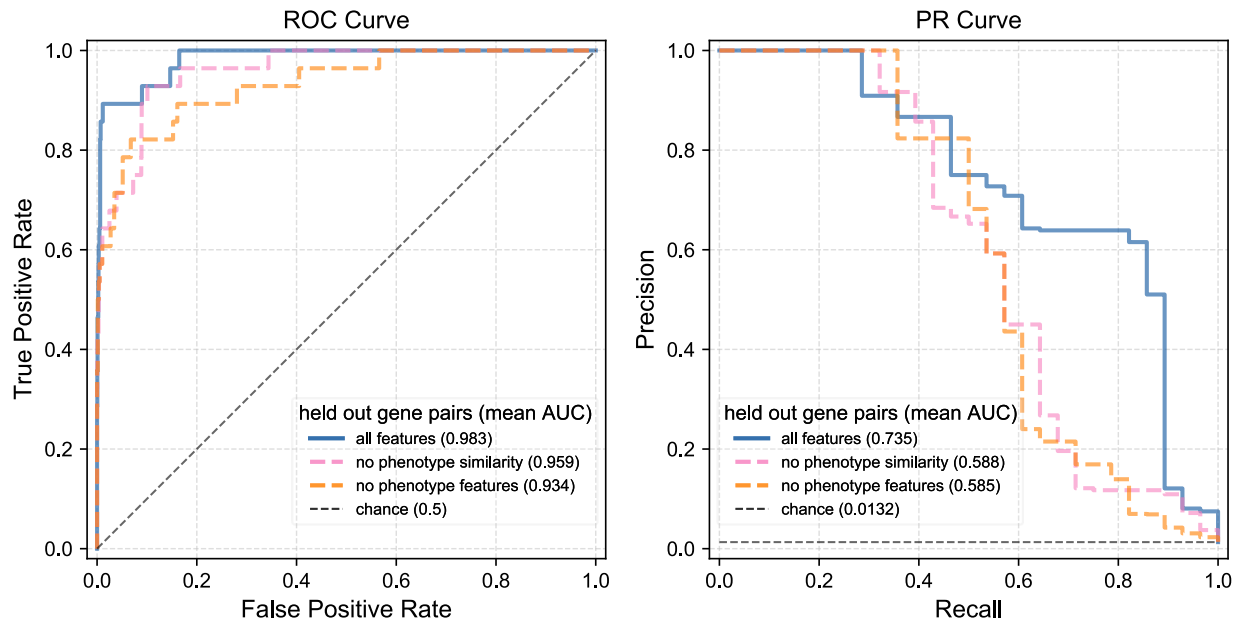
I also evaluated a gene pair from a solved UDN case in which variants in *FBN1* and *TRPS1* caused independent autosomal dominant conditions with some overlapping symptoms that produced a unique phenotype in the patient (Zastrow et al. 2017) <sup>136</sup>. Due to the lack of interaction, this pair does not meet the strict criteria for digenic pairs used here. DiGePred predicted that this gene pair was not digenic at the  $F_{0.5}$  threshold. Nonetheless, it was predicted at

the  $F_1$  threshold (**Figure 11-12, Dataset D2**), suggesting the potential of the classifier to highlight pairs of functionally related genes.

### *Discussion*

In this paper I describe DiGePred, a high-throughput machine-learning approach to identify gene pairs with the potential to cause digenic disease. I demonstrate the accuracy and robustness of the approach in several realistic scenarios. I was motivated to create DiGePred by the challenge of identifying causal variants in patients with rare disease that cannot be explained by a single variant. It is not feasible to experimentally evaluate all candidate pairs of variants in a patient of interest.

The DiGePred classifier trained using negatives derived from unaffected relatives is likely best suited to the purpose of identifying digenic pairs in patients with rare disease, because it reflects the baseline distribution of gene pairs with variants identified using clinical sequencing pipelines in individuals without severe disease. Moreover, classifiers trained using these negative sets performed well. However, this approach performs well at distinguishing digenic pairs from several additional sets of candidate non-digenic gene pairs, and the features used by these classifiers are similar unless the prediction problem is explicitly engineered to make them different (**Figure 11-12**).



**FIGURE 13:** Leaving out phenotype features reduces DiGePred performance, but it remains strong.

**(A)** ROC and **(B)** PR curves for random forest classifiers trained on all features (blue), leaving out phenotype similarity (pink), and leaving out all phenotype features (orange) using held-out testing digenic and unaffected no gene overlap pairs. The AUCs were significantly lower without phenotype features ( $p$  values  $< 1.34 \times 10^{-24}$ ), but the models maintain strong performance. The area under the ROC curves (AUROCs) were  $> 0.934$  in both cases, while the area under the PR curves (AUPRs) were  $> 0.585$  in all cases.

Positives: digenic pairs from DIDA; held-out testing  
 Negatives: non-digenic pairs from unaffected relatives of UDN patients  
 No gene level overlap between training and testing datasets.

The features prioritized by DiGePred support previous work<sup>39,42</sup> in that phenotypic similarity, number of phenotypes, and involvement in the same molecular pathways are the most important predictors. They also suggest that these may be more specific predictors of digenic gene pairs than similar co-expression profiles or close interaction network distance. The results using negatives that match the network and functional features between positives and negatives sets indicate that digenic gene pairs also have differences in their evolutionary attributes.

These analyses are based on the examples available in DIDA, but there are likely hundreds or even thousands of undiscovered digenic diseases. The strong performance of

DiGePred on the test set with no gene overlap with the training set and DiGePred's ability to identify new digenic pairs from the recent literature (**Figure 11-12**) suggest that the algorithm will generalize. However, I note that the performance estimates may be optimistic; known digenic pairs are unlikely to be an unbiased sample of the full spectrum of digenic mechanisms. I anticipate that the algorithms will further improve as more digenic diseases and their causal molecular mechanisms are determined.

DiGePred is based on functional, biological network, and evolutionary features in a Random Forest model. Phenotype similarity and other phenotype related features, such as the mean number of phenotypes associated with each individual gene, were the most important features. Given that our understanding of function of most genes is incomplete, the high reliance on a phenotype-based features could lead to a high performing model that does not generalize when these features are missing. I retrained and evaluated DiGePred leaving out either phenotype similarity or all phenotype related features. There was a decrease in performance on the held-out test set (P value  $< 1.34 \times 10^{-24}$ ) (**Figure 13**); however, the classifiers maintained substantial accuracy, with a ROC AUCs  $>0.93$  and PR AUCs  $>0.585$ . Thus, while the models are likely somewhat biased by existing knowledge, the strong performance is not only due to overlapping phenotypic annotations.



## **CHAPTER 3: Application of DiGePred to real world scenarios – predict digenic pairs in individuals suffering from rare diseases**

### *Summary*

I have developed DiGePred, a method for identifying gene pairs with digenic disease potential, and generated predictions for all pairs of human genes. The use of this tool on rare disease patients illustrates its potential to provide insight in real-world settings, and I anticipate that it will have broad utility in clinical genome interpretation. In contrast to other approaches, DiGePred also appropriately controls the number of false positives when applied in realistic clinical settings. Finally, to enable the rapid screening of variant gene pairs for digenic disease potential, we freely provide the predictions of DiGePred on all human gene pairs.

These findings were published as a part of the DiGePred manuscript published in AJHG in 2021. A novel digenic candidate identified was published in Mikhael S et al., 2021 (Mikhael S, Dugar S, Morton M, Chorich LP, Tam KB, Lossie AC, Kim HG, Knight J, Taylor HS, **Mukherjee S**, Capra JA, Phillips JA 3rd, Friez M, Layman LC. Genetics of agenesis/hypoplasia of the uterus and vagina: narrowing down the number of candidate genes for Mayer-Rokitansky-Küster-Hauser Syndrome. Hum Genet, 2021 140, 667–680 (2021); <https://doi.org/10.1007/s00439-020-02239-y>)

### *Introduction*

Application of machine learning methods to rare diagnosis has been performed previously.<sup>137</sup> With an increased focus on rare diseases research, there is an increase in rare disease cohorts.<sup>138,140</sup> The Undiagnosed Diseases Network (UDN)<sup>17,18</sup> performs genome sequencing

individuals suffering from rare diseases who have applied to and been accepted as a part of the UDN cohort for analysis. In addition, both parents and siblings, when available and other relatives, when applicable are sequenced as a part of the cohort as well. Often the relatives of the individual with rare disease show no clinical phenotypes and carry no rare deleterious variants. This provides sequencing data for unaffected individuals for the UDN cohort as well, and provides additional levels of analyses that can be performed to help identify rare and unique causative variants or variant combinations, that are not present in other unaffected relatives.

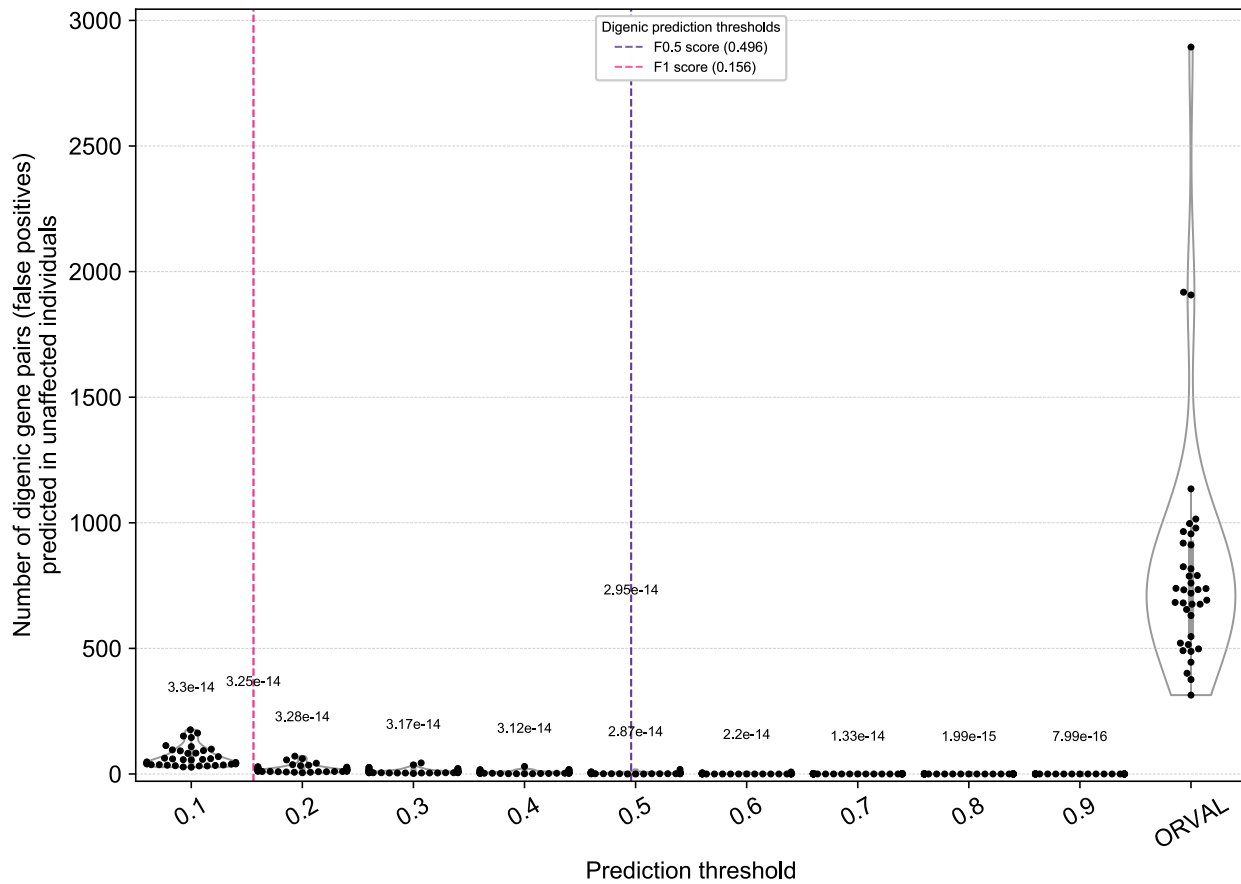
There is a comprehensive database of digenic diseases<sup>39</sup> and machine learning methods to identify digenic gene pairs from genome sequencing data currently available<sup>42-44</sup>. However, the efficacy and accuracy of these methods have not been evaluated in real world scenario, on rare disease cohorts.

I evaluated the accuracy of DiGePred on the UDN and other rare disease cohorts and demonstrate that it has a low false positive rate, which is essential for clinical applications. To aid in rapid screening of patients for potential digenic disease variants, I provide a classification of the digenic disease potential for all human gene pairs.

## Results

### DiGePred has a low false positive rate in real-world applications

Individuals often carry hundreds of protein-coding variants of unknown significance, which results in thousands of potential digenic disease pairs per individual. Thus, when considering the



**FIGURE 14:** *DiGePred has a low false positive rate and outperforms a recent digenic gene prediction method.*

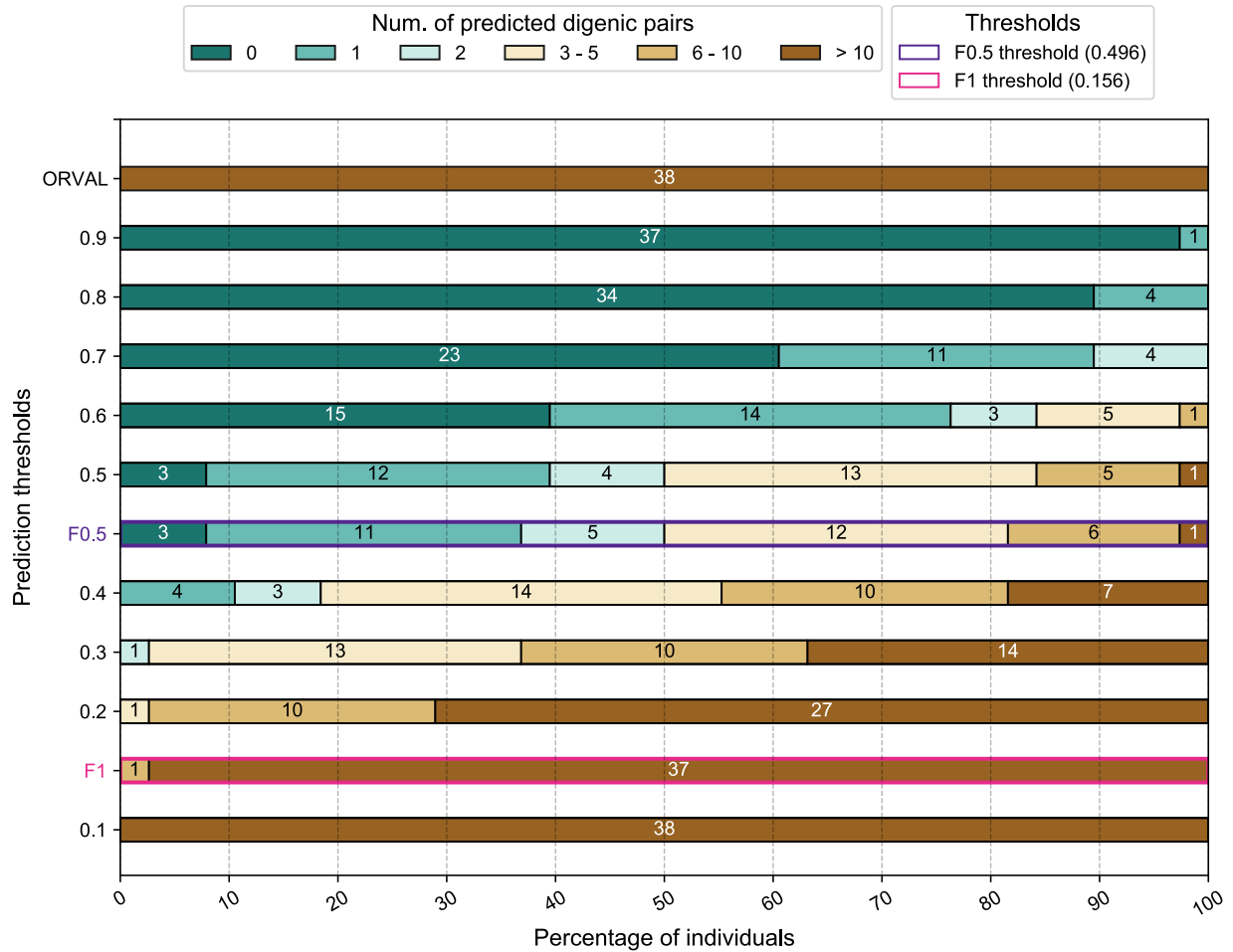
The number of digenic pairs identified for each of 38 healthy relatives of UDN patients is plotted at a range of DiGePred thresholds (x-axis) and for the highest confidence predictions (99% threshold) of the ORVAL/VarCOPP method. The DiGePred score thresholds that maximize the  $F_1$  and  $F_{0.5}$  metrics on the held-out data are shown in pink and purple, respectively. Since the individuals considered are healthy, any predicted digenic disease pairs are very likely false positives. DiGePred predicts significantly fewer digenic pairs at each threshold than ORVAL (Mann-Whitney U test, p-values above each bar). At the  $F_{0.5}$  threshold, DiGePred predicts an average of under four digenic pairs per healthy individual and none above the 0.9 threshold, while ORVAL predicts an average of 830 digenic pairs per healthy individual at its strictest threshold. Results were similar for classifiers trained on other negative sets (**Figures 15-23**).

application of classifiers to individuals' genomes, it is essential to understand and control the false positive rate. To this end, I evaluated DiGePred on gene pairs with rare variants predicted to disrupt protein function in 38 human genomes from unaffected parents and relatives of UDN patients not used in training the algorithm. These healthy individuals should not contain any true digenic disease pairs, so any positive predictions on gene pairs from these individuals are very likely to be false positives. The gene pairs from these individuals were not used in the training, validation, or held-out test sets.

At the  $F_{0.5}$  threshold, 8% of unaffected individuals had no predicted candidate digenic pairs and 29% had only one candidate digenic pair. On average, less than four digenic pairs were predicted per individual, and only six had more than five pairs (**Figure 14-15**). Furthermore, I emphasize that users can adjust the score threshold to reflect their tolerance for false positives in different applications; for example, the fraction of individuals with no digenic gene pairs predicted was 31%, 66% and 92% at score thresholds of 0.6, 0.7 and 0.8, respectively (**Figure 15**).

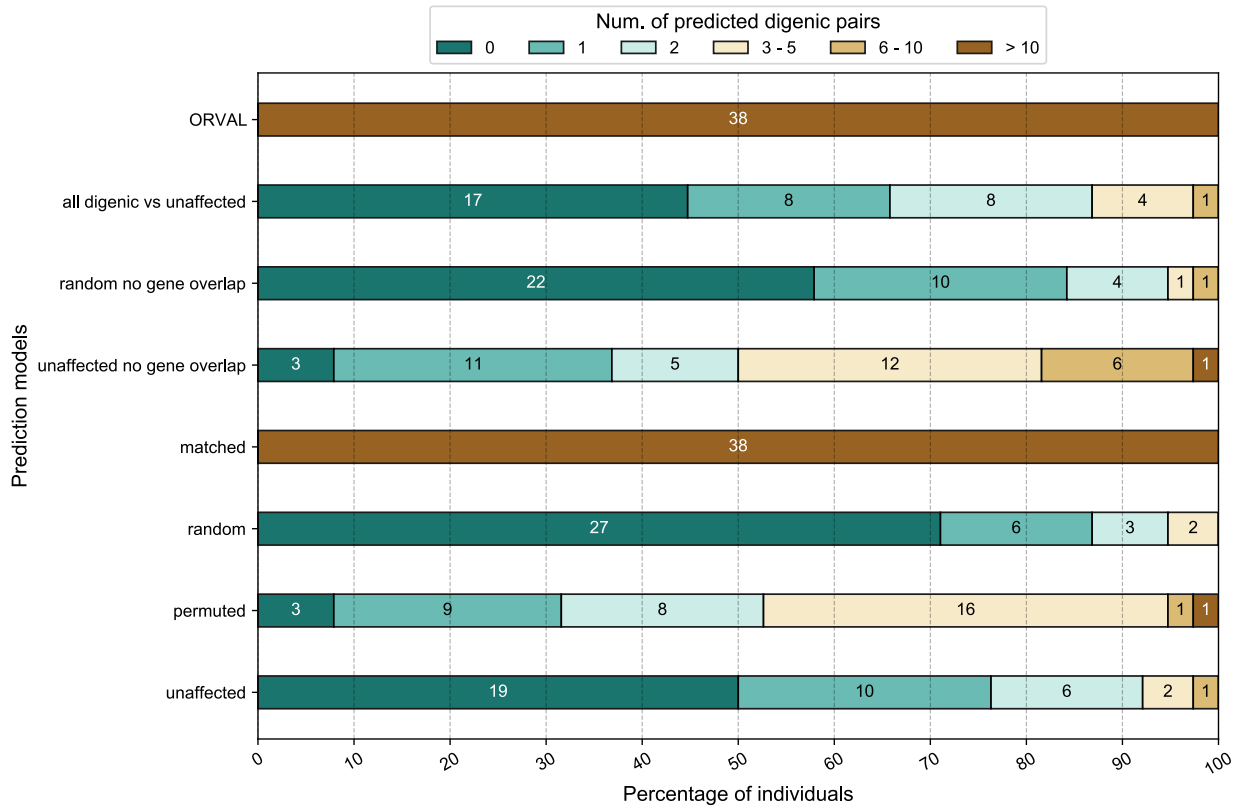
In contrast, I applied the ORVAL<sup>43,45</sup> method for identifying digenic disease pairs to variants from these same individuals. At its highest confidence threshold, ORVAL predicted that all these healthy individuals have digenic disease pairs, with an average of 830 highest confidence digenic pairs per individual. All individuals were predicted to have > 300 digenic pairs, and 5 (~13%) had more than a thousand digenic pairs predicted (**Figure 14**). This is a significantly larger number of candidate digenic disease pairs per individual than DiGePred ( $P = 2.95 \times 10^{-14}$ , MWU test), and these are very likely to be false positives given that these are healthy individuals. This difference in number of false positives was recapitulated for all gene selection

criteria, variant pathogenicity prediction approaches, and all models of training considered (Figure 15-23).



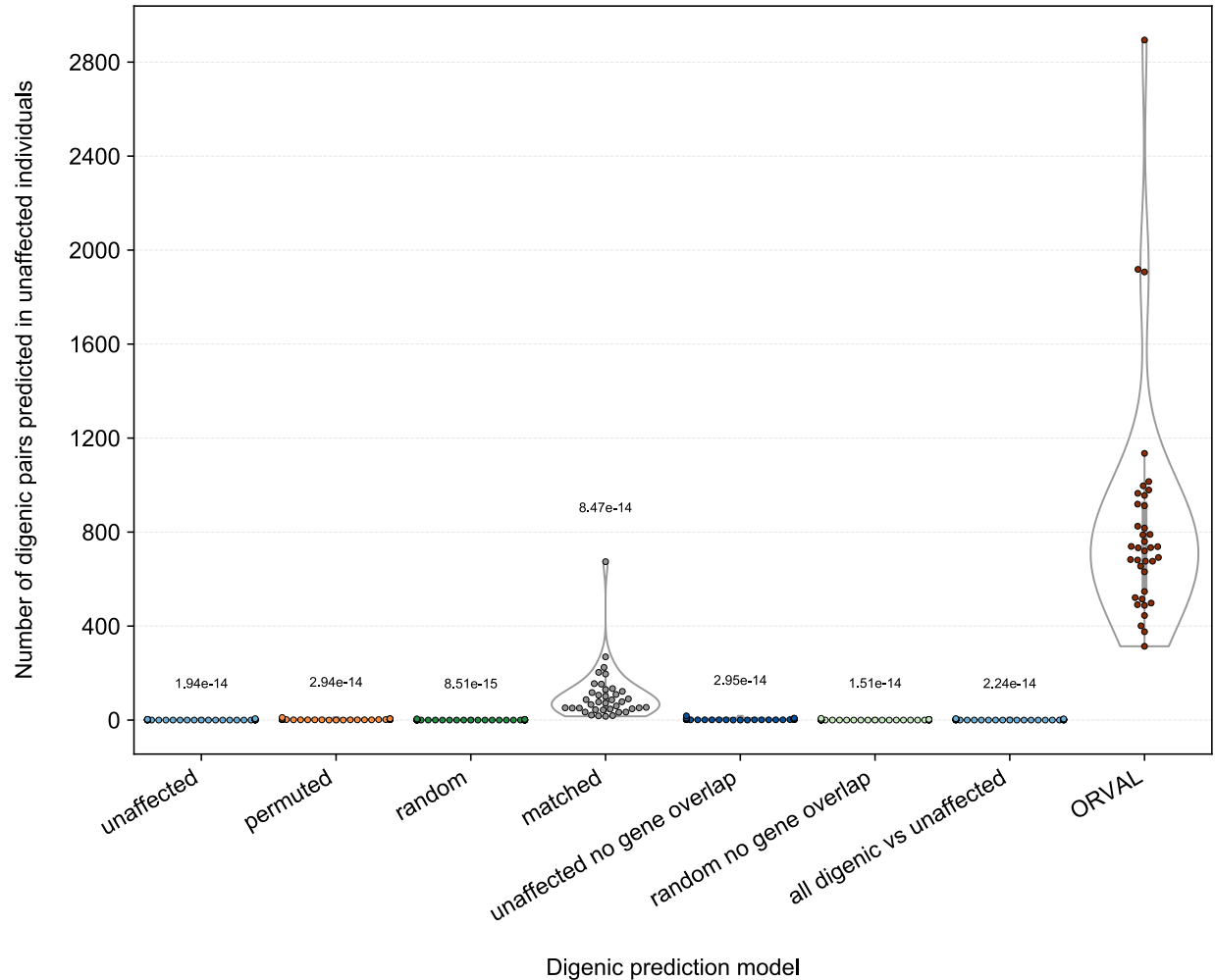
**FIGURE 15:** Low false positive rate of classifier on external negative test set of gene pairs from unaffected relatives of UDN patients; comparison with recently published variant combination pathogenicity predictor at various prediction thresholds

Number of gene pairs from individuals without digenic disease (unaffected relatives of UDN patients; n=38) identified to be digenic at varying predicted probability thresholds shown on X axis. Percentage of individuals with zero (dark green), one (green), two (light green), three - five (beige), six - ten (light brown) and > ten (dark brown) predicted digenic pairs shown on X axis; prediction thresholds shown on Y axis. ORVAL is a recently published variant combination pathogenicity predictor. The number in the box indicates number of individuals in each category. The prediction thresholds (F<sub>1</sub> score, shown in pink and F<sub>0.5</sub> score threshold shown in purple). At the F<sub>0.5</sub> threshold, three (7.9%) of unaffected individuals had no predicted digenic pairs, while 11 (29%) had only one predicted digenic pair. Only seven (18%) individuals had more than five digenic pairs and under four digenic pairs were predicted per individual on average. All individuals were predicted to have more than ten digenic pairs by ORVAL, with an average number of 830 predicted digenic pairs per individual.



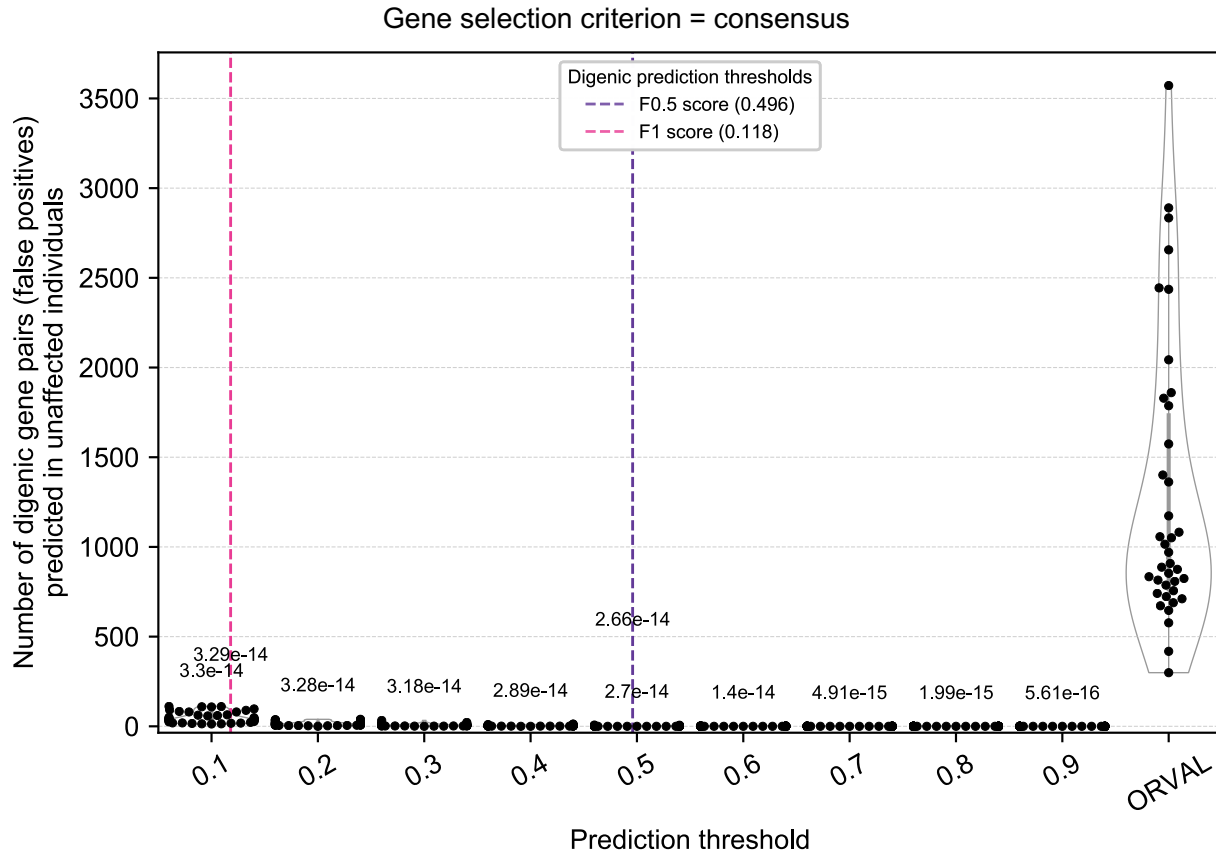
**FIGURE 16:** *Low false positive rate of classifier on external negative test set of gene pairs from unaffected relatives of UDN patients; comparison with recently published variant combination pathogenicity predictor using various models of training.*

Number of gene pairs from individuals without digenic disease (unaffected relatives of UDN patients; n=38) identified to be digenic at varying predicted probability thresholds shown on X axis. Percentage of individuals with zero (dark green), one (green), two (light green), three - five (beige), six - ten (light brown) and > ten (dark brown) predicted digenic pairs shown on X axis; models of training on Y axis ORVAL (VarCOPP) is a recently published variant combination pathogenicity predictor. The number in the box indicates number of individuals in each category. At the  $F_{0.5}$  threshold, all models, except matched, has an average of fewer than four digenic pairs, with at least 50% of unaffected individuals having fewer than three predicted digenic pairs. ORVAL predicts >300 digenic pairs for every individual.



**FIGURE 17:** Fewer false positives for DiGePred compared to ORVAL for other models of classifier.

The number of digenic pairs identified for each of 38 healthy relatives of UDN patients is plotted for different models of DiGePred, trained on different negative sets, (x-axis) and for the ORVAL/VarCOPP method. Since these individuals are healthy, any predicted digenic disease pairs a very likely false positives. DiGePred predicts significantly fewer digenic pairs for every model than ORVAL (MWU test, p-values above each bar). DiGePred trained on *Unaffected no gene overlap* (dark blue) pairs predicts an average of two digenic pairs per healthy individual, while the *Unaffected* (light blue), *Permuted* (orange), *Random* (dark green), *Matched* (grey), *Random no gene overlap* (light green) and *All digenic vs unaffected* (light blue) predict an average of 0.9, 2.7, 0.5, 103, 3.4, 0.7 and 1.2 digenic pairs per individual respectively. ORVAL predicts an average of 830 digenic pairs per healthy individual.

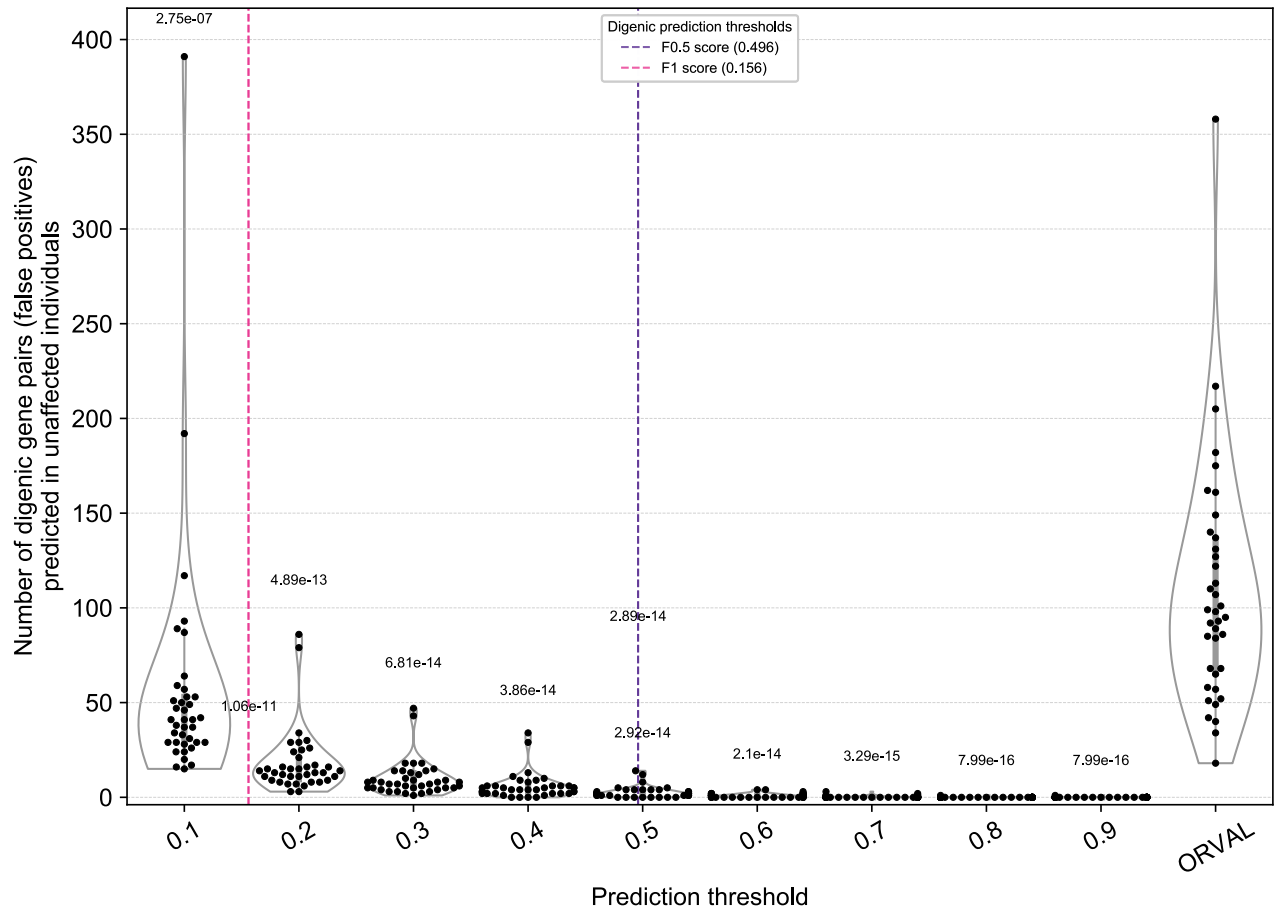


**FIGURE 18:** Fewer false positives for DiGePred compared to ORVAL when genes selected based on predicted deleterious variant effect.

The number of digenic pairs identified for each of 38 healthy relatives of UDN patients is plotted at a range of DiGePred thresholds (x-axis) and for the ORVAL/VarCOPP method. The score thresholds that maximize the F<sub>1</sub> and F<sub>0.5</sub> metrics on the held out data are shown in pink and purple, respectively. DiGePred predicts significantly fewer digenic pairs at each threshold than ORVAL (MWU test, p-values above each bar). The genes are selected by a Consensus pathogenic criterion. At the F<sub>0.5</sub> threshold, DiGePred predicts an average of 1.5 digenic pairs per healthy individual, while ORVAL predicts an average of 1286.

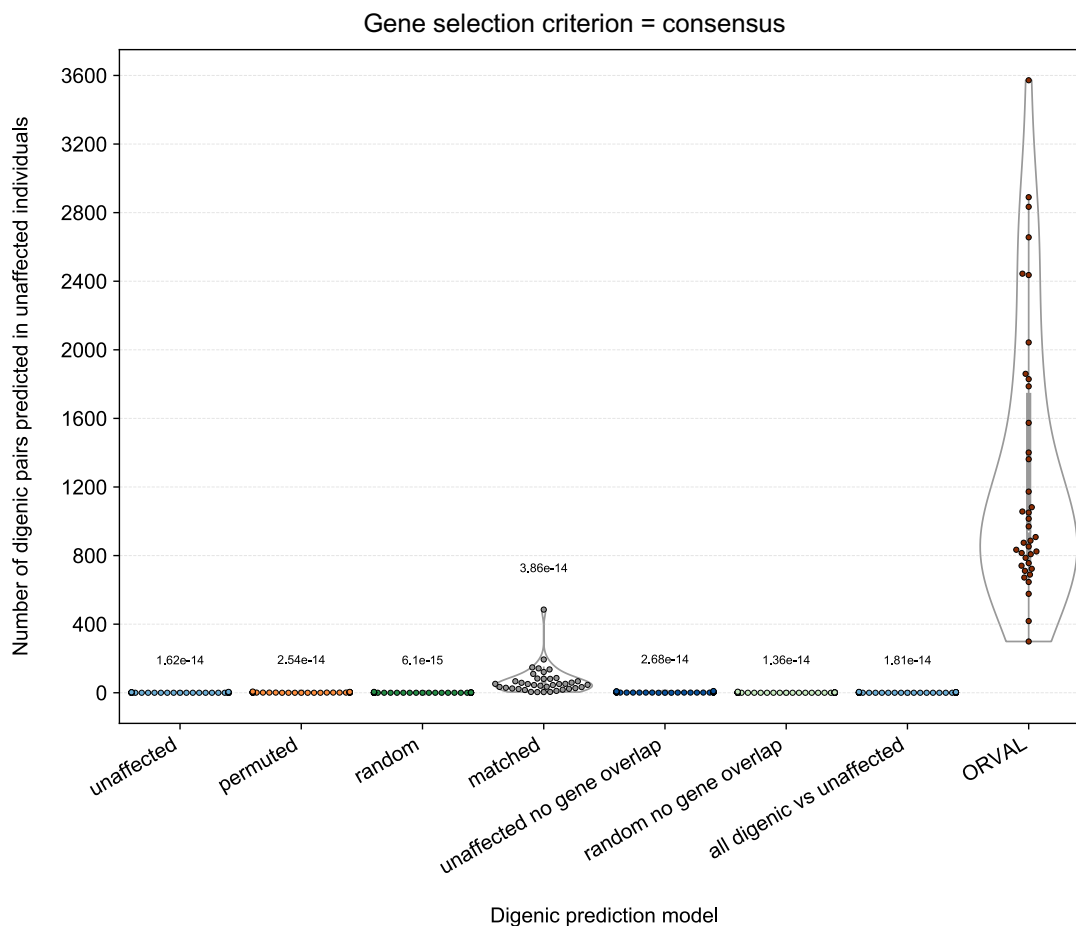


### Gene selection criterion = random



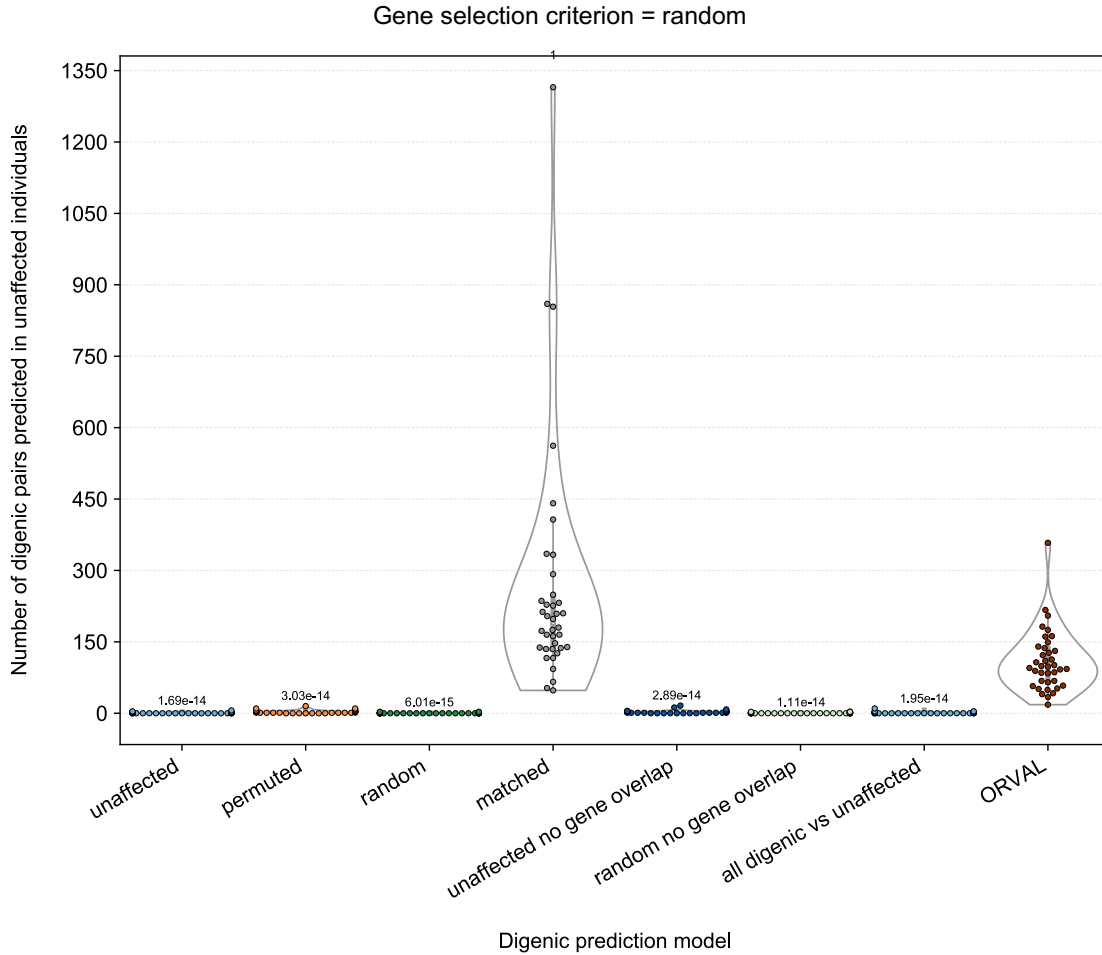
**FIGURE 19:** Fewer false positives for DiGePred compared to ORVAL when genes selected randomly.

The number of digenic pairs identified for each of 38 healthy relatives of UDN patients is plotted at a range of DiGePred thresholds (x-axis) and for the ORVAL/VarCOPP method. The score thresholds that maximize the  $F_1$  and  $F_{0.5}$  metrics on the held out data are shown in pink and purple, respectively. DiGePred predicts significantly fewer digenic pairs at each threshold than ORVAL (MWU test, p-values above each bar). The genes are selected by a Random selection. At the  $F_{0.5}$  threshold, DiGePred predicts an average of 2.7 digenic pairs per healthy individual, while ORVAL predicts an average of 108.



**FIGURE 20:** Fewer false positives for other models of DiGePred compared to ORVAL when genes selected based on predicted deleterious variant effect.

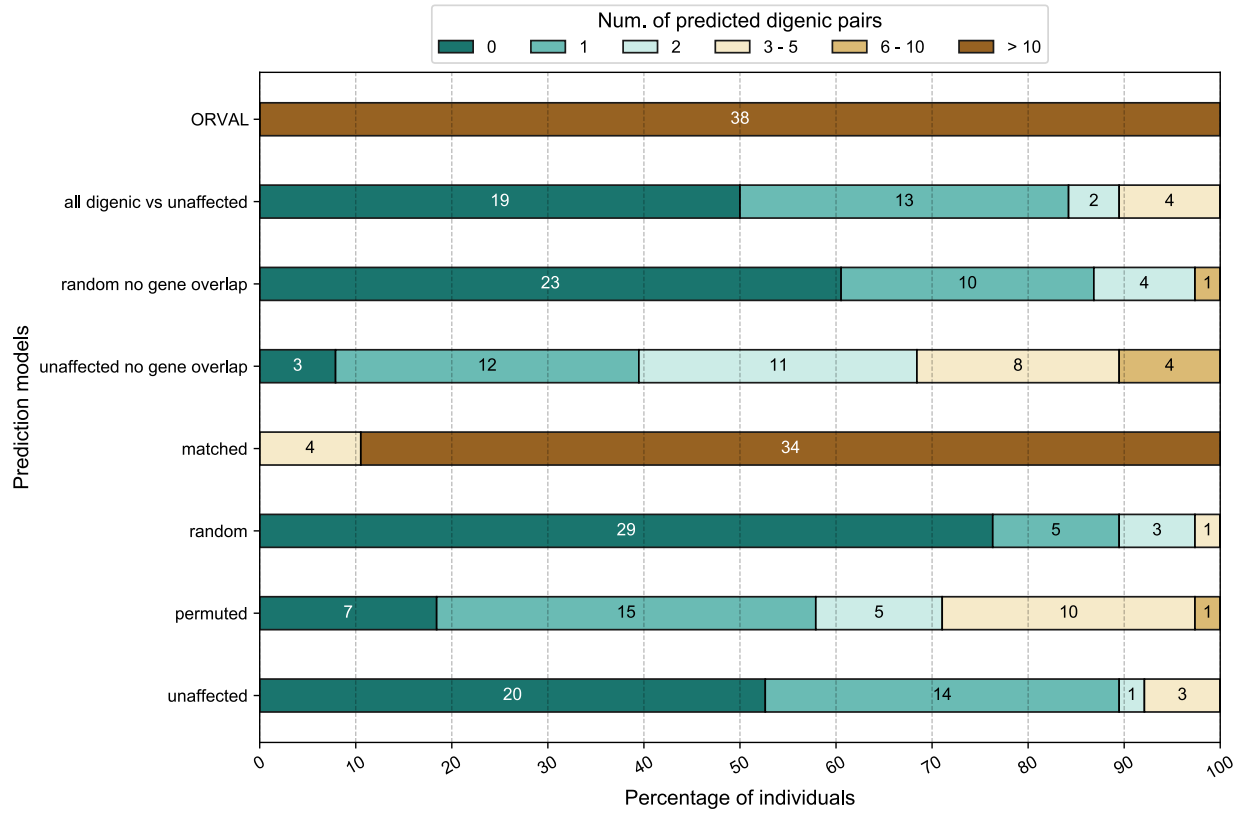
The number of digenic pairs identified for each of 38 healthy relatives of UDN patients is plot along the Y axis for training models of DiGePred among **Unaffected no gene overlap** (dark blue) pairs predicts an average of two digenic pairs per healthy individual, while the **Unaffected** (light blue), **Permuted** (orange), **Random** (dark green), **Matched** (grey), **Random no gene overlap** (light green) and **All digenic vs unaffected** (light blue), along the X-axis. DiGePred is compared to ORVAL (X axis), a recently published digenic predictor. DiGePred predicts significantly fewer digenic pairs for every model than ORVAL (MWU test, p-values above each bar). The genes are selected by a Consensus pathogenic criterion. DiGePred predicts an average of 0.7 (U), 1.8 (P), 0.4 (R), 68 (M), 2.7 (Un), 0.63 (Rn) and 0.84 (A). ORVAL predicts an average of 1285.



**FIGURE 21:** Fewer false positives for other models of DiGePred compared to ORVAL when genes selected randomly.

The number of digenic pairs identified for each of 38 healthy relatives of UDN patients is plot along the Y axis for training models of DiGePred among **Unaffected no gene overlap** (dark blue) pairs predicts an average of two digenic pairs per healthy individual, while the **Unaffected** (light blue), **Permuted** (orange), **Random** (dark green), **Matched** (grey), **Random no gene overlap** (light green) and **All digenic vs unaffected** (light blue), along the X-axis. DiGePred is compared to ORVAL (X axis), a recently published digenic predictor. DiGePred predicts significantly fewer digenic pairs for every model than ORVAL (MWU test, p-values above each bar). The genes are selected by a Random selection DiGePred predicts an average of 10.87 (U), 3.1 (P), 0.34 (R), 265 (M), 2.7 (Un). 0.58 (Rn), 1.07 (A), ORVAL predicts an average of 108.

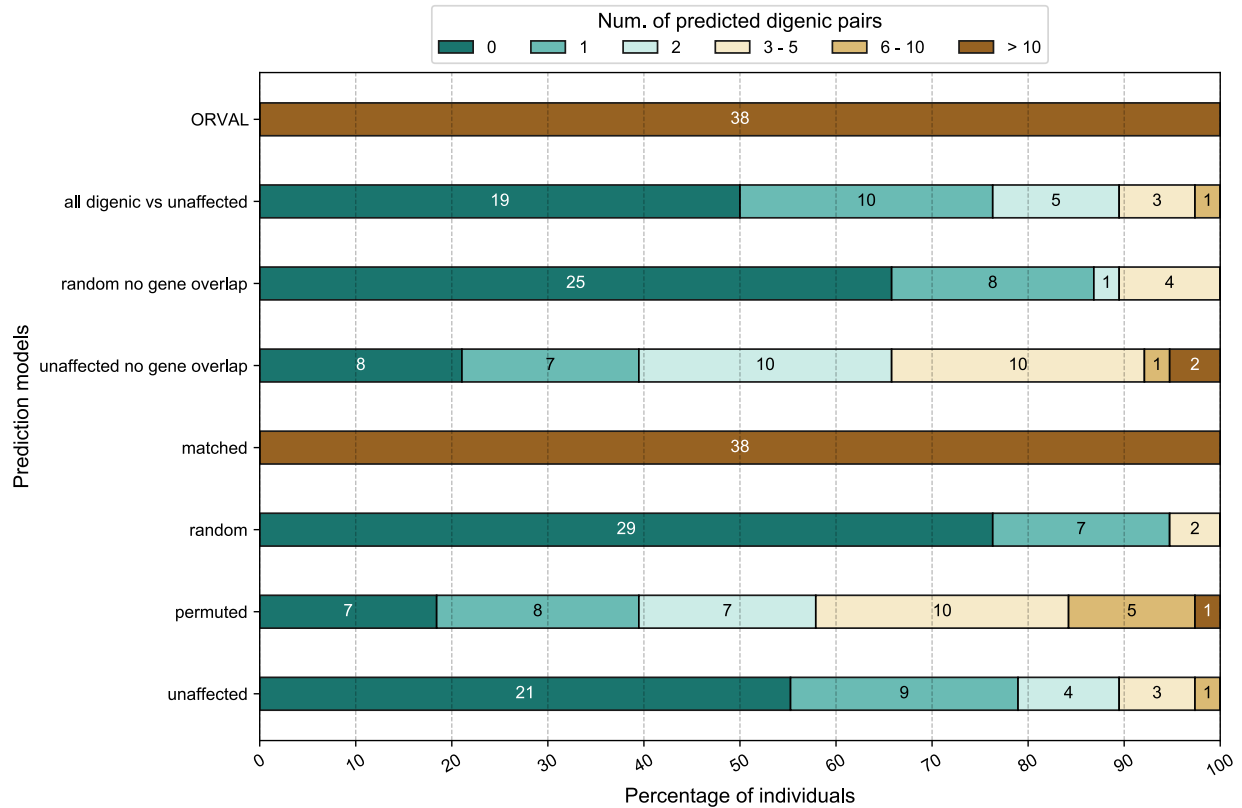
Gene selection criterion = consensus



**FIGURE 22:** *Low false positive rate of classifier on external negative test set of gene pairs from unaffected relatives of UDN patients; comparison with recently published variant combination pathogenicity predictor using various models of training, when genes selected based on predicted deleterious variant effect.*

Number of gene pairs from individuals without digenic disease (unaffected relatives of UDN patients; n=38) identified to be digenic at varying predicted probability thresholds shown on X axis. Percentage of individuals with zero (dark green), one (green), two (light green), three - five (beige), six - ten (light brown) and > ten (dark brown) predicted digenic pairs shown on X axis; models of training on Y axis ORVAL is a recently published variant combination pathogenicity predictor. The number in the box indicates number of individuals in each category. At the  $F_{0.5}$  threshold, all models except matched has fewer than three digenic pairs predicted for > 68% of unaffected individuals. ORVAL predicts >299 digenic pairs for every individual.

Gene selection criterion = random



**FIGURE 23:** Low false positive rate of classifier on external negative test set of gene pairs from unaffected relatives of UDN patients; comparison with recently published variant combination pathogenicity predictor using various models of training, when genes selected randomly.

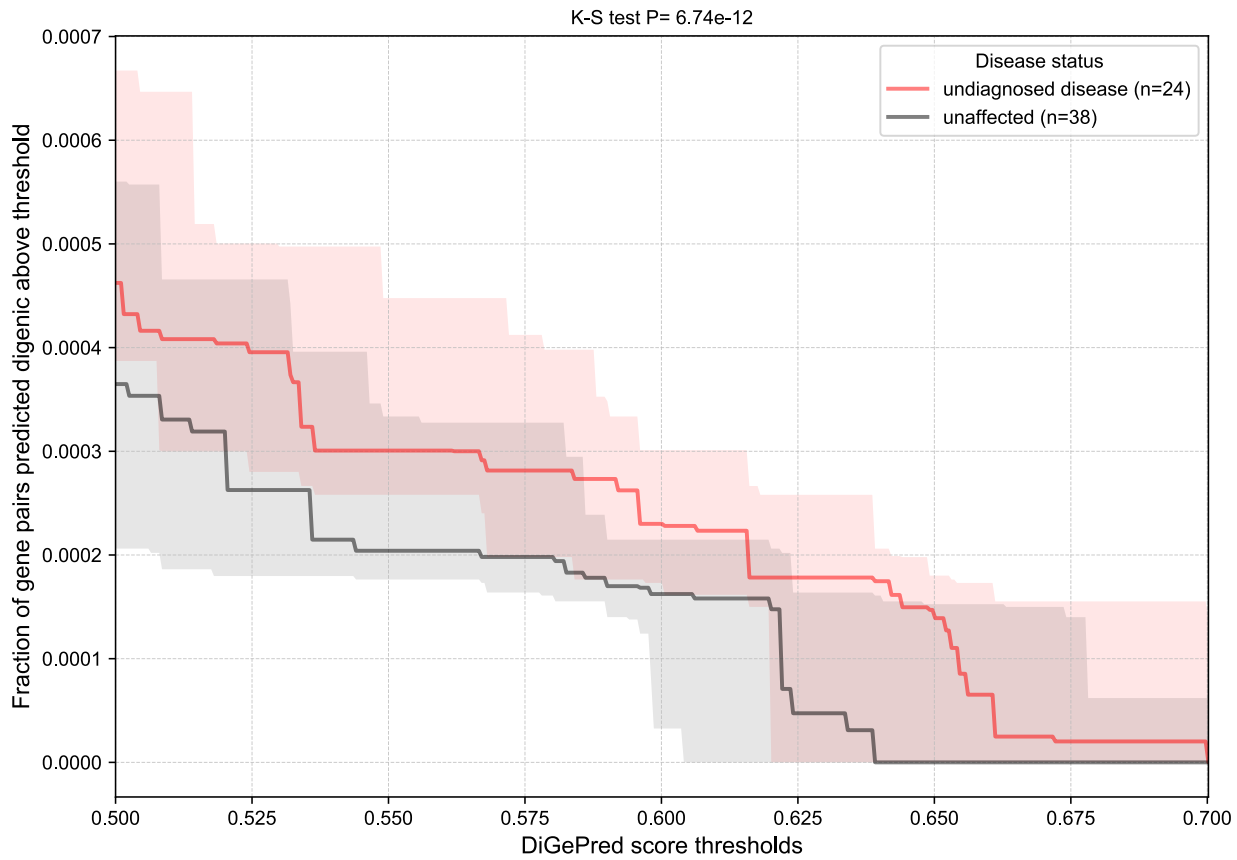
Number of gene pairs from individuals without digenic disease (unaffected relatives of UDN patients; n=38) identified to be digenic at varying predicted probability thresholds shown on X axis. Percentage of individuals with zero (dark green), one (green), two (light green), three - five (beige), six - ten (light brown) and > ten (dark brown) predicted digenic pairs shown on X axis; models of training on Y axis ORVAL is a recently published variant combination pathogenicity predictor. The number in the box indicates number of individuals in each category. At the F0.5 threshold, all models except matched has fewer than three digenic pairs predicted for > 58% of unaffected individuals. ORVAL predicts >18 digenic pairs for every individual.

I found that 11.8% of false positives in unaffected individuals (gene pairs incorrectly predicted as digenic by DiGePred) had at least one gene as a member of a known digenic pair in DIDA. Only 3.2% of all gene pairs evaluated by DiGePred had at least one gene overlapping with known digenic pairs from DIDA. This is an approximately 4-fold enrichment of such gene pairs among false positives compared to the genome-wide expectation (P-value=1.31x10<sup>-05</sup>).

#### Prediction of digenic pairs among patients with undiagnosed disease

To illustrate the application of DiGePred in patients with rare undiagnosed genetic disorders, I applied it to: 1) patients from the UDN site at Vanderbilt and 2) a cohort of 111 individuals with Mayer–Rokitansky–Küster–Hauser (MRKH) syndrome.

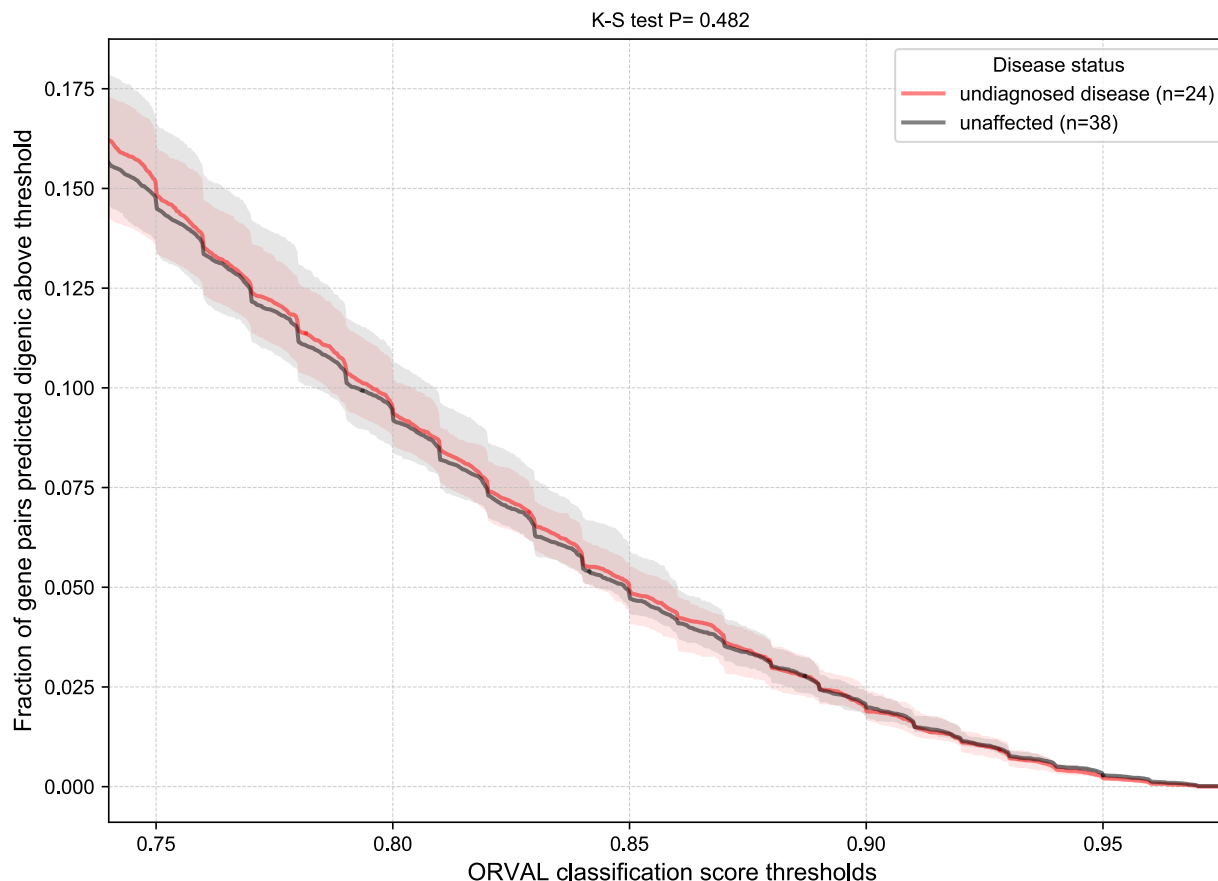
I first considered variants from ~50 UDN cases and identified several candidate digenic pairs based on DiGePred score integrated with analyses of variant effect, variant inheritance, and similarity of the gene's functions to the patient phenotype. Since these cases are still being actively evaluated, I cannot report full details here. Instead, I describe a representative example. I predicted a candidate digenic pair of *ATXN2* (Ataxin 2) and *FUS* (fused in sarcoma) for a patient with ALS (amyotrophic lateral sclerosis) and Parkinsonism like phenotypes. The variant in *ATXN2* was a polyglutamine (polyQ) repeat expansion variant, and there is evidence in literature for a functional interaction between these two genes<sup>139,141,142</sup>.



**FIGURE 24:** *DiGePred predicts that UDN patients have more digenic gene pairs above high confidence thresholds than unaffected relatives.*

The median fraction of gene pairs predicted to be digenic by DiGePred (Y-Axis) above the corresponding threshold (X-Axis) for 24 individuals with undiagnosed disease (red) and 38 unaffected individuals (black). The distribution of the median number of predicted digenic pairs for individuals with undiagnosed disease is significantly greater than for unaffected relatives (P-value of  $6.74 \times 10^{-12}$ , Kolmogorov–Smirnov (K-S) test). The shading around the lines indicates the 95% confidence interval around each median. All high confidence prediction thresholds ( $F_{0.5}$  threshold of 0.496 and greater) for which at least one of the medians was  $> 0$  were considered.

To explore the performance of DiGePred on UDN individuals more quantitatively, I compared the predictions on variants from 24 patients with 38 available unaffected relatives that were not used in the training of DiGePred. I tested whether the rare disease patients had a higher median of fraction of high-confidence predicted digenic pairs compared to related individuals without rare disease. At all thresholds considered ( $F_{0.5}$  or higher), DiGePred predicted a greater



**FIGURE 25:** *ORVAL predicts that UDN patients and unaffected relatives have similar numbers of digenic gene pairs at high confidence thresholds.*

The median fraction of gene pairs predicted to be digenic by ORVAL classification score (Y-Axis) above the corresponding threshold (X-Axis) for 24 individuals with undiagnosed disease (red) and 38 unaffected individuals (black). The distributions for the individuals with undiagnosed disease and unaffected individuals were not significantly different (P-value of 0.482, Kolmogorov–Smirnov (K-S) test). The shading around the lines indicate the 95% confidence interval around the median. All high confidence prediction thresholds (ORVAL classification scores of 0.74 and greater, corresponding to the 99% confidence zone) for which at least one of the medians was  $> 0$  were considered.

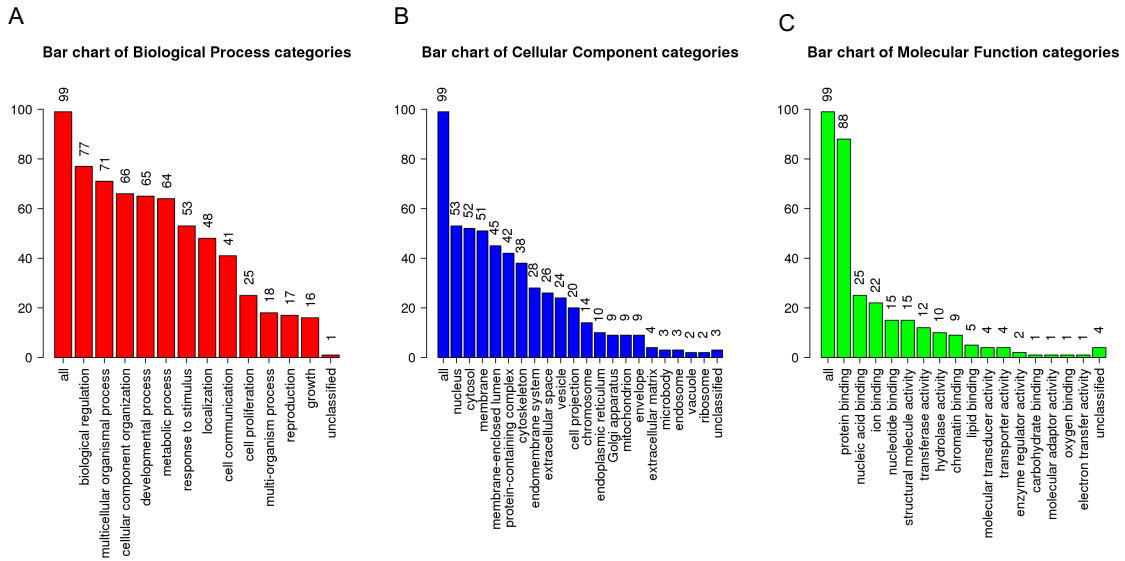
fraction of gene pairs with variants to be digenic for individuals with undiagnosed disease than the unaffected individuals. The difference between the distributions was significant ( $P=6.74 \times 10^{-12}$ , Kolmogorov–Smirnov test; **Figure 24**). In contrast, the fraction of predicted digenic pairs was similar for the individuals with undiagnosed disease compared to unaffected individuals across a range of ORVAL classification scores within the 99% confidence zone ( $P=0.482$ ; **Figure 25**).



Next, I applied DiGePred to variants from a cohort of 111 individuals with MRKH syndrome<sup>143</sup>, a developmental disorder primarily affecting the female reproductive system, often characterized by a congenital absence of a uterus or vagina<sup>144,145</sup>. I identified a potential digenic pair between *LAMC1* (Laminin Subunit Gamma 1), an extracellular matrix (ECM) glycoprotein that is a member of the Integrin pathways and plays a role in cell adhesion and signaling, and *MMP14* (Matrix Metalloproteinase 14), a protein involved in breaking down the extracellular matrix during embryonic development and tissue remodeling. The DiGePred prediction was driven by the two proteins being highly co-expressed with one another, directly interacting along the Integrin pathway, being only one protein away on the global PPI network, and having ~5% phenotype similarity. Furthermore, there is evidence in literature of functional interaction between *LAMC1* and *MMP14* that affects ECM remodeling via fibronectin deposition in zebrafish<sup>146</sup>.

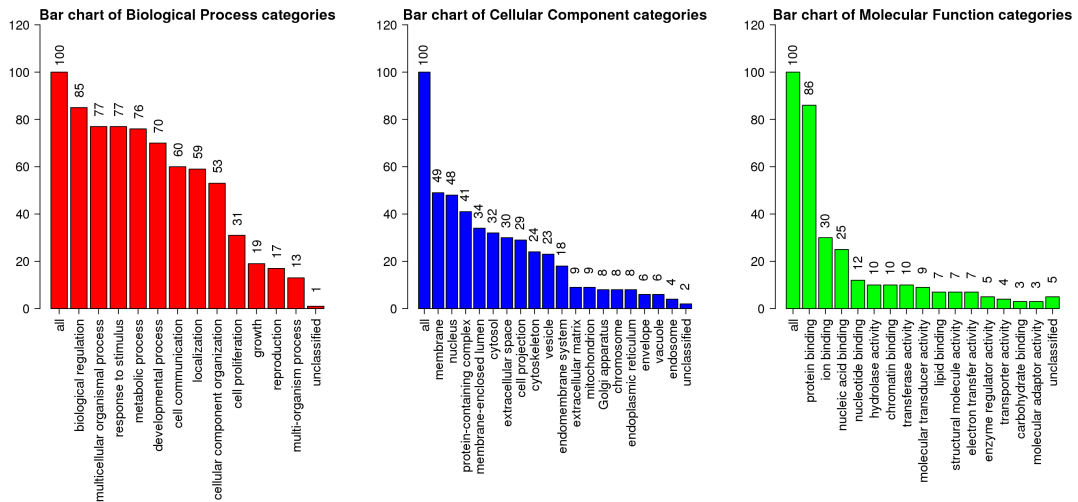
#### *Prediction of digenic pairs among all human gene pairs at various confidence thresholds*

To aid in the rapid evaluation of digenic disease potential for a pair of genes of interest, I trained a new DiGePred classifier using all digenic pairs from DIDA (to maximize use of available data) and variant gene pairs from healthy relatives of UDN patients. I applied DiGePred to all possible human gene pairs. A gene pair was deemed a candidate digenic pair if the digenic score met the



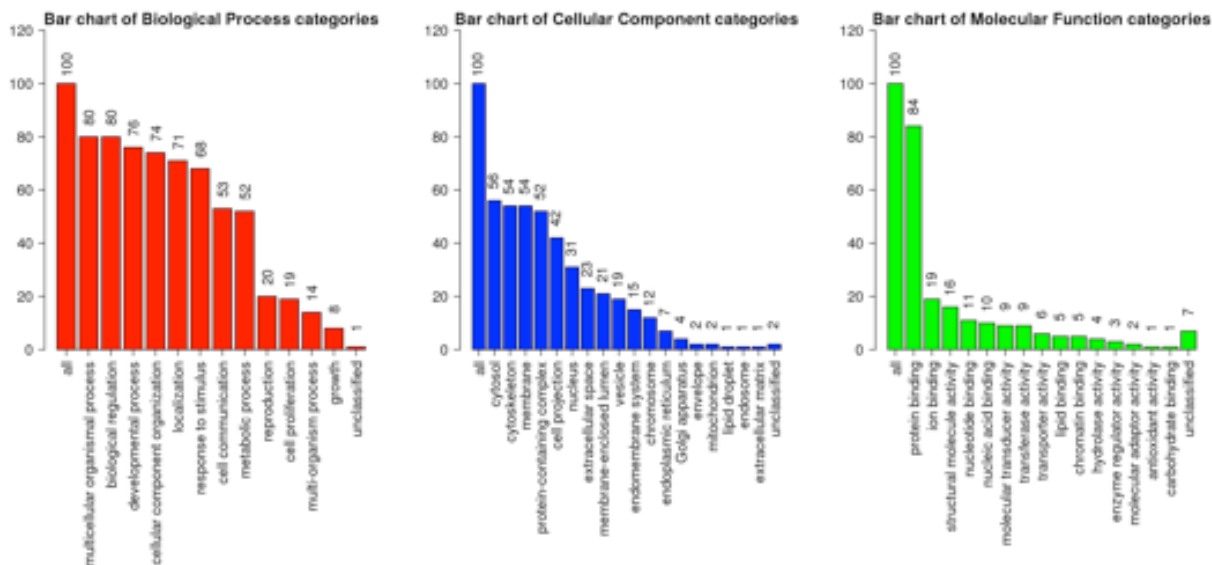
**FIGURE 26:** Gene Ontology enrichment for top 100 genes with most predicted digenic pairs.

Gene ontology (GO) enrichment using WebGestalt (WEB-based GENE SET ANALYSIS TOOLKIT). GO terms along X axis, and number along Y axis. (A) Biological process (red); (B) Cellular Component (blue); (C) Molecular function (green). Most genes are involved in metabolic processes, localized to the cell membranes and multi-protein complexes and have important binding domains.



**FIGURE 27:** Gene Ontology enrichment for top 100 genes with highest average predicted value.

Gene ontology (GO) enrichment using WebGestalt (WEB-based GENE SET ANALYSIS TOOLKIT). GO terms along X axis, and number along Y axis. (A) Biological process (red); (B) Cellular Component (blue); (C) Molecular function (green). Most genes are involved in metabolic processes, localized to the cell membranes and multi-protein complexes and have important binding domains.



**FIGURE 28:** Gene Ontology enrichment for genes in the top 100 gene pairs with highest predicted value.

Gene ontology (GO) enrichment using WebGestalt (WEB-based GENE SeT AnaLYsis Toolkit). GO terms along X axis, and number along Y axis. (A) Biological process (red); (B) Cellular Component (blue); (C) Molecular function (green). Most genes are involved in metabolic processes, localized to the cell membranes and multi-protein complexes and have important binding domains.

$F_{0.5}$  threshold as described above. As expected, the percentage of all possible gene pairs that were identified as digenic at the most confident threshold was very low (54,318 out of 155.33 million gene pairs, 0.035%). These predictions and the raw digenic scores are available in

### Dataset D3.

Overall, 7,970 unique genes are involved in at least one predicted digenic pair. This illustrates that DiGePred is not just prioritizing gene pairs that include a gene in a known digenic pair. In fact, only 3 of the top 100 genes with the most predicted digenic pairs occur in a DIDA pair. These genes are enriched for several essential developmental and molecular Gene Ontology functional annotations including “maintenance of cell number” (7.5x expected, FDR=0.005), “chromatin remodeling” (7.3x, FDR=0.005), and “membrane docking” (7.0x expected,

FDR=0.004; **Table T2**). For example, *FGF5*, a growth factor important for cell proliferation and differentiation, tissue development and repair, had the highest number of predicted digenic pairs above the  $F_{0.5}$  threshold with 370. *ARID1B*, which had the 2<sup>nd</sup> highest number of predicted digenic pairs, with 262, encodes a component of the SWI/SNF chromatin remodeling complex with broad regulatory functions across the genome. *CEP290*, a centrosome protein, with essential roles in centrosome and cilia development in many cell types had the 6<sup>th</sup> most predicted digenic interactions with 232. The genes with the most predicted digenic pairs were also enriched for several organ development and cell cycle processes. The top 100 gene pairs with the highest average DiGePred scores were enriched for tissue and organ development, ciliary function, and electron transfer activity (**Figure 26-28, Tables T2-4**).

**TABLE T2: Gene Ontology enrichment for top 100 genes with most predicted digenic pairs.**

<b>GO term</b>	<b>Description</b>	<b>Size</b>	<b>Expected</b>	<b>Enrichment Ratio</b>	<b>P- Value</b>	<b>FDR</b>
GO:0003713	transcription coactivator activity	316	2.0516	5.8491	8.3026e-7	0.00023413
GO:0051427	hormone receptor binding	184	1.1946	6.6967	2.4596e-5	0.0034681
GO:0048568	embryonic organ development	423	2.7139	4.7901	2.9198e-6	0.0024819
GO:0048880	sensory system development	355	2.2776	4.8296	1.6639e-5	0.0037935

GO:0022406	membrane docking	177	1.1356	7.0447	1.7852e-5	0.0037935
GO:0098727	maintenance of cell number	146	0.93672	7.4729	4.2022e-5	0.0051660
GO:0006338	chromatin remodeling	150	0.96238	7.2736	4.9927e-5	0.0051660
GO:0050953	sensory perception of light stimulus	209	1.3409	5.9661	5.8508e-5	0.0051660
GO:0006333	chromatin assembly or disassembly	155	0.99446	7.0390	6.1484e-5	0.0051660
GO:0044772	mitotic cell cycle phase transition	487	3.1245	3.8406	6.4116e-5	0.0051660
GO:0043583	ear development	212	1.3602	5.8816	6.4680e-5	0.0051660
GO:0044441	ciliary part	440	3.1871	3.7651	6.7244e-6	0.0038553
GO:0005819	spindle	328	2.3759	3.7881	5.6910e-4	0.024471

TABLE T3: Gene Ontology enrichment for top 100 genes with highest average predicted value.

GO term	Description	Size	Expected	Ratio	P- Value	FDR
GO:0030990	intraciliary transport particle	28	0.21671	18.458	5.9038e-5	0.010155
GO:0048018	receptor ligand activity	468	3.1144	4.4952	2.2023e-6	0.00062106
GO:0009055	electron transfer activity	111	0.73868	9.4764	8.6751e-6	0.0012232
GO:0070851	growth factor receptor binding	132	0.87843	6.8304	2.4251e-4	0.022796
GO:0016651	oxidoreductase activity, acting on NAD(P)H	106	0.70540	7.0881	6.9140e-4	0.039453
GO:0001228	DNA-binding transcription activator activity, RNA polymerase II-specific	444	2.9547	3.3844	6.9951e-4	0.039453
GO:0048732	gland development	434	2.8431	7.0345	4.8348e-12	2.0548e-9

GO:0060485	mesenchyme development	262	1.7164	9.3221	1.2885e-11	3.6508e-9
GO:0048568	embryonic organ development	423	2.7711	5.7740	1.4138e-8	0.0000024034

*TABLE T4: Gene Ontology enrichment for genes in the top 100 genes with highest average predicted value.*

<b>GO term</b>	<b>Description</b>	<b>Size</b>	<b>Expected</b>	<b>Ratio</b>	<b>P- Value</b>	<b>FDR</b>
GO:0070491	repressing transcription factor binding	73	0.43248	23.123	1.2986e-11	3.6619e-9
GO:0015631	tubulin binding	321	1.9017	5.7842	2.66203-6	0.00025023
GO:0008307	structural constituent of muscle	44	0.26067	15.345	1.2865e-4	0.0090698
GO:0007389	pattern specification process	433	2.8950	8.2900	6.6613e-16	1.8874e-13
GO:0044839	cell cycle G2/M phase transition	213	1.4241	10.533	1.0589e-11	1.2858e-9
GO:0044441	ciliary part	440	3.6674	8.7255	~0	~0

GO:0005874	microtubule	402	3.3507	3.2829	4.7853e-4	0.013718
------------	-------------	-----	--------	--------	-----------	----------

I found that 19,325 (35%) of predicted digenic gene pairs had at least one recessive phenotype associated in OMIM <sup>147-149</sup>. In almost a fifth of these cases (3,697; 19%), at least one phenotype was in common or with high semantic similarity <sup>150</sup> between the two genes. For most of these gene pairs (3,601; 97%), the two genes had different MIM numbers in OMIM. This indicates that the two genes have not been previously annotated as causing a digenic disease, <sup>147</sup> and thus, suggests that they are novel associations.

Existing knowledge provides plausible mechanisms underlying many of these predicted novel digenic gene pairs. For example, a digenic pair comprising *STIMI* and *ORAI1* had the 4<sup>th</sup> highest score over all human gene pairs. It has been previously reported that *STIMI* and *ORAI1* function together to form Ca<sup>2+</sup> release-activated Ca<sup>2+</sup> (CRAC) channels, which are responsible for Ca<sup>2+</sup> influx called store-operated Ca<sup>2+</sup> entry (SOCE) <sup>151</sup>. The proper functioning of these channels is necessary for maintaining the normal physiology of several cell types, including T cell receptors and human lymphocytes <sup>152-154</sup>. Missense variants in *STIMI* and *ORAI1*, individually, cause diseases with a great degree of phenotypic homogeneity <sup>155</sup>. Loss of function variants in *STIMI* and *ORAI1* have also been known to cause immunodeficiency, <sup>156-159</sup> under autosomal recessive conditions, as reported by OMIM. Therefore, it is possible that single loss of variants in both genes occurring simultaneously could lead to the autosomal recessive immunodeficiency.



## *Discussion*

To facilitate the rapid identification of candidate digenic gene pairs in patients, I have provided DiGePred predictions for all pairs of human genes at several confidence thresholds (**Dataset D4A-D**). DiGePred has demonstrated low false positive rate when applied to unaffected individuals, indicating its applicability to real world scenarios. I compared DiGePred to the recently published ORVAL/VarCOPP digenic disease prediction server. This method was also developed using DIDA as positive training data. Due to the challenge of running the web server on a large-scale, it was not possible to evaluate its performance in the training, validation, test framework.

I applied it to variant gene pairs from the 24 UDN patients and their 38 unaffected relatives not used in the training or initial evaluation of DiGePred. At its strictest (99%) prediction threshold, I found an average of 855 predicted digenic disease pairs per individual without disease. This false positive rate is too high for clinical use. In contrast, DiGePred predicts two or fewer digenic pairs for 47% of these individuals and an average of under four digenic pairs per individual overall. I also observed that ORVAL predicted a similar fraction of digenic pairs in the unaffected and patient groups at increasingly strict classification score thresholds (**Figure 14**). My analysis of the ORVAL method suggests that if one of the genes in a pair carries a variant that is predicted to be pathogenic by ORVAL's variant effect prediction component, then the gene pair is very likely to be predicted to be digenic. This suggests that strong variant-level effects may obscure signals specific to digenic disease.

As a part of our collaboration with the UDN, we help review individuals with rare diseases analyzed as a part of the UDN cohort by the clinical research team. This includes application of computational variant effect predictors, study of inheritance patterns, and clinical

expertise. Other models of DiGePred perform similarly well whether trained against gene pairs that have predicted disruptive variants or on all variant pairs from individuals (**Figure 15-23**), suggesting that they are not simply identifying pairs containing monogenic disease genes. Going forward, I will continue to refine this approach in collaboration with the UDN and other rare disease cohorts.

## **CHAPTER 4: A Personalized Structural Biology (PSB) approach reveals the molecular mechanisms underlying heterogeneous epileptic phenotypes caused by *de novo* KCNC2 variants**

### *Summary*

Next-generation whole exome sequencing (WES) is ubiquitous as an early step in the diagnosis of rare diseases and the interpretation of variants of unknown significance (VUS).

Developmental and epileptic encephalopathies (DEE) are a group of rare devastating epilepsies, many of which have unknown causes. Increasing WES in the clinic has identified several rare monogenic DEEs caused by ion channel variants. However, WES often fails to provide actionable insight, due to the challenges of proposing functional hypotheses for candidate variants.

Here, I have described a “personalized structural biology” (PSB) approach that addresses this challenge by leveraging recent innovations in the determination and analysis of protein 3D structures. I illustrated the power of the PSB approach in an individual from the Undiagnosed Diseases Network (UDN) with DEE symptoms who has a novel *de novo* VUS in *KCNC2* (p.V469L), the gene that encodes the Kv3.2 voltage-gated potassium channel. A nearby *KCNC2* variant (p.V471L) was recently suggested to cause DEE-like phenotypes. I found that both variants are located in the conserved hinge region of the S6 helix and likely to affect protein function. However, despite their proximity, computational structural modeling suggests that the V469L variant is likely to sterically block the channel pore, while the V471L variant is likely to stabilize the open state. Biochemical and electrophysiological analyses demonstrate heterogeneous loss-of-function and gain-of-function effects, respectively, as well as differential

inhibition in response to 4-aminopyridine (4-AP) treatment. Using computational structural modeling and molecular dynamics simulations, I illustrate that the pore of the V469L variant is more constricted increasing the energetic barrier for K<sup>+</sup> permeation, whereas the V471L variant stabilizes the open conformation

These results implicated *KCNC2* as a causative gene for DEE and guided the interpretation of a UDN case. They further delineate the molecular basis for the heterogeneous clinical phenotypes resulting from two proximal pathogenic variants. This demonstrates how the PSB approach can provide an analytical framework for individualized hypothesis-driven interpretation of protein-coding VUS suspected to contribute to disease.

This work has been communicated as a manuscript to the American Journal of Human Genetics (AJHG) Advances, and published on MedRxiv as Mukherjee et al., 2022 (**Mukherjee S, Cassini TA, Hu N, Yang T, Li B, Shen W, Moth CW, Rinker DC, Sheehan JH, Cogan JD, Undiagnosed Diseases Network, Newman JH, Hamid R, Macdonald RL, Roden DM, Meiler J, Kuenze G, Phillips JA, Capra JA. Personalized structural biology reveals the molecular mechanisms underlying heterogeneous epileptic phenotypes caused by de novo KCNC2 variants. MedRxiv doi: <https://doi.org/10.1101/2022.02.01.21268115>).**

### *Introduction*

The advent of cheaper and more accurate DNA sequencing technologies has enabled the integration of genetic information into diverse areas of medicine. For example, more than 70% of rare diseases are thought to have a genetic cause, and recent efforts have identified the causal variants for thousands of Mendelian diseases<sup>160–162</sup>. However, causal variants have not been identified for approximately half (~3000) of known rare genetic diseases<sup>163–165</sup>, and sequencing

often fails to lead to actionable insights, even after expert clinical evaluation through programs like the NIH's Undiagnosed Diseases Network (UDN) <sup>17,18,166</sup>.

Many computational methods have been developed for interpreting variants observed in clinical sequencing <sup>167-170</sup>. However, they have substantial weaknesses and often disagree <sup>171-175</sup>. In particular, commonly used tools provide only ill-defined, categorical variant classifications like “pathogenic” and “benign” and fail to propose specific hypotheses about the underlying molecular effects of variants. A prediction that a variant is “pathogenic” is not of much clinical use without a testable prediction of the mechanisms of its pathogenicity, pleiotropic effects and possible insights to treatment.

Motivated by the challenges of variant interpretation, recent advances in experimental approaches for protein structure determination <sup>176-181</sup> and recent improvements to the accuracy of prediction, modeling and analysis of native 3D protein structural models <sup>92,182-184</sup>, I propose a new variant interpretation paradigm. This “personalized structural biology” approach focuses on making mechanistic predictions about the effects of the variant(s) observed in patients in the context of their genetic background via computational and experimental evaluation of protein structure and function. I demonstrate the power of this approach on two candidate VUS in *KCNC2*, the gene encoding the homo-tetrameric voltage gated potassium channel Kv3.2, one variant from an individual with an unsolved epilepsy-like disease enrolled in the UDN and another variant from a recent case report <sup>185</sup> with epilepsy-like phenotypes; however no functional validation was done for the variant.

Developmental epileptic encephalopathies (DEE) are a group of devastating disorders in which epileptic activity contributes to cognitive and behavior impairment in addition to underlying developmental pathologies <sup>186,187</sup>. Genetic etiologies are thought to be the cause of a

substantial proportion of these DEE cases, and with recent advances in genetic testing technology, many DEE variants have been discovered. The underlying genetic mechanisms are diverse<sup>188</sup>, but defects in neuronal ion channels are thought to be a common cause of DEE. For example, the initial discovery that Dravet syndrome is caused by variants in *SCN1A*<sup>189</sup> has been followed by the demonstration that variants in many voltage-gated potassium (Kv) channels can cause DEE<sup>190</sup>. The largest family of these channels is the Kv family, with 12 subfamilies whose alpha subunits are encoded by approximately 40 genes. The Kv3 subfamily influences rapid firing of inhibitory interneurons in the central nervous system<sup>191</sup>. The general mechanism of Kv3.2 channel gating is thought to be similar to other closely related voltage-gated potassium ion channels. Kv channels consist of four homologous subunits, with each monomer having six transmembrane helical domains (S1-S6). S1-S4 form the voltage sensing domain (VSD) and S5-S6 from all four subunits form the membrane pore. A linker domain between S4 and S5 connects the VSD to the pore forming units<sup>192-196</sup>. The VSD undergoes conformational changes between the open and closed state of the channel<sup>197</sup>, and the coupling between the S4-S5 linker and the S6 pore forming unit is responsible for the voltage dependent gating of the channel<sup>198,199</sup>. With a pronounced inward movement of the positively charged S4 voltage sensor, the S4-S5 linker is pushed downwards, which causes the S6 helix to constrict the pore, thus closing the channel<sup>195</sup>. This gating mechanism is made possible by the presence of a Proline-Valine-Proline (PVP) motif, which acts as a hinge domain in the S6 helix. The hinge domain allows the S6 pore-forming helix to kink at the flexible PVP motif to open and close the channel pore.

Other members of this subfamily have been implicated as a potential causes of DEE and other epilepsy-like symptoms with discoveries of variants in genes encoding Kv3.1 and Kv3.3<sup>200-202</sup>. More recently variants in *KCNC2*, which is highly expressed in GABAergic interneurons

in the CNS, have been suggested to be linked to DEE-like phenotypes, with possibly dominant negative effects<sup>185,203,204</sup>. However, the links and their mechanisms are yet to be established.

MD simulations provided detailed insight into the molecular mechanism underlying the altered function of both variants. Our MD results agree well with the experimentally observed loss-of-function and gain-of-function phenotypes for V469L and V471L, respectively.

## *Results*

### *Undiagnosed Diseases Network patient with DEE-like symptoms*

A child at the Vanderbilt University UDN site presented with DEE-like phenotypes, including multiple types of refractory seizure and global developmental delay. Early on, they developed generalized tonic clonic seizures, and was diagnosed with Lennox-Gastaut syndrome, a severe form of DEE. However, he continued to have frequent myoclonic absence seizures and occasional generalized tonic clonic seizure.

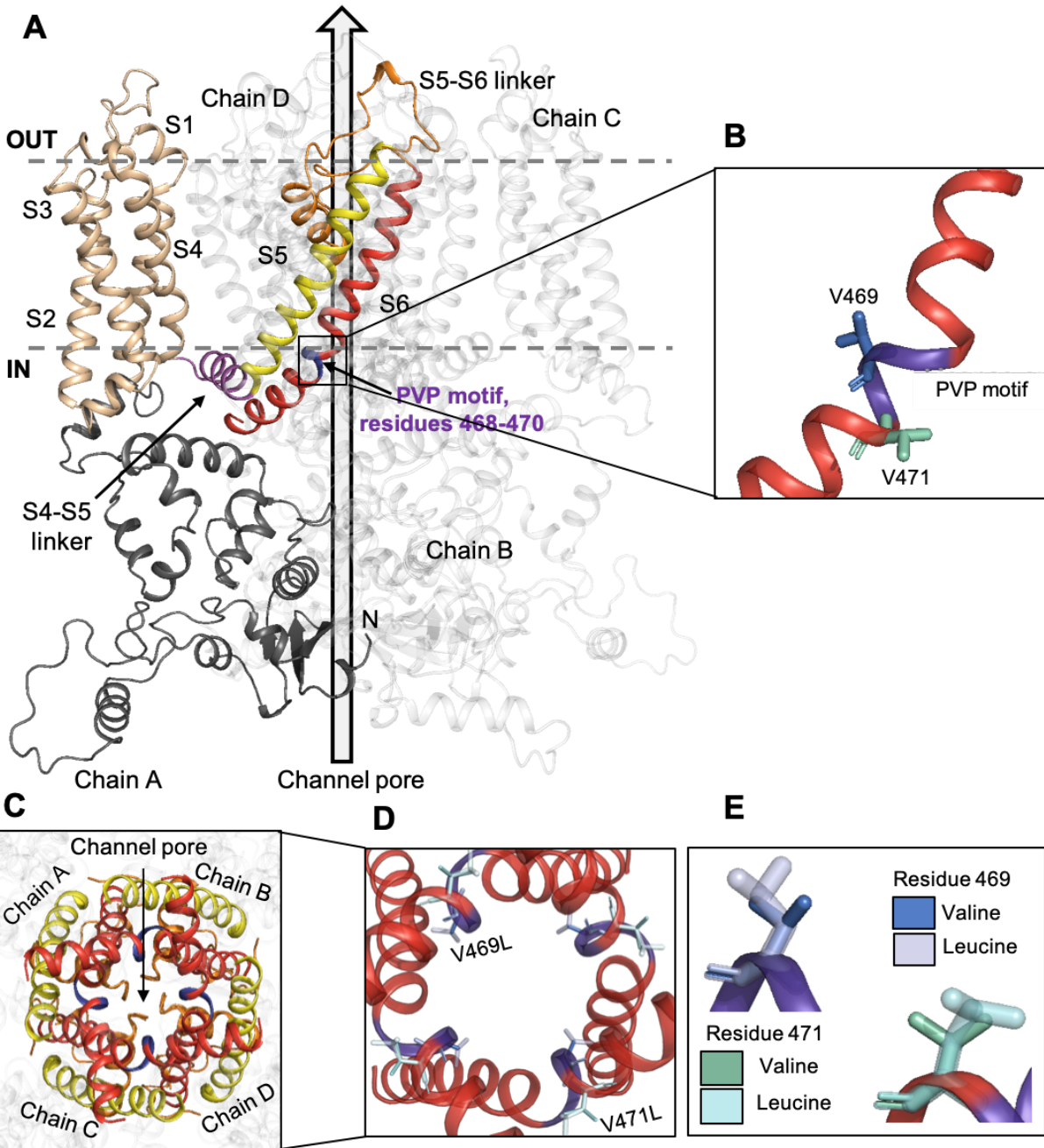
Initial sequencing of the individual on an epilepsy gene panel through Athena covered *ARHGEF9*, *ARX*, *CDKL5*, *CNTNAP2*, *FOXP1*, *GABRG2*, *GRIN2A*, *KCNT1*, *MECP2*, *NRXN1*, *PCDH19*, *PNKP*, *RNASEH2A*, *RNASEH2B*, *RNASEH2C*, *SAMHD1*, *SCN1A*, *SCN1B*, *SCN2A*, *SCN8A*, *SCN9A*, *SLC25A22*, *SLC2A1*, *SLC9AC*, *SPTAN1*, *STXBP1*, *SYNGAP1*, *TCF4*, *TREX1*, *UBE3A*, *ZEB2*. The only potentially significant finding was a heterozygous c.2985G>C variant in *GRIN2A*. Deletion analysis of *SCN1A* was negative as well. Secondary findings, metabolic screens, and mitochondrial DNA sequencing were also negative.

Following the negative epilepsy panel result, WES revealed a candidate variant in a voltage-gated potassium channel Kv3.2, *KCNC2* c.1405G>T (p.V469L). Sanger sequencing confirmed this variant. This variant was not seen in either of his parents, and therefore it was

presumed to be *de novo*. This variant is in the conserved hinge motif of the channel which is critical for channel gating (**Figure 29**). The potential relevance of this variant is supported by another recently reported discovered candidate heterozygous variant also located in the hinge domain of *KCNC2* (c.1411G>C, p.V471L) only two amino acids away from V469L variant found in the UDN subject<sup>185</sup>. The UDN subject and the previously reported case shared the phenotypes of DEE, seizures refractory to medications, developmental delay, and microcephaly. However, their phenotypes differed in that the reported case also had complete absence of speech, dystonia, decreased myelination around frontal and occipital horns of the lateral ventricles, spastic tetraplegia, myoclonic jerks, and opisthotonos attacks.

To evaluate the evidence for these VUS and propose specific functional hypotheses, I assessed the effects of these variants on protein expression, structure, and function with experimental and computational methods.





**FIGURE 29:** *Candidate pathogenic variants in KCNC2 are nearby, but have different structural contexts in Kv3.2.*

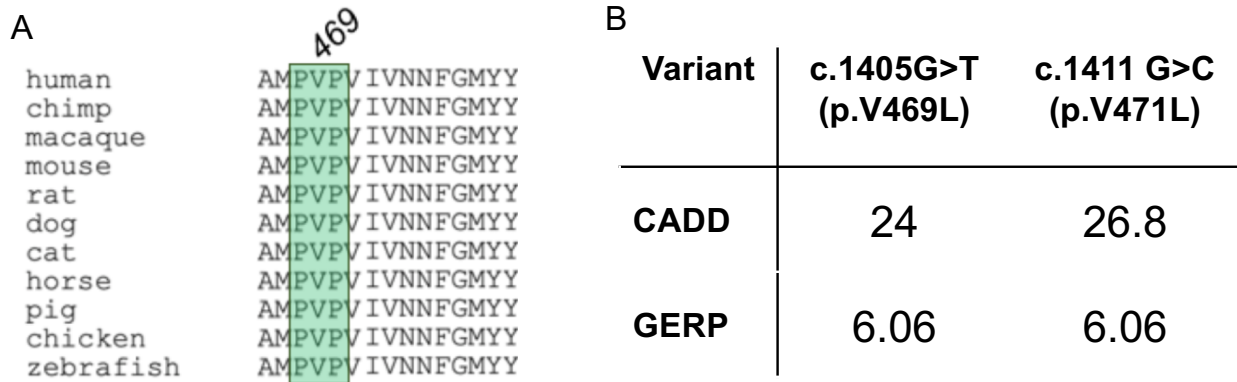
(A) Homo-tetrameric structure of Kv3.2. The complete structural model of Kv3.2 (*KCNC2*) was generated using RosettaCM from Kv1.2-Kv2.1 paddle chimera channel (PDB ID: 2R9R, 42.8% sequence identity). Four homologous subunits form the tetrameric channel pore structure; chain A is shown in color. Each monomeric subunit has intracellular N terminal domain (black) and six transmembrane helical domains. S1-S4 form the voltage sensing domain (VSD, beige). The S4-S5 linker (magenta) is the force transducer between the VSD and the channel pore, formed by the S5 (yellow) and the S6 (red) helices. The S5-S6 linker (orange) acts as the selectivity filter, allowing only potassium ions through the channel. The patient variant (V469L) is located in the PVP motif (purple; residues 468-470) which acts a hinge domain facilitating channel gating. The previously discovered variant (V471L) suspected to also cause DEE-like symptoms is located adjacent to the PVP motif. (B) A view of the carbon backbone of the S6 helix, showing the PVP hinge region. The valines at positions 469 (blue) and 471 (green) are shown. (C) A view of the channel pore formed by the tetrameric structure of Kv3.2, showing the selectivity filter (orange) and the hinge domain (purple). (D) A closer view of the channel pore with reference amino acids valine at positions 469 (deep blue) and 471 (teal) shown, alongside variant leucine residues (light blue and cyan). Residue 469 extends into the pore, while residue 471 faces away from the pore. (E) A view of the carbon backbone of the native (valine) and substituted (leucine) amino acids at positions 469 and 471.

### *Structural modeling suggests distinct functional effects for candidate KCNC2 variants*

To evaluate the potential effects of the *KCNC2* variants at the molecular level, my PSB colleague Dr. Bian Li and I constructed a homology model of its tetrameric structure based on a high-resolution structure of the Kv1.2-Kv2.1 paddle chimera channel (PDB ID: 2R9R)<sup>198</sup> using the Rosetta molecular modeling suite (Methods)<sup>87</sup>.

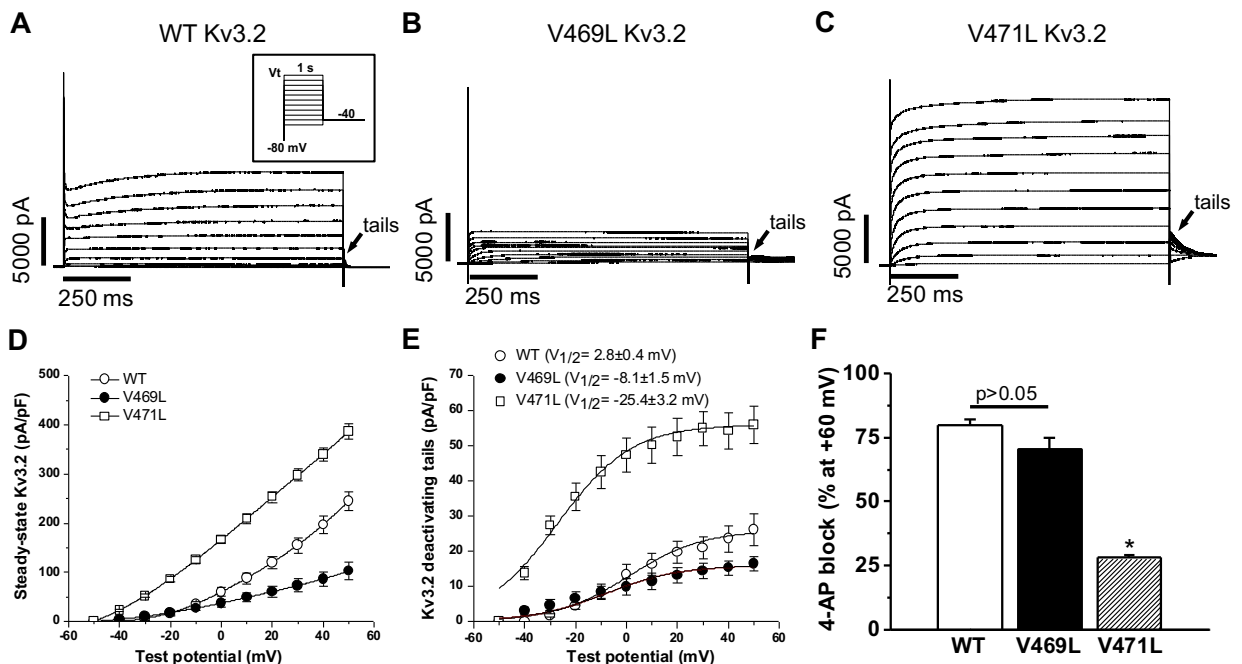
The homo-tetrameric Kv3.2 model, with six transmembrane helical domains (S1-S6) is shown in **Figure 29A**. The PVP motif ranges between residues 468 and 470 on the S6 helix and facilitates channel gating. The variants of interest p.V469L and p.V471L are adjacent to the PVP motif (**Figure 29B**). The channel pore is formed by the S5 and S6 helices of all four chains together (**Figure 29C**), and the PVP motifs on all four helices act together for channel gating. This region is almost entirely conserved among vertebrates **Figure 30A**). Furthermore, previous

studies have shown that altering the central hydrophobic residues in Kv1.1 channels from valine to isoleucine affects channel kinetics, stability, and conformational dynamics<sup>205</sup>.



**FIGURE 30:** KCNC2 variants are located in the evolutionarily conserved hinge domain and predicted to be deleterious.

(A) The amino acid sequence flanking the PVP motif (green box) and the V469L and V471L are deeply conserved across vertebrate species. (B) V469L and V471L are both predicted to be functional by the combined annotation dependent depletion (CADD) and genomic evolutionary rating profile (GERP) scores.

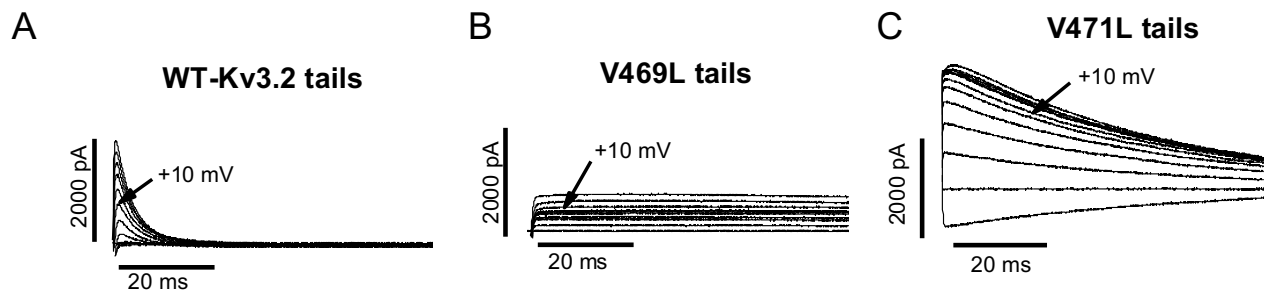


**FIGURE 31:** *Candidate Kv3.2 variants cause loss and gain of channel function.*

(A) Representative potassium current traces for wild type (WT) Kv3.2 in CHO cells recorded by voltage clamp. WT shows the characteristic current amplitude and fast deactivation (short deactivating tail currents). (B) Current traces for the V469L variant show lower peak current and very slow deactivation. (C) Current traces for the V471L variant demonstrate much higher current and moderately slowed deactivation compared to WT. Each voltage clamp used the protocol shown in the insert of panel A. The deactivation tails are compared in greater detail in Figure 32A-C. (D) Current vs. steady-state voltage (I-V) curves for WT Kv3.2, V469L, and V471L. The V469L variant showed a much lower current at steady-state, while the V471L variant showed increased current at steady-state. (E) Current vs. voltage plots for the tails for each variant. V469L has a slight negative shift (~10 mV), while V471L has a large negative voltage shift (~28 mV). Each group considered 6 to 10 cells. (F) Percentage of channel activity (steady-state current) blocked by 200  $\mu$ M 4-aminopyridine (4-AP), a known voltage gated potassium channel blocker, for WT, V469L, and V471L. The WT and V469L Kv3.2 were similarly blocked ( $p > 0.05$ ,  $n=6$  for each), but V471L was resistant to 4-AP blockage ( $p < 0.0001$ ,  $n=6$ ). Figure 33A-C shows the protocol and representative traces for each variant. Altogether, these data demonstrate loss- and gain-of-function effects on channel activity for V469L and V471L, respectively.

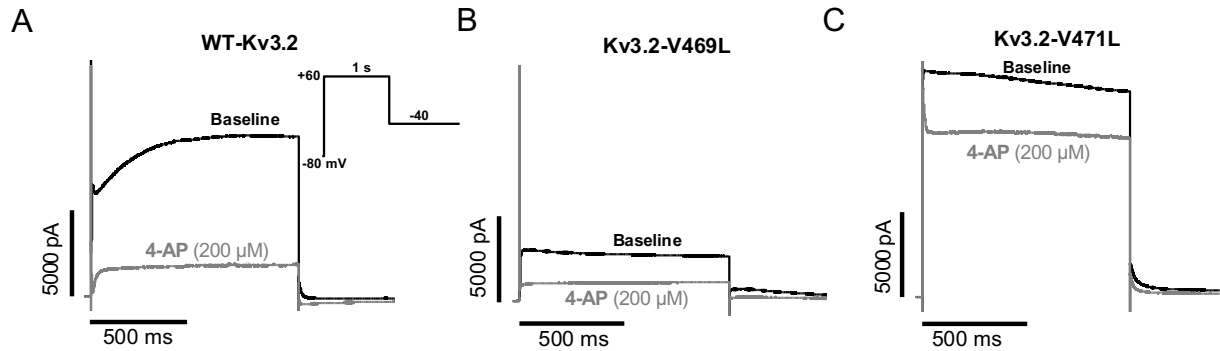
The UDN subject's variant (p.V469L) results in a conservative substitution, which changes the hydrophobic valine at the core of the PVP motif to leucine, another hydrophobic amino acid. However, the p.V469L variant is predicted to have a deleterious effect on the protein by commonly used variant effect prediction tools like CADD and GERP (**Figure 29B**). The residue at position 469 faces into the channel pore (**Figure 28D**), and while conservative amino acid substitutions do not usually have severe effects, in this case, the bulkier leucine amino acid (**Figure 28E**) at the center of the PVP motif on the hinge domain could influence ion transfer. I hypothesize that it could sterically obstruct the pore resulting in a decreased pore radius and slower kinetics of channel gating. The steric hindrance of the curving of the S6 helix could lead to fewer ions passing through the pore.

The second recently reported candidate *KCNC2* variant<sup>185</sup> (p.V471L) is immediately adjacent to the PVP motif (**Figure 29A**); however, the structural context of this variant in the model suggests potentially different effects from that of our UDN subject's p.V469L variant. Residue p.V471 faces away from the pore (**Figure 29D**), and therefore, the substitution of the bulkier amino acid leucine (**Figure 29E**) is less likely to lead to a decrease in the pore radius as the residue faces outward. In this case, I hypothesized that the molecular effect of the p.V471L variant would widen the pore and increase its tendency to remain open, leading to more ions passing through than normal and thus a gain-of-function phenotype. I also predicted that the channel gating will be affected by the bulkier leucine residue; however, not to the extent of the p.V469L variant since p.V471L faces away from the channel pore.



**FIGURE 32:** Candidate *KCNC2* variants have different effects on Kv3.2 deactivation tail kinetics.

(A) Current traces for the deactivating tails for wild type (WT) Kv3.2. The short peak indicates fast deactivation, which is a hallmark of Kv3.2, and required for fast depolarization of membrane potential in the central nervous system. (B) Current traces for the deactivating tails of V469L Kv3.2 show a much slower deactivation, indicated by the long peak which does not return to zero. (C) Current traces for the deactivating tails of V471L Kv3.2 demonstrate a much higher current and a somewhat slowed deactivation, indicated by the higher and longer peak, compared to the WT.



**FIGURE 33:** 4-AP differentially blocks variant Kv3.2 channels.

The effect of 4-aminopyridine (4-AP), a voltage gated potassium channel blocker, on WT Kv3.2 (A), V469L (B), and V471L (C) channel activity. Representative baseline (absence of 4-AP) current traces are shown in black for each protein, and representative traces with 200  $\mu$ M 4-AP are shown in gray. 4-AP blocked the WT and V469L at similar levels. However, activity of the V471L variant was only modestly reduced by 4-AP and maintained high activity. The voltage-clamp protocol is shown as insert in panel A.

*V469L causes loss of channel function, while V471L causes gain of function*

To quantify the effects of the candidate variants our experimental collaborator Dr. Tao Yang, from the Dan Roden group, quantified the electrophysiological function of Kv3.2, potassium channel currents for WT Kv3.2 and the two disease causing variants (p.V469L and p.V471L) in CHO cells. The proteins were expressed in a homo-tetrameric model, with all four chains carrying the variant, in each case. The WT form of Kv3.2 showed a very fast deactivation, in accordance with previously characterized behavior of the Kv3.2 channel (**Figure 31A, 32**). *KCNC2* is primarily expressed in the brain, where its product Kv3.2 contributes to the fast repolarization of action potentials in neurons of the central nervous system<sup>191,206,207</sup>. Therefore, short spike duration and rapid deactivation of Kv3.2 channels are important for normal physiology.

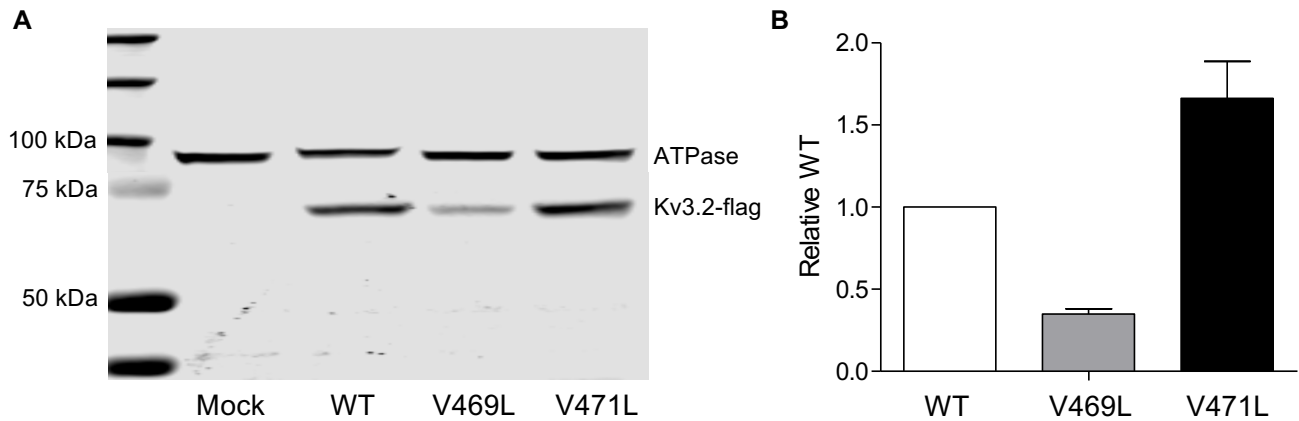
The electrophysiological profiles for the p.V469L variant (**Figure 31-32**) compared to the WT Kv3.2 had a lower peak current and much longer deactivation tails. The peak current for the

p.V469L mutant was less than half that of WT (**Figure 31D**), with a much slower deactivation and a slight negative shift of  $\sim 10\text{mV}$  (**Figure 31E**). In contrast, the electrophysiological profile for p.V471L (**Figure 31C**, **Figure 32**) showed a much higher peak current than WT, with a longer deactivation tail. The p,V471L peak current was 1.5 times that of the WT channel (**Figure 31D**), with a moderately slower deactivation, but a drastic negative shift of  $\sim 28\text{mV}$  (**Figure 31E**). The very slow deactivation and slightly negative voltage shift observed for p.V469L and moderately slow deactivation and dramatically negative voltage shift for p.V471L align with the structural hypothesis of loss-of-function and gain-of-function phenotypes, respectively. For each variant, the same amount of plasmid was injected, and the behavior of the proteins at their native levels of expression was analyzed.

Dr. Yang further characterized the effect of the variants on channel function by administration of the voltage-gated potassium channel blocker 4-aminopyridine (4-AP). Kv3.2 is very sensitive to 4-AP<sup>208,209,211</sup>. 4-AP is known to approach the channel lumen from the cytoplasmic side<sup>208,210,212</sup>; and bind the open channel weakly. Once bound, the channel becomes biased towards the closed state<sup>210</sup>, and 4-AP binds strongly to the closed conformation, blocking the channel.

Interestingly, 4-AP blocked the channel activity similarly for the WT and p.V469L Kv3.2 (**Figure 31F**, **Figure 33**). Both experienced  $>70\%$  decreases in activity ( $p > 0.05$ ,  $n=6$ ). In contrast, the gain-of-function p.V471L variant was resistant to 4-AP compared to WT and p.V469L ( $p < 0.0001$  for both,  $n=6$ ), showing a less than 30% reduction in activity (**Figure 31F**, **Figure 33**). This could be due to the V471L variant stabilizing the channel in the open conformation, thereby making 4-AP less effective in closing the channel and less likely to bind.

These results further supported the contrasting loss- and gain-of-function mechanisms I proposed for p.V469L and p.V471L.



**FIGURE 34:** Candidate KCNC2 variants modify Kv3.2 expression levels.

(A) Western blot showing expression of WT Kv3.2, V469L and V471L variants in CHO stable cells. The mock well shows only the ATPase antibody tag, while the WT, V469L and V471L labeled wells have the corresponding version of Kv3.2 loaded. Kv3.2 has a molecular weight of ~70 kDa and shows up as one band, below the 75 kDa marker, thus confirming the presence of the protein in its native state. The V471L band is larger and more intense than WT, while V469L is faint. This suggests higher protein levels for V471L and lower levels for V469L compared to WT. (B) Protein expression estimated from Western blot band intensity. The protein levels for V469L were roughly half that of the WT, and the proteins levels for the V471L were more than one and a half times that of WT. These results support a loss-of-function phenotype for V469L and a potential gain-of-function phenotype for V471L.

V469L expression is lower while V471L expression is higher than wild type

Rare pathogenic protein-coding variants, in addition to causing changes to protein structure and molecular function, can also lead to altered protein expression in cells. Dr. Ningning Hu, from the Robert Macdonald group, compared the levels of protein present in cells for the WT and two Kv3.2 variants with an Immunoblot analysis (Western Blot).

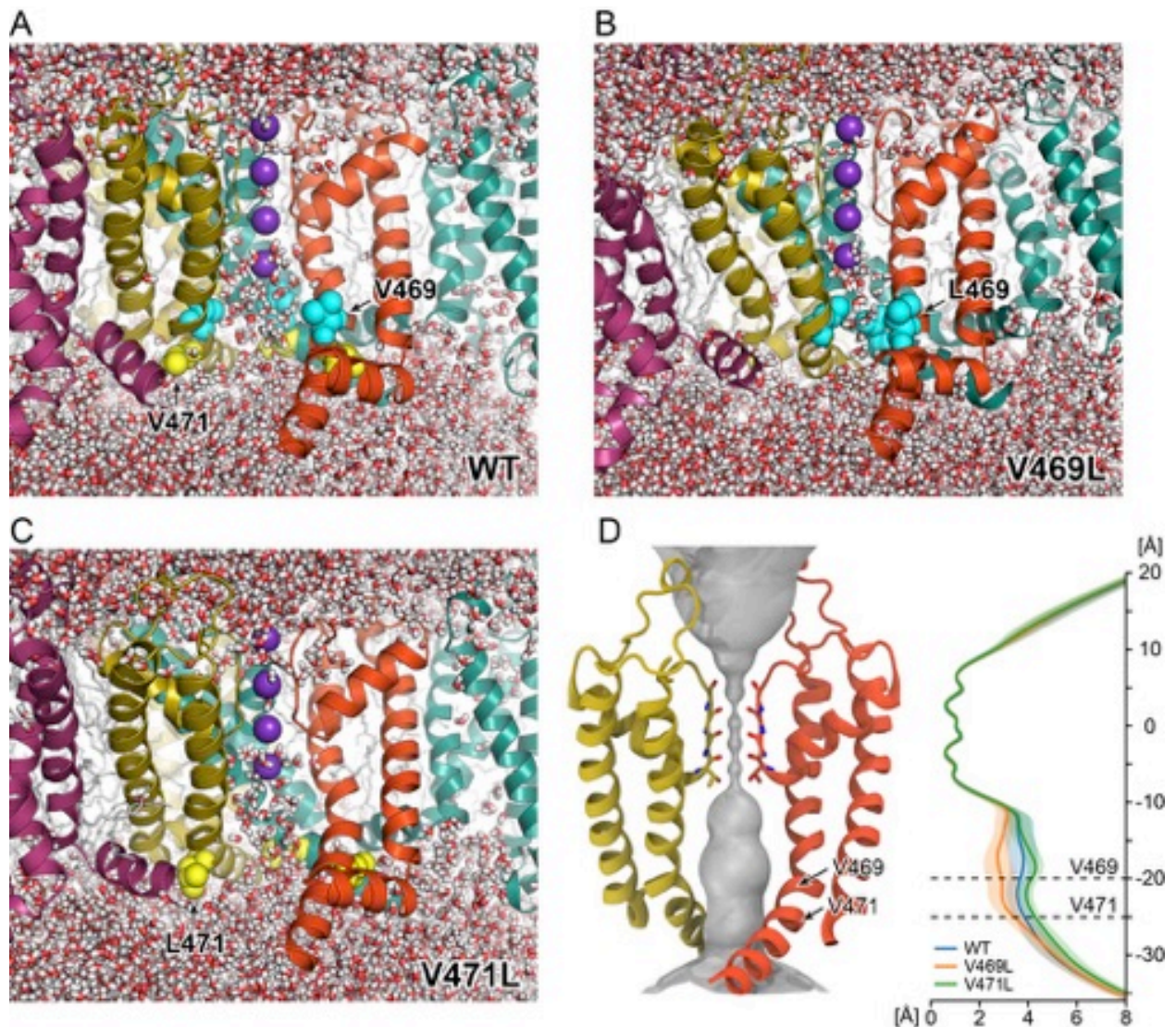
The homomeric V469L Kv3.2 was present at less than half the amount for homomeric WT, while the levels for homomeric V471L Kv3.2 were greater than one and a half times that of WT (**Figure 34**). Thus, expression differences likely contribute to the loss- and gain-of-function



effects for the two variants, respectively. However, while the expression levels could cause the observed differences in the peak currents for the two variants (**Figure 31**), differences in protein levels alone cannot explain the slowed deactivation dynamics of the V469L channel. Thus, both differences in the molecular function and expression levels of these *KCNC2* variants contribute to their phenotypic effects.

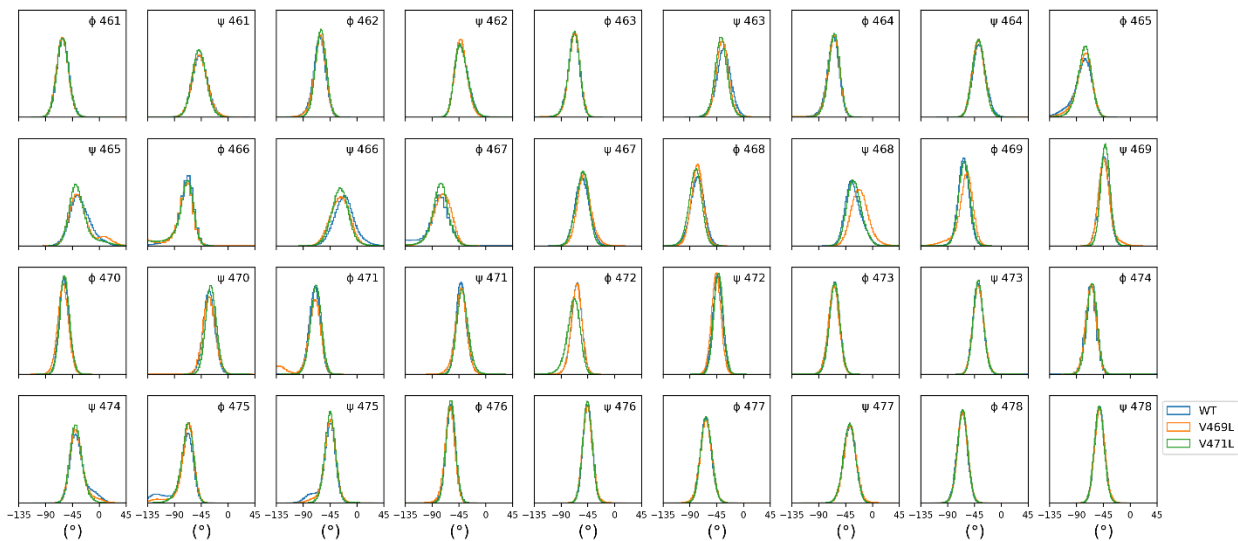
*V469L constricts the channel pore in molecular dynamics simulations increasing the energetic barrier for K<sup>+</sup> ion permeation*

To explore the molecular basis for the functional changes caused by V469L and V471L, Dr. Georg Kuenze and I performed molecular dynamics simulations of these ion channel variants and of WT Kv3.2 in POPC membranes. Each system was simulated for more than 1 $\mu$ s in total. **Figure 35A-C** displays simulation snapshots of the three ion channel systems and **Videos V1-V6** show representative MD trajectories. I observed that the inner pore helices in V469L moved closer together at their hinge motif sites such that the pore became more constricted and fewer water molecules were able to enter the inner channel cavity through the cytosolic gate. This effect was most likely driven by increased attractive interactions between leucine 469 residues on adjacent and opposite S6 helices that led to a ‘de-wetting’ of the channel pore. Calculation of the pore radius along the channel axis (z-axis) (**Figure 35D**) confirmed that the K<sup>+</sup> ion permeation pathway in the V469L channel is more constricted compared to WT and V471L Kv3.2 channels.



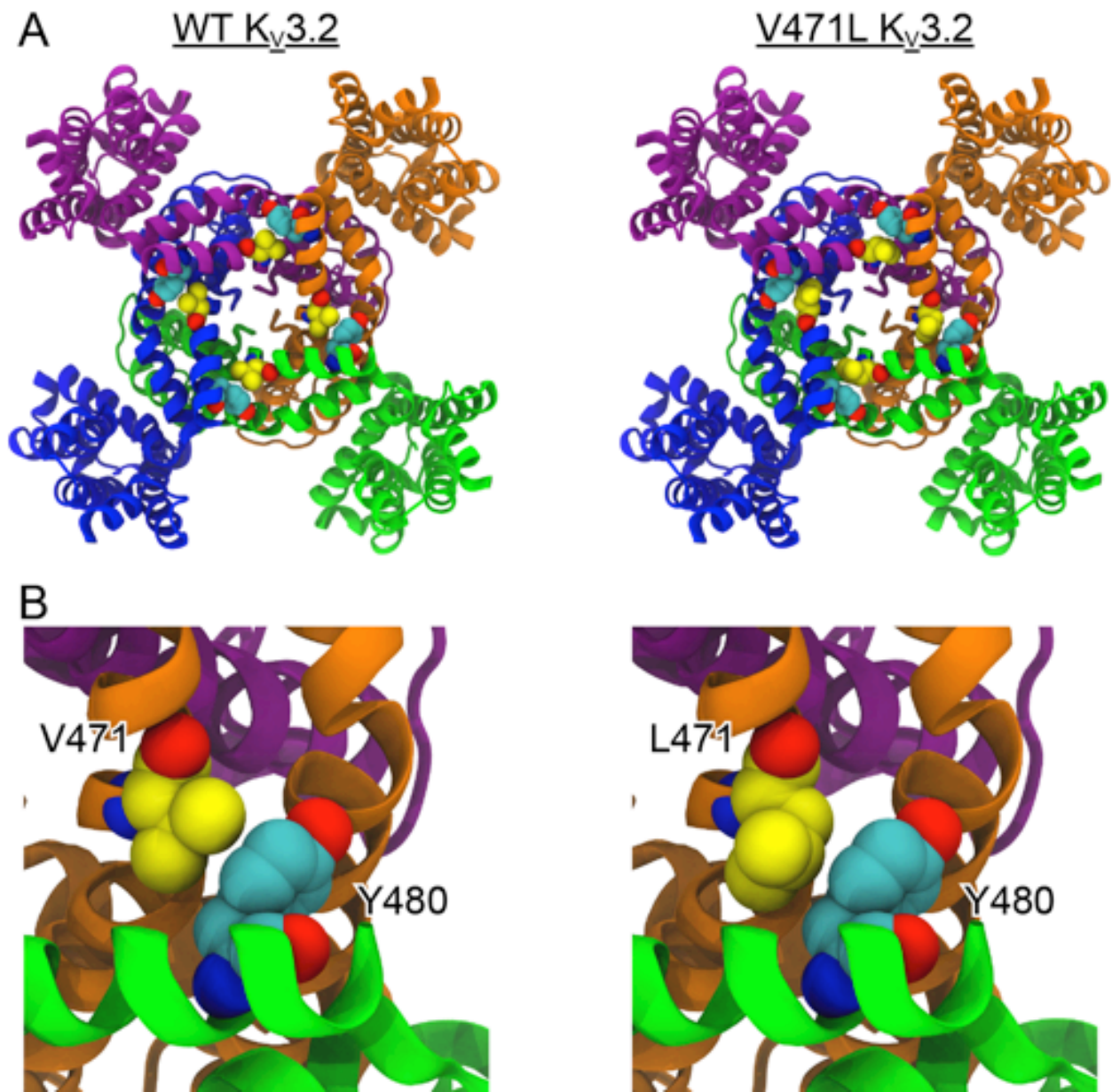
**FIGURE 35:** *The channel pore of Kv3.2 V469L becomes constricted in MD simulations whereas the pore of V471L adopts a stable open conformation.*

(A) – (C) Snapshots from MD simulations of Kv3.2 WT, V469L, and V471L. The entire MD trajectories are available at <https://vanderbilt.box.com/s/al6y4ezhmquw8il3wsvhhesdazeqmbyi>. For each protein, the channel-membrane system was simulated in four replicas with a total simulation time of more than 1 $\mu$ s. The protein is represented as ribbon with each chain shown in a different color. One domain in the front is not shown to better see the channel cavity. The amino acids at positions 469 and 471 are depicted as spheres and colored cyan and yellow, respectively. Water molecules are shown as sticks (red-white). (D) Surface representation of the pore radius of Kv3.2 WT (left) and 1D pore radius profiles (right) along the channel axis (z-axis) for WT, V469L, and V471L. The solid line and shaded area represent the average radius and standard deviation from four independent MD replicas.



**FIGURE 36:** *Distribution of  $\phi$  and  $\psi$  backbone angles of pore helix residues 461 to 478 sampled in MD simulation of Kv3.2 WT, V469L, and V471L.*

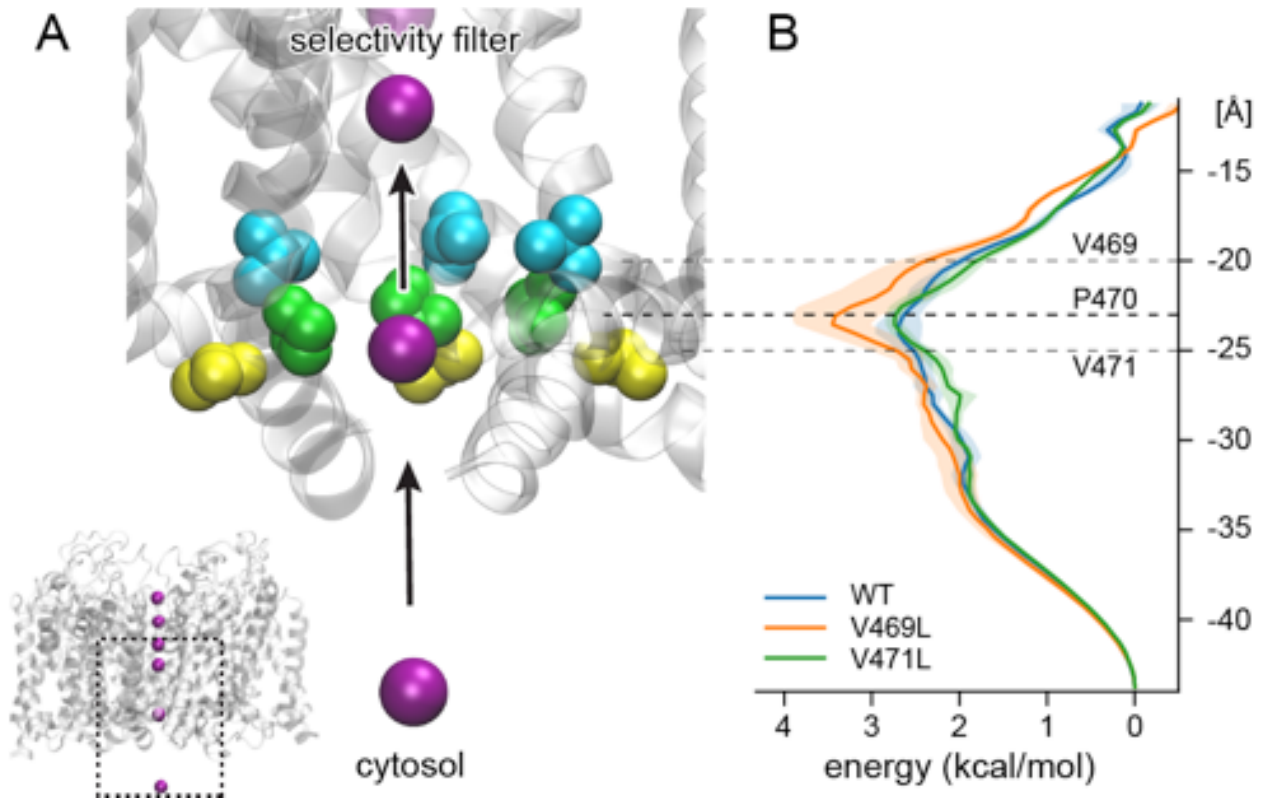
For Kv3.2 V469L variant, small changes are found at  $\psi$  466,  $\phi$  467,  $\psi$  468 and  $\phi$  469, and for Kv3.2 V471L at  $\psi$  466 (**TABLE T5**).



**FIGURE 37:** *Adjacent pore helices in Kv3.2 form a contact between residues 471 and 480.*

A) Cytosolic view of structural models of Kv3.2 WT and V471L. Each subunit is shown with a different color. V471 in WT Kv3.2 and L471 in the pV471L variant, respectively, are represented as spheres and colored yellow. Residue Y480 is shown in cyan. (B) Close-up view of residues V471 (left) or L471 (right) interacting with Y480.





**FIGURE 38:** *V469L increases the energy required for K<sup>+</sup> ion transfer through the cytosolic gate of Kv3.2 compared to WT and V471L.*

(A) I estimated the energy required to transfer a K<sup>+</sup> ion (purple sphere) through the cytosolic channel gate, from the bulk solvent into the cavity below the selectivity filter by umbrella MD simulation. A close-up view of the gate and aqueous cavity of WT Kv3.2 channel is shown. The subunit in the front is not depicted to better see the K<sup>+</sup> permeation pathway which is lined by residues on S6. The side chains of reference amino acids V469, P470, and V471 are drawn as cyan, green, and yellow spheres, respectively. (B) 1D PMF of K<sup>+</sup> transfer through the channel for Kv3.2 WT, V469L, and V471L. The solid line and shaded area represent the average PMF and standard deviation of four independent MD simulations. The V469L variant induces a greater energetic barrier to ion transfer compared to WT and V471L. The increased energetic requirement is focused on the conserved P470 residue in the hinge region. This supports the relevance of the disruption of this element to function and the functional difference between V469L and V471L in spite of their spatial proximity.

TABLE T5: Summary of  $\phi$  and  $\psi$  backbone angles of pore helix residues 463 to 475 sampled in MD simulation of Kv3.2 WT, V469L, and V471L.

Residue	Dihedral	WT median	V469L median	V71L median	WT - V469L	WT - V471L	V469L - V471L
463	$\phi$	-68.214	-67.374	-67.786	0.840	0.428	0.412
	$\psi$	-32.987	-35.295	-37.187	2.308	4.2	1.892
464	$\phi$	-65.619	-64.638	-64.060	0.981	1.559	0.578
	$\psi$	-36.610	-38.107	-37.833	1.497	1.223	0.274
465	$\phi$	-78.080	-76.485	-76.110	1.595	1.97	0.375
	$\psi$	-32.992	-34.092	-36.759	1.100	3.767	2.667
466	$\phi$	-69.493	-72.195	-71.071	2.702	1.578	1.124
	$\psi$	-21.215	-27.900	-28.040	6.685	6.825	0.140
467	$\phi$	-78.686	-72.731	-75.937	5.955	2.749	3.206
	$\psi$	-54.420	-51.163	-53.854	3.257	0.566	2.691
468	$\phi$	-76.931	-77.422	-80.585	0.491	3.654	3.163
	$\psi$	-32.838	-21.468	-30.719	11.370	2.119	9.251
469	$\phi$	-62.922	-58.835	-61.602	4.087	1.32	2.767
	$\psi$	-43.718	-43.248	-41.395	0.470	2.323	1.853
470	$\phi$	-58.551	-60.526	-57.736	1.975	0.815	1.160
	$\psi$	-31.664	-31.793	-28.719	0.129	2.945	2.816
471	$\phi$	-70.613	-71.664	-68.606	1.051	2.007	0.956
	$\psi$	-40.992	-41.250	-39.090	0.258	1.902	1.644
472	$\phi$	-63.507	-63.709	-68.455	0.202	4.948	4.746

	$\psi$	-43.176	-45.512	-41.865	2.336	1.311	1.025
473	$\phi$	-63.365	-64.722	-63.215	1.357	0.15	1.207
	$\psi$	-38.527	-38.229	-38.088	0.298	0.439	0.141
474	$\phi$	-65.186	-65.446	-63.958	0.260	1.228	0.968
	$\psi$	-36.940	-38.138	-38.796	1.198	1.856	0.658
475	$\phi$	-70.334	-67.447	-67.946	2.887	2.388	0.499
	$\psi$	-46.164	-45.022	-44.724	1.142	1.44	0.298

For Kv3.2 V469L variant, small changes are found at  $\psi$  and  $\phi$  angles for several residues. The alterations for  $\psi$  466 were consistent between both the variants, indicating a strain on the S6 helix resulting from both variants. However, the difference between V469L and WT for  $\psi$  468 and  $\phi$  469 were specific for V469L variant, and not observed for the V471L variant. Conversely, the alterations to  $\psi$  471 and  $\phi$  472 were observed specifically for the V471L variant, indicating these changes were specific for the position and resulting amino acid of the variants.

Furthermore, I noticed small but distinct differences of the backbone structure at residue 469 and preceding residues (**Figure 36, Table T5**). By contrast, the pore radius of the V471L channel was slightly wider than that of WT Kv3.2 indicating that V471L adopted a stable open conformation in MD. One possible explanation for this observation is the difference in the types of interactions made by residues at positions 469 and 471. While L469 side chains are oriented towards each other and towards the pore, L471 residues are oriented outward and interact with residues on S5 in the same subunit and with residues on S6 in a neighboring subunit. The largest number of atom contacts of L471 are made with Y480 on an adjacent S6 helix (**Figure 37**). In the variant, the number of heteroatom contacts (within 4Å radius) for the L471-Y480 interaction more than doubled relative to the V471-Y480 interaction in WT Kv3.2 (average of ~3.2 contacts in WT Kv3.2 to ~7.0 contacts in the V471L channel). This finding offers a plausible explanation

for how this amino acid change at position 471 leads to stabilization of the open channel conformation.

To assess the energetic cost more directly for K<sup>+</sup> ion permeation in WT Kv3.2 and both channel variants, I used umbrella MD simulations and calculated the potential of mean force (PMF) for moving a K<sup>+</sup> ion from the cytosolic site of the channel through the S6 helix gate into the water-filled cavity below the selectivity filter (**Figure 38A**). Compared to WT, the energetic barrier for ion transfer of the V469L variant increased by ~0.8 kcal/mol (**Figure 38B**). The V471L variant, however, required an energy for ion transfer similar to WT. The highest peak in the PMF and the V469L-specific energy increase occurred at position P470. This indicates that V469L, but not V471L, constricts the channel pore at the PVP hinge region, which is in line with our pore radius measurements.

### *Discussion*

Our team has illustrated the power of a “personalized structural biology” pipeline that places candidate VUS into 3D structural models tailored to the patient. The integration of cycles of computational and experimental analysis enabled us to provide mechanistic molecular insights into the different mechanisms by which two proximal candidate *KCNC2* VUSs lead to DEE-like phenotypes. DEE has been previously linked to dysfunction of other ion channels<sup>213,215</sup>. Moreover, Kv3 channel family members have been previously associated with neurological disorders such as ataxias, epilepsies, schizophrenia, and Alzheimer’s disease<sup>214</sup>.

The V469L variant occupied the central hydrophobic residue of the PVP motif and the flexibility of this hinge region is critical for channel opening and closing kinetics<sup>216</sup>. Previous studies in other channels supported this hypothesis, as altering the central hydrophobic residues



in Kv1.1 channels from valine to isoleucine, a constitutional isomer of leucine, affected channel kinetics, stability, and conformational dynamics<sup>205</sup>. The V469L variant resulted in > 50% decrease in peak current and a very slow deactivation with a slight negative shift (~10mV). Interestingly, the V469L variant caused the Kv3.2 channel to be expressed at < 50% of WT. Molecular dynamics simulations showed that the V469L variant had a smaller pore radius and a higher energetic barrier to ion transfer. The constriction was likely the result of hydrophobic interactions between bulkier L469 residues causing part of the channel pore to become devoid of water molecules<sup>217,218</sup>.

The V471 variant is immediately adjacent to the PVP hinge motif and resulted in a > 50% increase in the peak current, with moderately slow deactivation, but a drastic negative shift (~28mV). There was also an increase in protein expression to > 150% of WT and the pore remained fully open in MD and the pore radius for the V471L variant of Kv3.2 channel was slightly wider. An increased number of inter-subunit contacts made by V471L suggests a possible mechanism how this mutant could selectively stabilize the open channel state. The V469L and V471L variants had opposite loss-of-function and gain-of-function effects respectively.

The lower current for V469L versus higher for V471L compared to WT could be explained by the differences in protein expression levels. However, alterations in the protein level would not cause changes in the kinetics of deactivation. The two variants could affect the energy required for the protein to undergo conformational changes between the open and closed states. The adjacency of the variants to the PVP motif would lead to changes to the ability of the helix to kink at the hinge domain and facilitate channel gating. The V469L variant, which results in the central hydrophobic valine of the PVP motif to be substituted by a bulkier leucine

extending into the pore, results in a lower tendency for the helix to kink, and therefore, leads to slower channel gating and slower deactivation. In contrast, the stabilization of the V471L channel in the “open” conformation, which would also affect channel gating, is consistent with its moderately slow deactivation. Thus, these results indicate that the variants influence both expression and channel function. However, they do not identify the cause of the differences in the expression levels of the two variants. It is likely that these result from differences in protein folding or stability. Further studies, such as analysis of protein trafficking in cells, are needed to identify the causes. For experimental simplicity, we carried out *in vitro* and *in silico* analyses with all four chains carrying the variant of interest. In the future, it would be valuable to evaluate the spectrum of effects for channels carrying different combinations of variant and WT chains; though, similar effects can be anticipated.

The ability of the channel blocker 4-AP to inhibit the V469L Kv3.2 channel aligned with the loss-of-function phenotype because 4-AP approaches the channel lumen from the cytoplasm<sup>208,210,212</sup>; therefore the steric hinderance of the pore cavity by V469L should not affect its mechanism of action. Furthermore, the decreased ability of 4-AP to block the V471L channel supports the gain-of-function hypothesis. Channel closing is destabilized in V471L, so 4-AP may not bind as efficiently to the open channel. These results also illustrate how the personalized structural biology approach can help evaluate the effects of possible pharmacological interventions. For example, 4-AP is not likely to help individuals with the V469L variant, it is possible that it could counteract some effects of the V471L variant.

Taken together, the clinical features of the UDN patient, and the reported and the combined experimental and molecular modeling of their *de novo* KCNC2 variants prove their functional role as a cause of DEE. The phenotypes associated with the two variants in *KCNC2*

are the result of two fundamentally different molecular mechanisms, even though the residues are only two amino acids away. Analyses of these variants in their structural context was key to revealing their mechanistic and functional heterogeneity. In addition to the contribution to diagnosis, these results also suggest that drugs that modulate the activity of Kv3.2 could be potential treatments. This case study demonstrates the strength of personalized structural biology as a diagnostic method to predict precise molecular hypotheses by taking the context of the variant of interest in the 3D structure of the protein into account.

## CHAPTER 5: Conclusion

Using machine learning to develop a classifier for digenic gene pairs has yielded success. Nonetheless, there is still much to learn about the mechanisms underlying digenic and other rare diseases. Other machine learning approaches and integrating additional features could further improve performance. For example, I have used Gene Ontology functional annotation enrichment as a way of categorizing the most confident digenic predictions, but GO ontology relationships between the genes would likely help prioritize potential digenic interactions. Since protein-protein interactions (PPI) were an indicative feature for DiGePred, protein family and domain similarity, derived from the Pfam<sup>219</sup> database, could be considered as a relevant feature as well. I used a Random Forest model as it suited the ensemble approach based on many features on disparate scales and limited training data. Alternatively, a support vector machine (SVM) or linear regression approaches could be used with feature normalization. As discussed in the next paragraph, I also believe that approaches that incorporate genetic variants into the prediction are promising; however, the small amount of available training data pose challenges. As more digenic disease pairs are identified, I anticipate that better predictive models will be developed and that these models will yield insight into the genes, pathways, evolutionary histories, and phenotypes associated with digenic disease.

The approach used to design DiGePred could be expanded to consider oligogenic combinations of greater than two genes. Trigenic and oligogenic cases are beginning to be identified<sup>220,222</sup>, and previous work has identified exclusive gene hubs that cause disease in combination<sup>221,223</sup>. In fact, many previously considered monogenic diseases are now being classified as oligogenic or multigenic, with a range of phenotypes depending upon which genes and how many carry variants<sup>224,225</sup>. I also believe that there is the potential to integrate

information from large-scale screens of genetic and synthetic lethal interactions in human cell lines and model organisms<sup>226–231</sup>.

This approach intentionally separates the prediction of variants' effects on gene function from the identification of gene pairs that could cause disease when their functions are disrupted simultaneously. The focus on gene pairs is reflected in this use of gene level and gene-pair level systems biology, biological network, and evolutionary features that represent genes as a whole. The question of whether a variant affects gene function has been studied extensively. There are many methods for interpreting variants of unknown significance,<sup>232–238</sup> but there is low concordance between them<sup>239,240</sup>. The decoupling of these tasks enables users to apply the approaches they believe to be most appropriate to identify gene pairs of interest before screening for digenic disease potential.

In the future, it may be beneficial to incorporate variant-level and gene-level information into a single algorithm, in particular in cases where there is structural information about the proteins of interest. Indeed, our team has had success incorporating 3D modeling of variants and their interactions with the UDN. Deriving actionable information for patient diagnosis and treatment from clinical sequencing data is a fundamental challenge in genetics and medicine. Current methods for analyzing sequencing data often fail due to the inability to predict the effects of the detected VUS on protein function.

The Personalized Structural Biology (PSB) approach to interpreting *de novo* potassium ion channel variants and attempting to uncover disease mechanisms looks to make four main contributions. First, I demonstrated via expression and electrophysiology analyses that the two candidate *KCNC2* variants (p.V469L, p.V471L) have loss-of-function and gain-of-function effects, respectively, despite both affecting the essential hinge region of Kv3.2 responsible for

channel gating. Second, the protein structural modeling and molecular dynamics simulations rationalized the mechanistic basis for the phenotypic heterogeneity of these variants. Third, the results combined to validate links between *KCNC2* variants and heterogenous DEE phenotypes. Finally, the PSB analyses provide a blueprint for integrating genetics, expression analysis, electrophysiology, and protein structural modeling to develop mechanistic understanding of the molecular effects of *de novo* variants in rare disease.

I have not simulated the entire dynamics involved in channel activation and deactivation. This process is too long to be studied by conventional MD methods. Enhanced MD protocols, which aim at representing the free energy landscape of the molecular system by a set of low-dimensional collective variables, have been used to simulate conformational changes for some ion channel systems<sup>241</sup>. These methods could be helpful for explaining the observed changes in activation potential and deactivation time. However, no general protocol for deriving a set of collective variables that capture the whole activation and deactivation cycle of Kv channels like Kv3.2 is available yet. Going forward, this approach has broad applicability across VUS observed in studies from rare disease to cancer.

MD simulations can elucidate the mechanisms of protein folding and predict the effects of variants to protein dynamics and stability. However, these are computationally generated results and not experimentally validated. Thus, we collaborated experimentally extensively to determine the functional consequences of missense variants in the potassium ion channel. There is very little experimental evidence for the functional impact of single nucleotide variants or single amino acid missense variants in human proteins<sup>104,105</sup>. To improve the scope of precision and personalized medicine, deep mutational scanning (DMS) analyses have been performed that record a sequence library with all possible variants, or combination of variants and the functional

impact score, determined by the effect of the variant on the organism fitness.<sup>106</sup> These experiments have been historically conducted in single cellular systems, but are being expanded to more complex organism and tissue systems.<sup>106-109</sup> Deep mutational scanning (DMS) analyses derive information from saturation mutagenesis and high-throughput experimental functional analyses, leading to a framework conveying data for multiple mutations simultaneously.<sup>110</sup> DMS studies have the potential for rare disease causing variant prioritization and identification.

Recently DMS datasets are being applied to help design computational variant interpretation methods.<sup>106,108,111,112</sup> The experimental mutation level data serves as an informative independent training set that has been demonstrated to have high predictive power and facilitate development of more accurate variant effect prediction tools, which also perform better on unseen variants.<sup>107,113,114</sup> I am currently using DMS data derived on human proteins to develop a tool to predict the functional effect of rare variants using computational exhaustive mutagenesis in Rosetta<sup>87,115</sup> and FoldX.<sup>89</sup> This tool will help integrate experimental variant level data into a computational variant interpretation tool.

## CHAPTER 6: Methods

### *Digenic gene pairs*

I obtained known digenic disease gene pairs from the Digenic Diseases Database (DIDA; using the latest version as of April 2021, which was updated in July 2017) <sup>39</sup>. There were 140 unique gene pairs in DIDA. These pairs served as the “positive” training data for the machine learning classifier and were termed the *digenic* set of gene pairs. DIDA provides information about the genes mutated together in cases of digenic disease, the variants in the genes, the number of variants on both alleles, as well as information concerning the connectivity of the genes forming a gene pair such as distance on PPI network, whether expressed in same tissue, whether members of the same biochemical pathway, and whether annotated to have the same function. The additional list of digenic pairs discussed in a follow up paper by the group that produced DIDA <sup>42</sup> were not used for training.

### *Non-digenic gene pairs*

I generated several sets of putative non-digenic gene pairs that served as the “negative” data in training different classifiers. The *unaffected* non-digenic set was created from genes with variants in the sequenced exomes or genomes of relatives of UDN patients deemed unaffected by the UDN. Thus, I consider any combination of genes observed to be mutated simultaneously in any one “unaffected” individual to be non-digenic. Combining gene pairs from 55 individuals, the unaffected set contains 1.8 million gene pairs. I considered validation sets both with and without gene-level overlap with the training/validation sets. The *random* non-digenic set was created by selecting random pairs from the list of all human genes. The *permuted* non-digenic set was created by generating all possible pairs of two genes from the DIDA genes excluding actual



DIDA pairs; this resulted in 13,390 permuted gene pairs. I created the *matched* non-digenic gene pair set from the random gene pairs by selecting gene pairs such that the distribution of the six NFFs match those of the digenic set. The digenic gene pairs were binned by dividing the distribution of features into equal sized intervals, such that every feature value data interval had an equal number of gene pairs. I selected random gene pairs for the matched set such that the distributions of feature values for all the selected pairs recapitulated the overall distribution for all features of the digenic set, simultaneously.

### *Six Network and Functional Features*

#### *Pathway similarity*

The pathway annotations for the genes were derived from KEGG <sup>117</sup> and Reactome <sup>118</sup>. The Jaccard similarity metric <sup>116</sup> was used to calculate the proportion of pathway overlap between the two genes. The Jaccard similarity is measured by the ratio between the intersection of two sets and the union of two sets. In this case, the pathway similarity was calculated by taking the ratio of pathways annotations in common with both genes and pathway annotations associated with either. If both genes did not have pathway annotation, the similarity value was 0.

#### *Phenotype similarity*

The phenotype annotations from the Human Phenotype Ontology (HPO) <sup>120</sup> for the genes were used as features. The phenotypic overlap between the two genes was calculated similarly, as above, using the Jaccard similarity metric. The value for missing phenotype annotations was 0.

### Co-expression

The co-expression data was derived from the COXPRESdb web server version 7.3<sup>121</sup>. The data is in the form of a mutual co-expression rank, which indicated how likely it was for a pair of genes to be co-expressed in the same tissue and the same level compared to other gene pairs. A lower rank indicated high co-expression. The inverse of the rank was used as the feature and if either gene was not found in the co-expression database, the value used was 0. The network data was downloaded from the UCSC gene and pathway interaction browser<sup>122</sup>, which in turn was derived from other sources of data, such as protein-protein interaction (PPI) databases<sup>242–245</sup>, functional annotation databases<sup>246</sup> and others.

### PPI distance

The PPI network was based on experimental data regarding protein interactions. The inverse of the shortest path between a pair of genes on this network was used as the PPI distance feature.

### Pathway distance

The pathways interaction network was based on interactions between the various curated biochemical pathways. The inverse of the shortest path between a pair of genes on this network was used as the pathway distance feature.

### Literature distance

The literature mined interaction network was made up of interactions derived from reported interactions or predicted associations in published biomedical literature. The inverse of the shortest path between a pair of genes on this network was used as the literature distance feature.

For each network (PPI, Pathway, and Literature), a value of 0 indicates the absence of a path between the gene pair in the network and was thus assigned to pairs with missing data.

### *Five Evolutionary Features*

#### *Evolutionary age*

I obtained the evolutionary ages of the proteins coded by the genes using ProteinHistorian<sup>247</sup>. This estimates the ancestral branch on which the gene first appeared and the age in millions of years. The quadratic mean of the values for each gene in a pair was used as a combined feature.

#### *Gene essentiality*

The gene essentiality scores provide a rank of how important and vital a gene is for normal physiology, viability and survival. They were derived from the OGEE webserver<sup>248,250</sup>. The essentiality scores are based on knockout (KO) experiments in model organisms and cell-based assays. The quadratic mean of the values for each gene was used as a combined feature.

#### *Loss of function intolerance (pLI)*

I added the loss of function intolerance (pLI) scores<sup>249</sup>, obtained from the EXAC consortium. These scores were based on the difference between actual mutation incidence and expected mutation frequency. A depletion of mutation incidence, compared to expected frequency, could mean the inability of the organism to survive if the gene was mutated. The quadratic mean was used as a combined feature.

### Selection pressure (dN/dS)

I used measures of selection pressure in the form of dN/dS scores for the genes. These were derived from the EVOLA web server <sup>252</sup>. dN/dS ratios give a measure of the ratio between the non-synonymous mutations and synonymous mutations during evolution. This ratio tells us whether the gene has been evolving under strong positive, negative or neutral selection. The quadratic mean was used as a combined feature.

### Haploinsufficiency

I used the Haploinsufficiency scores <sup>251</sup> which were in the form of predictions of which genes were haploinsufficient, based on observed mutations. The quadratic mean was used as a combined feature.

### Gene-focused network and functional features

Several additional gene-level attributes in the network and functional data sources described above were used as features.

### Number of pathways

The feature used for the classifier was the quadratic mean of the number of pathways associated with gene A and the number of pathways associated with gene B.

### Number of phenotypes

Similar to the pathways, the feature used for the classifier was the quadratic mean of number of phenotypes associated with gene A and with gene B, individually.

### Network neighbors

The numbers of shared network neighbors, defined as the number of genes directly connected to both gene A and B, were also considered. For each gene pair, I computed the quadratic mean of the number of genes in the network directly connected to gene A and to gene B. These features were defined for all three types of interaction networks.

### Number co-expressed

The number of genes highly co-expressed with both gene A and gene B were identified as the top 500 co-expressed genes (out of possible 20,000) for each. The feature used in the classifier was the quadratic mean number of genes highly co-expressed with gene A and gene B, individually.

### Encoding gene level features

Several of the evolutionary, genomic, and network features are attributes of individual genes rather than gene pairs. I combined these gene-level features into a single feature for each gene pair by computing their quadratic mean. Results were similar when using the arithmetic mean (Figure S5).

### Performance Quantification

Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves were computed to evaluate the performance of the classifiers. The ROC curve plots the False Positive Rate (FPR) on the x-axis and the True Positive Rate (TPR) on the y-axis. The area under each curve (AUC) was used to summarize performance.

### *Training and Testing the DiGePred Random Forest Models*

I trained several random forest (RF) classifiers to distinguish digenic and non-digenic gene pairs. I selected RFs because they can integrate diverse features, perform well on unbalanced positive and negative sets, and provide interpretable models. The sci-kit learn (sklearn) python module was used for all training, evaluation, and prediction<sup>253</sup>. Hyper-parameters were selected by nested cross validation on 80% of the labeled gene pairs. A stratified shuffle split was used for 10-fold cross validation. This method involved splitting the data into 10 equal parts, with each part of the data containing approximately the same ratio of positives and negatives as the other parts. The optimum number of trees was found to be 500 and the maximum depth was found to be 15. Based on these analyses, I selected the classifier trained with the unaffected negative pairs and all features as the best model, and I refer to this as the DiGePred classifier.

The remaining 20% of the combined labeled data was held out for final performance validation of this best model from the cross-validation. These pairs had not been previously evaluated by the classifier. I also considered held-out test sets that had no overlap with the training/validation sets at the gene level (“no gene overlap” classifiers). In addition to the held-out positive digenic pairs, I generated 100 sets of held-out non-digenic pairs for evaluation. This enabled us to evaluate the best classifier 100 fold, with the same positive digenic pairs used in every iteration, but a unique non-overlapping set of held-out non-digenic pairs in every iteration.

### *Evaluation using additional digenic pairs not in DIDA*

The classifier was further evaluated using an external set, made up of gene pairs considered to be digenic that were reported after DIDA was compiled. The external evaluation set was used in the

previously published variant combination pathogenicity predictor (VarCOPP/ ORVAL) <sup>43,45</sup>. This set had three unique gene pairs, which did not overlap with DIDA pairs. These gene pairs (AHI1, CEP290), (CEP290, CRB1) and (CEP290, RPE65) was labeled Papadimitrou et al., 19 validation set. I included recently discovered novel digenic inheritance of profound non-syndromic hearing impairment caused by (PCDH15, USH1G) <sup>126</sup>. In addition, three recently reported cases of digenic inheritance in immune disorders were used. Ameratunga et al., 17 identified epistatic interactions between TACI and TCF3 (or TNFRSF13B) resulting in severe primary immunodeficiency disorder and systemic lupus erythematosus <sup>127</sup>. Hoyos-Bachiloglu et al., 17 discussed how human immunodeficiency was caused by mutations in IFNAR1 and IFNGR2 <sup>128</sup>. More recent digenic findings such as (LAMA4, MYH7) linked to infantile dilated cardiomyopathy (Abdallah et al., 2019) from Abdallah et al., 19; (KCNE2, KCNH2) linked to long QT syndrome types 2 and 6 <sup>135</sup> from Heida et al., 2019; (CLCNKB, SLC12A3) linked to Gitelman syndrome <sup>133</sup> from Kong et al., 2019; (CACNA1C, SCN5A) linked to Long QT phenotype <sup>131</sup> from Nieto-Marín et al., 2019; (FGFR1, KLB) linked to insulin resistance <sup>132</sup> and diabetes from Stone et al., 2019; (CLCNKA, CLCNKB) linked to Bartter syndrome with sensorineural deafness <sup>129</sup> from Nozu et al., 2008; and (CLCN7, TCIRG1) linked to osteoporosis <sup>134</sup> from Yang et al., 2018 were used to assess the classifier as well.

I also included gene pairs not characterized as digenic, but displaying functional synergy associated with disease or adverse phenotypes. I derived the gene pair from the previously reported UDN study that found mutations in TRPS1 and FBN1 to be responsible for the patient phenotype and it was labeled *Zastrow et al., 17 (UDN)* <sup>136</sup>.

### *Feature Importance*

To identify the most important features I used the classifier feature importance function in scikit-learn, which uses the GINI impurity approach to quantify the relative feature importance for all features. Owing to possible biases in the GINI-based approaches when diverse features are considered,<sup>125</sup> I also used a permutation approach to calculate feature importance. This involved scrambling the feature values and comparing the error in classification between using the actual and permuted values for each individual feature<sup>255</sup>.

### *Prediction Score Thresholds*

I determined a digenic score threshold for the DiGePred classifier for classifying gene pairs digenic based on the  $F_{0.5}$  metric. This is a modification of the  $F_1$  statistic, designed to attenuate the effect of false negatives. It is calculated as  $F_{\beta} = (1 + \beta^2) \times TP / (1 + \beta^2 \times TP + \beta^2 \times FP + FP)$ , where  $\beta=0.5$ , TP=true positives, FP=false positives. The score that yielded the highest  $F_{0.5}$  value was 0.534.

### *Estimating the False Positive Rate at various score thresholds*

I evaluated the DiGePred classifier with an external set of non-digenic gene pairs as well. These gene pairs were obtained from 38 unaffected relatives of UDN patients. The genes were preliminarily selected if the variant in the gene had an ExAC<sup>256,257</sup> minor allele frequency of < 1%. A gene was further selected if it received a pathogenicity score of 'D' ("*probably damaging*") from Polyphen2 (Kircher et al., 2014) Only genes passing this Polyphen2 filter were selected to limit the predictions to pairs of genes with variants that likely affected molecular function.



Additionally, genes with rare variants were selected based on a consensus pathogenicity approach if at least two out of Polyphen2, SIFT<sup>258,259</sup>, CADD (Kircher et al., 2014; Rentzsch et al., 2019) and PhyloP<sup>261</sup> agreed that the variant(s) in the gene was pathogenic. A Polyphen2 selection criteria was similar to above. A variant was deemed pathogenic by SIFT if the score was  $\leq 0.05$ . a CADD score  $\geq 30$  was considered pathogenic, while a PhyloP score of  $\leq -10$  for a variant deemed it pathogenic. All possible gene pairs were used as the consensus pathogenic gene pairs for an individual.

The fraction of gene pairs predicted to be digenic was compared for individuals with undiagnosed disease vs. unaffected members of UDN cohorts. The comparison of these fractions was done for the most confident DiGePred thresholds ( $F_{0.5}$  and higher), with the MWU P-value being calculated for each and every threshold.

#### *Comparison with ORVAL*

I submitted the list of gene pairs for all the unaffected individuals to the ORVAL<sup>43,45</sup> server. I compared the number of pairs predicted to be digenic by ORVAL, according to its highest confidence threshold, to the number predicted by our method to be digenic at the  $F_{0.5}$  threshold. I obtained the list of genes for each unaffected individual as mentioned in the previous section. I evaluated the statistical significance of the number of digenic pairs predicted as false positives per individual between DiGePred and ORVAL using a MWU test.

Furthermore, 20% of all genes with rare variants in the individual were chosen at random. All possible gene pairs were generated to constitute the random set of gene pairs for each individual. I calculated the number of digenic pairs predicted per individual at different score thresholds. This was done to compare the number of false positives between ORVAL and

DiGePred fairly. As ORVAL includes variant effects as a feature, selecting for genes with variants that were predicted pathogenic by Polyphen2 or by a consensus of several predictors of variant effect could bias against ORVAL, though it reflects common clinical practice. Therefore, I also compared DiGePred and ORVAL on pairs of genes selected at random.

For the purpose of comparing ORVAL predictions on individuals with undiagnosed disease and unaffected members of UDN cohorts, I further ranked ORVAL predictions using the ORVAL classification score as a prediction threshold. According to the authors, a pairs with a classification score of  $> 0.74$  with a support score of 100 were scored in the 99% confidence zone. I compared the fraction of gene pairs predicted to be digenic at varying ORVAL classification score thresholds, ranging from 0.74 and higher, for diseased vs. unaffected individuals. The Mann-Whitney U test P-value was calculated for the distributions at each and every threshold.

#### *Gene ontology (GO) enrichment*

The GO enrichment was computed using a web resource WebGestalt (WEB-based GENE SeT AnaLysis Toolkit) <sup>262</sup>. A list of genes was prepared for each selected set of predicted digenic pairs based on highest score, highest average score, or most predicted pairs. This list of genes was ranked based on the selection criteria and the GO enrichment for biological process, cellular component and molecular function categories was performed using the online tool.

#### *Structural modeling of Kv3.2*

The tetrameric structural model of human Kv3.2 (UniProtKB accession number: Q96PR1-1, modeled residues: 1-484) was generated by homology modeling using the molecular modeling

software suite Rosetta (version 3.10) <sup>263</sup>. The shaker family voltage dependent potassium channel (Kv1.2-Kv2.1 paddle chimera channel) resolved to 2.4 Å (PDB ID: 2R9R) was used as a template. The percent identity between the aligned positions of the sequences of Kv3.2 and the template structure was 42.8%, sufficiently high for the chimera channel structure to serve as a reliable template. A starting partial tetrameric model of Kv3.2, which only covered aligned residues, was generated by threading the sequence of Kv3.2 onto the template structure using the corresponding sequence alignment as a guide. Full models were created using the Rosetta comparative modeling (RosettaCM) protocol <sup>115</sup> guided by the RosettaMembrane energy function <sup>264</sup> in a C4 symmetry mode <sup>265</sup>. The boundaries of membrane-spanning segments were calculated using the PPM server <sup>267</sup> based on the starting model. The boundaries were used to impose membrane-specific Rosetta energy terms on residues within the theoretical membrane bilayer. A total of 1000 full tetrameric models were generated using RosettaCM. The lowest-energy model was selected as the final model for structure-based analysis in this work. This work by conducted in collaboration with Dr. Bian Li.

### *MD system setup*

The Kv3.2 channel domain (residues L211 – M484) was embedded in a POPC (palmitoyl-oleoyl-phosphatidylcholine) bilayer (~240 lipid molecules per leaflet) using the membrane builder tool of CHARMM-GUI <sup>266</sup>. The system was solvated in TIP3P water containing 0.15 M of neutralizing KCl. Three K<sup>+</sup> ions were placed in the channel selectivity filter at coordination sites S0, S2, and S4 by inferring their positions from the crystal structure of the Kv1.2-2.1 chimeric channel (PDB: 2R9R) <sup>199</sup>. Another K<sup>+</sup> ion was placed below the selectivity filter in the aqueous channel cavity (termed SCav site) and used for running umbrella simulations. During

conventional MD simulations, the position of the cavity  $K^+$  ion was constrained by imposing distance restraints to the selectivity filter residue T437. Dr. Georg Kuenze helped me with this setup.

### *Conventional MD simulations*

MD simulations of the Kv3.2 channel in POPC membranes were performed with AMBER16<sup>269</sup> using the ff14SB force field for proteins<sup>268</sup> and the Lipid17 force field. The system was simulated in four replicas with a total simulation time of  $\sim 1 \mu\text{s}$ . Bonds involving hydrogen atoms were constrained with SHAKE<sup>270</sup>. Nonbonded interactions were evaluated with a  $10 \text{ \AA}$  cutoff, and long-range electrostatic interactions were evaluated by the particle-mesh Ewald method<sup>271</sup>. Each MD system was first minimized using a four-step energy minimization procedure: Minimization of only lipids was followed by minimization of only water + ions, and minimization of protein before the whole system was minimized. With protein and ions restrained to their initial coordinates, the lipid and water were heated to 50 K over 1000 steps with a step size of 1 fs in the NVT ensemble using Langevin dynamics with a rapid collision frequency of  $10,000 \text{ ps}^{-1}$ . The system was then heated to 100 K over 50,000 steps with a collision frequency of  $1000 \text{ ps}^{-1}$  and finally to 310 K over 200,000 steps and a collision frequency of  $100 \text{ ps}^{-1}$ . After changing to the NPT ensemble, restraints on protein and ions were gradually removed over 500 ps. The system was equilibrated for another 10 ns at 310 K with weak positional restraints (with a force constant of  $5 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ ) applied to protein  $C\alpha$  atoms. The protein restraints were then gradually removed over 20 ns, and production MD was conducted for 260 ns using a step size of 2 fs, constant pressure periodic boundary conditions, anisotropic pressure

scaling, and Langevin dynamics. I was assisted by Dr. Georg Kuenze to help set up the simulations.

Subsequent to running production MD, molecules were reimaged back into the simulation box using CPPTRAJ<sup>272</sup> and the final 200 ns of each MD replica were analyzed. The Kv3.2 channel was aligned to the first MD frame and the channel pore radius was measured with HOLE<sup>273</sup> by taking conformations of Kv3.2 at every 1 ns.

### *Umbrella MD simulations*

In order to estimate the free energy of K<sup>+</sup> ion permeation through the cytosolic gate of Kv3.2 WT, V469L, and V471L channels, I calculated the potential of mean force (PMF) of moving a K<sup>+</sup> ion up the pore axis past the cytosolic constriction site and into the cavity below the selectivity filter. The center of mass of the backbone atoms of the selectivity filter residues (T437 – Y440 of all four subunits) was defined as the origin of the pore axis. Umbrella potentials (with a spring constant of 10 kcal mol<sup>-1</sup> Å<sup>-2</sup>) were placed at 0.5 Å intervals in the range from Z = -11 Å (below selectivity filter) to Z = -44 Å (in cytosol), making a total of 67 umbrella simulations for each 1D PMF. In addition, to ensure that the K<sup>+</sup> ion remained in the vicinity of the pore axis when it was no longer constrained by the S6 helices (i.e., was in bulk solvent), I used a method described in Fowler et al.<sup>274</sup> and applied a flat-bottomed cylindrical constraint with a radius of 8 Å and a spring constant of 10 kcal mol<sup>-1</sup> Å<sup>-2</sup>. Starting configurations for each umbrella window were prepared by taking the last frame from the conventional MD simulation and gradually pulling the cavity K<sup>+</sup> ion from SCav into cytosolic solvent over 10 ns using a spring constant of 10 kcal mol<sup>-1</sup> Å<sup>-2</sup>. To ensure that the direction of the pore axis was well-defined and did not change during the simulation, the positions of the backbone atoms of the first two helix turns of

S6 (W448 – G454) were constrained by a harmonic potential with a force constant of 5 kcal mol<sup>-1</sup> Å<sup>-2</sup> during the pulling and umbrella simulations. Each umbrella simulation was run for 10 ns and repeated twice for each of the original four MD replicas. The WHAM method<sup>275</sup> as implemented in the program by Grossfield<sup>276</sup> was used to remove the umbrella biases and calculate 1D PMFs. A final 1D PMF was calculated for Kv3.2 WT, V469L, and V471L, respectively, by averaging the individual PMFs for each variant, and the height of the free energy barrier relative to the bulk solvent was measured.

#### *Heterologous expression of Kv3.2 ion channel and whole-cell voltage clamp electrophysiology*

Wild-type Kv3.2, V469L, and V471L channel plasmids (1 ng/μl for each plasmid) were separately transfected with fluoresced green protein (GFP as marker to identify successful ion channel expression) into Chinese Hamster Ovary (CHO) cells using 10 μl Fugene 6 (Promega), following manufacturer's cell transfection instructions. Two days post transfection, cells with green color were selected for patch clamp experiments.

Whole-cell voltage clamp experiments were performed at room temperature (22-23°C) with 3~5 mΩ patch microelectrodes, by using a MultiClamp 700B amplifier and DigiData 1550B low-noise data acquisition system (Molecular Devices Inc., Sunnydale, California). The extracellular solution contained (in mmol/L) NaCl 145, KCl 4.0, MgCl<sub>2</sub> 1.0, CaCl<sub>2</sub> 1.8, glucose 10, and HEPES 10; the pH was 7.4, adjusted with NaOH. The pipette (intracellular) solution contained (in mmol/L) KCl 110, MgCl<sub>2</sub> 1.0, ATP-K<sub>2</sub> 5.0, BAPTA-K4 5.0, and HEPES 10; the pH was 7.2, adjusted with KOH. Data acquisition was performed using pClamp 10.7 software (Molecular Devices Inc.), sampling at 1 kHz and low-pass-filtered at 5 kHz. Activating current was elicited with 1-second depolarizing pulses from a holding potential of -80 mV at a 10-mV

increments, and tail current was recorded on return to  $-40$  mV. The voltage-clamp protocol is shown in Figure EP. Pulses were delivered every 15 seconds. The current-voltage (I-V) relationships were analyzed by fitting the Boltzmann equation to the data:

$I = I_{\max} / \{1 + \exp [(V_t - V_{0.5}) / k]\}$ , where  $I_{\max}$  is the maximal current,  $V_t$  is the test potential,  $V_{0.5}$  is the membrane potential at which 50% of the channels are activated, and  $k$  is the slope factor. Current densities (pA/pF) were obtained after normalization to cell surface area calculated by the Membrane Test in pClamp 10.7. A potassium channel blocker 4-aminopyridine (4-AP at  $200 \mu\text{M}$ ; Sigma-Aldrich Co., St. Louis, MO, USA) was used to test the sensitivity of wildtype Kv3.2 and two variant (V469L and V471L) channels to drug block by 1-second repetitive pulsing protocol from a holding potential of  $-80$  mV to a testing potential of  $+60$  mV (Figure S3). This work was performed by Dr. Tao Yang, our collaborator and a part of the Dan Roden group.

#### *DNA constructs for WT and variants of KCNC2*

The coding sequences DNA of human *Homo sapiens* potassium voltage-gated channel subfamily C member 2 *KCNC2* (NM\_139137.4) was subcloned into pcDNA3.1+ /C-(k)-DYK expression vector with an equipped Flag tag (DYKDDDDK) in C-terminal (GenScript, NJ, USA). The mutant *KCNC2* variants Kv3.2-V469L and Kv3.2-V471L cDNA constructs were generated by using a pair of designed overlapping primers for the PCR in the QuikChange Site-Directed Mutagenesis Kit (Agilent USA Cat. # 200523) and by PCR Overlap Extension method to introduce the mutation site in. Both variants were confirmed by DNA sequencing.

### *Western Blot*

To detect the Kv3.2 protein expression and to perform protein functional analysis, the wild type and two mutated cDNA plasmids were transfected to CHO stable cells (ATCC, USA) by XtremeGENE 9 DNA Transfection Reagent (Roche). The transfected cells were collected and lysed in modified Radio-Immunoprecipitation Assay (RIPA) lysis buffer (50 mM Tris pH = 7.4, 150mM NaCl, 1% NP-40, 0.2% Sodium Deoxycholate, 1mM EDTA), and 1% protease inhibitor cocktail (Sigma-Aldrich Co., USA). Collected protein samples were subjected to gel electrophoresis using 4–12% BisTris NuPAGE precast gels (Invitrogen Life Technologies, USA) and transferred to PVDF-FL membranes (MilliporeSigma, USA). Primary antibody against Flag epitope tag located on FLAG fusion proteins (Sigma-Aldrich, polyclonal ANTI-FLAG, rabbit host. F7425) was used to detect the Kv3.2 protein by indirect immunofluorescent staining at a 1:500 dilution. Anti-Na<sup>+</sup>/K<sup>+</sup> ATPase antibody (Developmental Studies Hybridoma Bank, Antibodies at the University of Iowa for use in research, USA) at a 1:1000 dilution was used as an internal quality control. IRDye conjugated secondary anti-rabbit antibody (LI-COR Biosciences Inc. USA) was used at a 1:10000 dilution. The membranes were scanned using the Odyssey Infrared Imaging System, and the integrated density value of bands was determined using the Odyssey Image Studio software (LI-COR Biosciences Inc. USA). This work was conducted by Dr. Ningning Hu and others with the Dr. Robert MacDonald group.



## ***DATA AND CODE AVAILABILITY***

The data and code used to train and evaluate DiGePred and other models considered are available at <https://github.com/CapraLab/DiGePred>. The trained DiGePred models are also available in the repository. In addition, digenic pairs from recent literature are provided as **Dataset D2**. The gene pairs predicted to be digenic above our most confident  $F_{0.5}$  threshold are listed in **Dataset D3**, and the predictions using all models of DiGePred on all human gene pairs are in **Datasets D4A-D**. A website that enables the user to access all DiGePred predictions is available at [http://www.meilerlab.org/index.php/servers/show?s\\_id=28](http://www.meilerlab.org/index.php/servers/show?s_id=28).

The data files are available at:

<https://vanderbilt.box.com/shared/static/h5s94d9qhd79mre2a0mgj57z4ljgmtwq>

DATASET D1: *Held-out digenic gene pairs*

DiGePred predictions on held-out digenic gene pairs from DIDA (n=28). These pairs were not used for training and used to test the trained classifier.

<https://vanderbilt.box.com/s/ufsbb48tnkkz5tkckfay23pmrkn3qfqq>

DATASET D2: *Novel digenic gene pairs from recent literature*

DiGePred predictions on novel digenic gene pairs from recent literature (n=13). These pairs were not included in DIDA.

<https://vanderbilt.box.com/s/3cq5f5ldl4h8h8w8hmnd8rmesj6in1hm>

DATASET D3: *Predicted digenic pairs with highest confidence from all possible gene pairs*

Gene pairs predicted to be digenic by DiGePred at most confident threshold. (n=54,318)

<https://vanderbilt.box.com/s/n1nzdyj8i5fa55vultyq4xn6rsp792a7>

DATASET D4A: *Digenic predictions on all human gene pairs*

<https://vanderbilt.box.com/s/459ethsqv339nqiarhm0j227jdjb0whq>

DATASET D4B: *Digenic predictions on all human gene pairs*

<https://vanderbilt.box.com/s/acdqyjuihj3932c6msi5py82rvr5kam3>

DATASET D4C: *Digenic predictions on all human gene pairs*

<https://vanderbilt.box.com/s/kb3vzubfxjctxt8x0y1vytu59x8r8no>

DATASET D4D: *Digenic predictions on all human gene pairs*

DiGePred predictions on all possible human gene pairs (n=155.32 mil.)

The entire MD trajectory videos (V1-V6) for simulations of V469L and V471L *KCNC2* are available at <https://vanderbilt.box.com/s/al6y4ezhmquw8il3wsvhhesdazeqmbyi>.

## **UNDIAGNOSED DISEASES NETWORK *CONSORTIUM***

Maria T Acosta, David R Adams, Pankaj Agrawal, Mercedes E Alejandro, Patrick Allard, Justin Alvey, Ashley Andrews, Euan A Ashley, Mahshid S Azamian, Carlos A Bacino, Guney Bademci, Eva Baker, Ashok Balasubramanyam, Dustin Baldrige, Jim Bale, Deborah Barbouth, Gabriel F Batzli, Pinar Bayrak-Toydemir, Alan H Beggs, Gill Bejerano, Hugo J Bellen, Jonathan A Bernstein, Gerard T Berry, Anna Bican, David P Bick, Camille L Birch, Stephanie Bivona, John Bohnsack, Carsten Bonnenmann, Devon Bonner, Braden E Boone, Bret L Bostwick, Lorenzo Botto, Lauren C Briere, Elly Brokamp, Donna M Brown, Matthew Brush, Elizabeth A Burke, Lindsay C Burrage, Manish J Butte, John Carey, Olveen Carrasquillo, Ta Chen Peter Chang, Hsiao-Tuan Chao, Gary D Clark, Terra R Coakley, Laurel A Cobban, F Sessions Cole, Heather A Colley, Cynthia M Cooper, Heidi Cope, William J Craigen, Precilla D'Souza, Surendra Dasari, Mariska Davids, Jyoti G Dayal, Esteban C Dell'Angelica, Shweta U Dhar, Naghmeh Dorrani, Daniel C Dorset, Emilie D Douine, David D Draper, Laura Duncan, David J Eckstein, Lisa T Emrick, Christine M Eng, Cecilia Esteves, Tyra Estwick, Liliana Fernandez, Carlos Ferreira, Elizabeth L Fieg, Paul G Fisher, Brent L Fogel, Irman Forghani, Laure Fresard, William A Gahl, Rena A Godfrey, Alica M Goldman, David B Goldstein, Jean-Philippe F Gourdine, Alana Grajewski, Catherine A Groden, Andrea L Gropman, Melissa Haendel, Neil A Hanchard, Nichole Hayes, Frances High, Ingrid A Holm, Jason Hom, Yong Huang, Alden Huang, Rosario Isasi, Fariha Jamal, Yong-Hui Jiang, Jean M Johnston, Angela L Jones, Lefkothea Karaviti, Emily G Kelley, Dana Kiley, David M Koeller, Isaac S Kohane, Jennefer N Kohler, Susan Korrick, Mary E Koziura, Deborah Krakow, Donna M Krasnewich, Joel B Krier, Jennifer E Kyle, Seema R Lalani, Byron Lam, Brendan C Lanpher, Ian R Lanza, C Christopher Lau, Pace Laura, Jozef Lazar, Kimberly LeBlanc, Brendan H Lee, Hane Lee, Roy Levitt, Shawn

E Levy, Richard A Lewis, Sharyn A Lincoln, Pengfei Liu, Xue Zhong Liu, Nicola Longo, Sandra K Loo, Joseph Loscalzo, Richard L Maas, Calum A MacRae, Ellen F Macnamara, Valerie V Maduro, Marta M Majcherska, May Christine V Malicdan, Laura A Mamounas, Teri A Manolio, Rong Mao, Thomas C Markello, Ronit Marom, Gabor Marth, Beth A Martin, Martin G Martin, Julian A Martínez-Agosto, Shruti Marwaha, Thomas May, Jacob McCauley, Allyn McConkie-Rosell, Colleen E McCormack, Alexa T McCray, Thomas O Metz, Matthew Might, Eva Morava-Kozicz, Paolo M Moretti, Marie Morimoto, John J Mulvihill, David R Murdock, Avi Nath, Stanley F Nelson, J Scott Newberry, Sarah K Nicholas, Donna Novacic, Devin Oglesbee, James P Orengo, Stephen Pak, J Carl Pallais, Christina G S Palmer, Moretti Paolo, Jeanette C Papp, Neil H Parker, Jennifer E Posey, John H Postlethwait, Lorraine Potocki, Barbara N Pusey, Aaron Quinlan, Archana N Raja, Genecee Renteria, Chloe M Reuter, Lynette C Rives, Amy K Robertson, Lance H Rodan, Jill A Rosenfeld, Robb K Rowley, Maura Ruzhnikov, Ralph Sacco, Jacinda B Sampson, Susan L Samson, Mario Saporta, Judy Schaechter, Timothy Schedl, Kelly Schoch, Daryl A Scott, Lisa Shakachite, Prashant Sharma, Vandana Shashi, Kathleen Shields, Jimann Shin, Rebecca H Signer, Catherine H Sillari, Edwin K Silverman, Janet S Sinsheimer, Kathy Sisco, Kevin S Smith, Lilianna Solnica-Krezel, Rebecca C Spillmann, Joan M Stoler, Nicholas Stong, Jennifer A Sullivan, Shirley Sutton, David A Sweetser, Holly K Tabor, Cecelia P Tamburro, Queenie K-G Tan, Mustafa Tekin, Fred Telischi, Willa Thorson, Cynthia J Tifft, Camilo Toro, Alyssa A Tran, Tiina K Urv, Matt Velinder, Dave Viskochil, Tiphonie P Vogel, Colleen E Wahl, Melissa Walker, Nicole M Walley, Chris A Walsh, Jennifer Wambach, Jijun Wan, Lee-Kai Wang, Michael F Wangler, Patricia A Ward, Katrina M Waters, Bobbie-Jo M Webb-Robertson, Daniel Wegner, Monte Westerfield, Matthew T Wheeler, Anastasia L Wise, Lynne A Wolfe, Jeremy D Woods, Elizabeth A Worthey, Shinya

Yamamoto, John Yang, Amanda J Yoon, Guoyun Yu, Diane B Zastrow, Chunli Zhao, Stephan  
Zuchner.

## REFERENCES

1. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nat.* 2001 4096822 *409*, 860–921.
2. Hood, L., and Rowen, L. (2013). The human genome project: Big science transforms biology and medicine. *Genome Med.* *5*, 1–8.
3. Boycott, K.M., Vanstone, M.R., Bulman, D.E., and MacKenzie, A.E. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat. Rev. Genet.* 2013 1410 *14*, 681–691.
4. Klimova, B., Storek, M., Valis, M., and Kuca, K. (2017). Global View on Rare Diseases: A Mini Review. *Curr. Med. Chem.* *24*,
5. Schaaf, J., Sedlmayr, M., Schaefer, J., and Storf, H. (2020). Diagnosis of Rare Diseases: a scoping review of clinical decision support systems. *Orphanet J. Rare Dis.* *15*, 263.
6. Wright, C.F., FitzPatrick, D.R., and Firth, H. V. (2018). Paediatric genomics: diagnosing rare disease in children. *Nat. Rev. Genet.* 2018 195 *19*, 253–268.
7. Boycott, K.M., Rath, A., Chong, J.X., Hartley, T., Alkuraya, F.S., Baynam, G., Brookes, A.J., Brudno, M., Carracedo, A., den Dunnen, J.T., et al. (2017). International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. *Am. J. Hum. Genet.* *100*, 695–705.
8. Chong, J.X., Buckingham, K.J., Jhangiani, S.N., Boehm, C., Sobreira, N., Smith, J.D., Harrell, T.M., McMillin, M.J., Wiszniewski, W., Gambin, T., et al. (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am. J. Hum. Genet.* *97*, 199–215.
9. Kaiser, J. (2010). Affordable “exomes” fill gaps in a catalog of rare diseases. *Science* (80- ).

330, 903.

10. Stankiewicz, P., and Lupski, J.R. (2010). Structural Variation in the Human Genome and its Role in Disease. <https://doi.org/10.1146/annurev-med-100708-204735> *61*, 437–455.

11. Green, E.D., and Guyer, M.S. (2011). Charting a course for genomic medicine from base pairs to bedside. *Nat.* 2011 4707333 *470*, 204–213.

12. Bainbridge, M.N., Wiszniewski, W., Murdock, D.R., Friedman, J., Gonzaga-Jauregui, C., Newsham, I., Reid, J.G., Fink, J.K., Morgan, M.B., Gingras, M.C., et al. (2011). Whole-genome sequencing for optimized patient management. *Sci. Transl. Med.* *3*,.

13. Cano-Gamez, E., and Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front. Genet.* *11*, 424.

14. Lewis, C.M., and Vassos, E. (2020). Polygenic risk scores: From research tools to clinical instruments. *Genome Med.* *12*, 1–11.

15. Product, Development, I. of M. (US) C. on A.R.D.R. and O., Field, M.J., and Boat, T.F. (2010). Innovation and the Orphan Drug Act, 1983-2009: Regulatory and Clinical Characteristics of Approved Orphan Drugs.

16. Office of Inspector General, H. (2001). OFFICE OF INSPECTOR GENERAL THE ORPHAN DRUG ACT.

17. Ramoni, R.B., Mulvihill, J.J., Adams, D.R., Allard, P., Ashley, E.A., Bernstein, J.A., Gahl, W.A., Hamid, R., Loscalzo, J., McCray, A.T., et al. (2017). The Undiagnosed Diseases Network: Accelerating Discovery about Health and Disease. *Am. J. Hum. Genet.* *100*, 185–192.

18. Gahl, W.A., Mulvihill, J.J., Toro, C., Markello, T.C., Wise, A.L., Ramoni, R.B., Adams, D.R., and Tiffit, C.J. (2016). The NIH Undiagnosed Diseases Program and Network: Applications to modern medicine. *Mol. Genet. Metab.* *117*, 393–400.

19. UDN <https://undiagnosed.hms.harvard.edu/about-us/facts-and-figures/>.
20. Liu, N., Schoch, K., Luo, X., Pena, L.D.M., Bhavana, V.H., Kukolich, M.K., Stringer, S., Powis, Z., Radtke, K., Mroske, C., et al. (2018). Functional variants in TBX2 are associated with a syndromic cardiovascular and skeletal developmental disorder. *Hum. Mol. Genet.* *27*, 2454–2465.
21. Chao, H.-T., Davids, M., Burke, E., Pappas, J.G., Rosenfeld, J.A., McCarty, A.J., Davis, T., Wolfe, L., Toro, C., Tiff, C., et al. (2017). A Syndromic Neurodevelopmental Disorder Caused by De Novo Variants in EBF3. *Am. J. Hum. Genet.* *100*, 128–137.
22. Tokita, M.J., Chen, C.-A., Chitayat, D., Macnamara, E., Rosenfeld, J.A., Hanchard, N., Lewis, A.M., Brown, C.W., Marom, R., Shao, Y., et al. (2018). De Novo Missense Variants in TRAF7 Cause Developmental Delay, Congenital Anomalies, and Dysmorphic Features. *Am. J. Hum. Genet.* *103*, 154–162.
23. Schoch, K., Meng, L., Szelinger, S., Bearden, D.R., Stray-Pedersen, A., Busk, O.L., Stong, N., Liston, E., Cohn, R.D., Scaglia, F., et al. (2017). A Recurrent De Novo Variant in NACCI1 Causes a Syndrome Characterized by Infantile Epilepsy, Cataracts, and Profound Developmental Delay. *Am. J. Hum. Genet.* *100*, 343–351.
24. Bostwick, B.L., McLean, S., Posey, J.E., Streff, H.E., Gripp, K.W., Blesson, A., Powell-Hamilton, N., Tusi, J., Stevenson, D.A., Farrelly, E., et al. (2017). Phenotypic and molecular characterisation of CDK13-related congenital heart defects, dysmorphic facial features and intellectual developmental disorders. *Genome Med.* *9*, 73.
25. Küry, S., van Woerden, G.M., Besnard, T., Proietti Onori, M., Latypova, X., Towne, M.C., Cho, M.T., Prescott, T.E., Ploeg, M.A., Sanders, S., et al. (2017). De Novo Mutations in Protein Kinase Genes CAMK2A and CAMK2B Cause Intellectual Disability. *Am. J. Hum. Genet.* *101*,



768–788.

26. Pomerantz, D.J., Ferdinandusse, S., Cogan, J., Cooper, D.N., Reimschisel, T., Robertson, A., Bican, A., McGregor, T., Gauthier, J., Millington, D.S., et al. (2018). Clinical heterogeneity of mitochondrial NAD kinase deficiency caused by a *NADK2* start loss variant. *Am. J. Med. Genet. Part A* *176*, 692–698.

27. Oláhová, M., Yoon, W.H., Thompson, K., Jangam, S., Fernandez, L., Davidson, J.M., Kyle, J.E., Grove, M.E., Fisk, D.G., Kohler, J.N., et al. (2018). Biallelic Mutations in *ATP5F1D*, which Encodes a Subunit of ATP Synthase, Cause a Metabolic Disorder. *Am. J. Hum. Genet.* *102*, 494–504.

28. Oláhová, M., Yoon, W.H., Thompson, K., Jangam, S., Fernandez, L., Davidson, J.M., Kyle, J.E., Grove, M.E., Fisk, D.G., Kohler, J.N., et al. (2018). Biallelic Mutations in *ATP5F1D*, which Encodes a Subunit of ATP Synthase, Cause a Metabolic Disorder. *Am. J. Hum. Genet.* *102*, 494–504.

29. Johnston, J.J., van der Smagt, J.J., Rosenfeld, J.A., Pagnamenta, A.T., Alswaid, A., Baker, E.H., Blair, E., Borck, G., Brinkmann, J., Craigen, W., et al. (2018). Autosomal recessive Noonan syndrome associated with biallelic *LZTR1* variants. *Genet Med.*

30. Poli, M.C., Ebstein, F., Nicholas, S.K., de Guzman, M.M., Forbes, L.R., Chinn, I.K., Mace, E.M., Vogel, T.P., Carisey, A.F., Benavides, F., et al. (2018). Heterozygous Truncating Variants in *POMP* Escape Nonsense-Mediated Decay and Cause a Unique Immune Dysregulatory Syndrome. *Am. J. Hum. Genet.* *102*, 1126–1142.

31. Machol, K., Jankovic, J., Vijayakumar, D., Burrage, L.C., Jain, M., Lewis, R.A., Fuller, G.N., Xu, M., Penas-Prado, M., Gule-Monroe, M.K., et al. (2018). Atypical Alexander disease with dystonia, retinopathy, and a brain mass mimicking astrocytoma. *Neurol. Genet.* *4*, e248.

32. Marcogliese, P.C., Shashi, V., Spillmann, R.C., Stong, N., Rosenfeld, J.A., Koenig, M.K., Martínez-Agosto, J.A., Herzog, M., Chen, A.H., Dickson, P.I., et al. (2018). IRF2BPL Is Associated with Neurological Phenotypes. *Am. J. Hum. Genet.* *103*, 245–260.
33. Auer, F., Lin, M., Nebral, K., Gertzen, C.G.W., Haas, O.A., Kuhlen, M., Gohlke, H., Izraeli, S., Trka, J., Hu, J., et al. (2018). Novel Recurrent Germline JAK2 G571S Variant in Childhood Acute B-Lymphoblastic Leukemia: A Double Hit One Pathway Scenario. *Blood* *132*, 387–387.
34. Pehlivan, D., Bayram, Y., Gunes, N., Coban Akdemir, Z., Shukla, A., Bierhals, T., Tabakci, B., Sahin, Y., Gezdirici, A., Fatih, J.M., et al. (2019). The Genomics of Arthrogyrosis, a Complex Trait: Candidate Genes and Further Evidence for Oligogenic Inheritance. *Am. J. Hum. Genet.* *105*, 132–150.
35. Badano, J.L., and Katsanis, N. (2002). Beyond mendel: An evolving view of human genetic disease transmission. *Nat. Rev. Genet.* *3*, 779–789.
36. van Heyningen, V., and Yeyati, P.L. (2004). Mechanisms of non-Mendelian inheritance in genetic disease. *Hum. Mol. Genet.* *13*, R225–R233.
37. Kajiwara, K., Berson, E.L., and Dryja, T.P. (1994). Digenic retinitis pigmentosa due to mutations at the unlinked peripherin/RDS and ROM1 loci. *Science* *264*, 1604–1608.
38. Schäffer, A.A. (2013). Digenic inheritance in medical genetics. *J. Med. Genet.* *50*, 641–652.
39. Gazzo, A.M., Daneels, D., Cilia, E., Bonduelle, M., Abramowicz, M., Van Dooren, S., Smits, G., and Lenaerts, T. (2016). DIDA: A curated and annotated digenic diseases database. *Nucleic Acids Res.* *44*, D900–D907.
40. Lupski, J.R. (2012). Digenic inheritance and Mendelian disease. *Nat. Genet.* *44*, 1291–1292.
41. Deltas, C. (2018). Digenic inheritance and genetic modifiers. *Clin. Genet.* *93*, 429–438.
42. Gazzo, A., Raimondi, D., Daneels, D., Moreau, Y., Smits, G., Van Dooren, S., and Lenaerts,

- T. (2017). Understanding mutational effects in digenic diseases. *Nucleic Acids Res.* *45*, e140.
43. Papadimitriou, S., Gazzo, A., Versbraegen, N., Nachtegael, C., Aerts, J., Moreau, Y., Van Dooren, S., Nowé, A., Smits, G., and Lenaerts, T. (2019). Predicting disease-causing variant combinations. *Proc. Natl. Acad. Sci. U. S. A.* *116*, 11878–11887.
44. Boudellioua, I., Kulmanov, M., Schofield, P.N., Gkoutos, G. V., and Hoehndorf, R. (2018). OligoPVP: Phenotype-driven analysis of individual genomic information to prioritize oligogenic disease variants. *Sci. Rep.* *8*,.
45. Renaux, A., Papadimitriou, S., Versbraegen, N., Nachtegael, C., Boutry, S., Nowé, A., Smits, G., and Lenaerts, T. (2019). ORVAL: a novel platform for the prediction and exploration of disease-causing oligogenic variant combinations. *Nucleic Acids Res.* *47*, W93–W98.
46. Murphy, K.P. (2012). *Machine Learning: A Probabilistic Perspective* - Kevin P. Murphy - Google Books.
47. Franklin, J. (2008). The elements of statistical learning: data mining, inference and prediction. *Math. Intell.* *2005* *27* *27*, 83–85.
48. Carbonell, J.G., and Michalski, R.S. (2013). *Machine Learning: An Artificial Intelligence Approach* - Google Books.
49. Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Mol. Syst. Biol.* *12*, 878.
50. Xu, C., and Jackson, S.A. (2019). Machine learning and complex biological data. *Genome Biol.* *20*,.
51. Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., and Hoffman, M.M. (2019). *Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities.* *Inf. Fusion* *50*, 71.

52. Altman, N., and Krzywinski, M. (2018). The curse(s) of dimensionality. *Nat. Methods* *15*, 399–400.
53. Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine Learning in Medicine. *N. Engl. J. Med.* *380*, 1347–1358.
54. Wong, E. (2021). Media Review: Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. *InnovAiT Educ. Inspir. Gen. Pract.* 175573802110182.
55. Réda, C., Kaufmann, E., and Delahaye-Duriez, A. (2020). Machine learning applications in drug development. *Comput. Struct. Biotechnol. J.* *18*, 241–252.
56. Schaefer, J., Lehne, M., Schepers, J., Prasser, F., and Thun, S. (2020). The use of machine learning in rare diseases: A scoping review. *Orphanet J. Rare Dis.* *15*,.
57. Lehne, M., Sass, J., Essenwanger, A., Schepers, J., and Thun, S. (2019). Why digital medicine depends on interoperability. *Npj Digit. Med.* *2*, 1–5.
58. Rath, A., Olry, A., Dhombres, F., Brandt, M.M., Urbero, B., and Ayme, S. (2012). Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. *Hum. Mutat.* *33*, 803–808.
59. Robinson, P.N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.* *83*, 610–615.
60. Groza, T., Köhler, S., Moldenhauer, D., Vasilevsky, N., Baynam, G., Zemojtel, T., Schriml, L.M., Kibbe, W.A., Schofield, P.N., Beck, T., et al. (2015). The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease. *Am. J. Hum. Genet.* *97*, 111–124.
61. Amiri, H., and Kohane, I.S. (2021). Machine Learning of Patient Characteristics to Predict Admission Outcomes in the Undiagnosed Diseases Network. *JAMA Netw. Open* *4*,.

62. Parmar, A., Katariya, R., and Patel, V. (2019). A Review on Random Forest: An Ensemble Classifier. In *Lecture Notes on Data Engineering and Communications Technologies*, (Springer, Cham), pp. 758–763.
63. Breiman, L. (2001). Random forests. *Mach. Learn.* *45*, 5–32.
64. Sarica, A., Cerasa, A., and Quattrone, A. (2017). Random forest algorithm for the classification of neuroimaging data in Alzheimer’s disease: A systematic review. *Front. Aging Neurosci.* *9*, 329.
65. Menze, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., and Hamprecht, F.A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics* *10*, 1–16.
66. Chen, X., Wang, M., and Zhang, H. (2011). The use of classification trees for bioinformatics. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* *1*, 55–63.
67. Calle, M.L., Urrea, V., Boulesteix, A.L., and Malats, N. (2011). AUC-RF: A new strategy for genomic profiling with random forest. *Hum. Hered.* *72*, 121–132.
68. Caruana, R., and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *ACM International Conference Proceeding Series*, (New York, New York, USA: ACM Press), pp. 161–168.
69. Strobl, C., Boulesteix, A.L., and Augustin, T. (2007). Unbiased split selection for classification trees based on the Gini Index. *Comput. Stat. Data Anal.* *52*, 483–501.
70. Kuhlman, B., and Bradley, P. (2019). Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.* *20*, 681–697.
71. Anson, M.L., and Mirsky, A.E. (1930). Protein coagulation and its reversal the preparation of

- insoluble globin, soluble globin and heme. *J. Gen. Physiol.* *13*, 469–476.
72. Anfinsen, C.B., and Scheraga, H.A. (1975). Experimental and theoretical aspects of protein folding. *Adv. Protein Chem.* *29*, 205–300.
73. Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. *Science* (80-). *181*, 223–230.
74. ANFINSEN, C.B., HABER, E., SELA, M., and WHITE, F.H. (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. U. S. A.* *47*, 1309–1314.
75. Karplus, M. (1997). The Levinthal paradox: Yesterday and today. *Fold. Des.* *2*,.
76. Dill, K.A., Ozkan, S.B., Shell, M.S., and Weikl, T.R. (2008). The protein folding problem. *Annu. Rev. Biophys.* *37*, 289–316.
77. Bryngelson, J.D., Onuchic, J.N., Socci, N.D., and Wolynes, P.G. (1995). Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins Struct. Funct. Bioinforma.* *21*, 167–195.
78. Levitt, M., and Warshel, A. (1975). Computer simulation of protein folding. *Nature* *253*, 694–698.
79. Nealon, J.O., Philomina, L.S., and McGuffin, L.J. (2017). Predictive and experimental approaches for elucidating protein-protein interactions and quaternary structures. *Int. J. Mol. Sci.* *18*,.
80. Emwas, A.H.M. (2015). The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. *Methods Mol. Biol.* *1277*, 161–193.
81. Kikhney, A.G., and Svergun, D.I. (2015). A practical guide to small angle X-ray scattering

- (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett.* *589*, 2570–2577.
82. Nogales, E. (2015). The development of cryo-EM into a mainstream structural biology technique. *Nat. Methods* *13*, 24–27.
83. Bai, X. chen, McMullan, G., and Scheres, S.H.W. (2015). How cryo-EM is revolutionizing structural biology. *Trends Biochem. Sci.* *40*, 49–57.
84. Wang, H.W., and Wang, J.W. (2017). How cryo-electron microscopy and X-ray crystallography complement each other. *Protein Sci.* *26*, 32–39.
85. Chatham, J.C., and Blackband, S.J. (2001). Nuclear magnetic resonance spectroscopy and imaging in animal research. *ILAR J.* *42*, 189–208.
86. Brünger, A.T. (1997). X-ray crystallography and NMR reveal complementary views of structure and dynamics. *Nat. Struct. Biol.* *4*, 862–865.
87. Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P.D., Smith, C.A., Sheffler, W., et al. (2011). Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. In *Methods in Enzymology*, pp. 545–574.
88. Barlow, K.A., Conchuir, S.O., Thompson, S., Suresh, P., Lucas, J.E., Heinonen, M., Kortemme, T., and Biohub, C.Z. (2018). Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein–Protein Binding Affinity upon Mutation. *J. Phys. Chem. B* *122*, 5389–5399.
89. Delgado, J., Radusky, L.G., Cianferoni, D., and Serrano, L. (2019). FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics* *35*, 4168–4169.
90. Buß, O., Rudat, J., and Ochsenreither, K. (2018). FoldX as Protein Engineering Tool: Better Than Random Based Approaches? *Comput. Struct. Biotechnol. J.* *16*, 25–33.
91. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein

structure prediction with AlphaFold. *Nature* 596, 583–589.

92. Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W.R., Bridgland, A., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature* 577, 706–710.

93. Hospital, A., Goñi, J.R., Orozco, M., and Gelpí, J.L. (2015). Molecular dynamics simulations: Advances and applications. *Adv. Appl. Bioinforma. Chem.* 8, 37–47.

94. Hollingsworth, S.A., and Dror, R.O. (2018). Molecular Dynamics Simulation for All. *Neuron* 99, 1129–1143.

95. McCammon, J.A., Gelin, B.R., and Karplus, M. (1977). Dynamics of folded proteins. *Nature* 267, 585–590.

96. Salomon-Ferrer, R., Götz, A.W., Poole, D., Le Grand, S., and Walker, R.C. (2013). Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh ewald. *J. Chem. Theory Comput.* 9, 3878–3888.

97. Stone, J.E., Hallock, M.J., Phillips, J.C., Peterson, J.R., Luthey-Schulten, Z., and Schulten, K. (2016). Evaluation of emerging energy-efficient heterogeneous computing platforms for biomolecular and cellular simulation workloads. In *Proceedings - 2016 IEEE 30th International Parallel and Distributed Processing Symposium, IPDPS 2016*, (Institute of Electrical and Electronics Engineers Inc.), pp. 89–100.

98. Lindorff-Larsen, K., Best, R.B., DePristo, M.A., Dobson, C.M., and Vendruscolo, M. (2005). Simultaneous determination of protein structure and dynamics. *Nature* 433, 128–132.

99. Lindorff-Larsen, K., Piana, S., Dror, R.O., and Shaw, D.E. (2011). How fast-folding proteins fold. *Science* (80-. ). 334, 517–520.

100. Lindorff-Larsen, K., Maragakis, P., Piana, S., Eastwood, M.P., Dror, R.O., and Shaw, D.E.



(2012). Systematic validation of protein force fields against experimental data. *PLoS One* 7, e32131.

101. Shaw, D.E., Deneroff, M.M., Dror, R.O., Kuskin, J.S., Larson, R.H., Salmon, J.K., Young, C., Batson, B., Bowers, K.J., Chao, J.C., et al. (2008). Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM* 51, 91–97.

102. Case, D.A., Darden Thomas Cheatham III Carlos Simmerling Junmei Wang, T.E., Duke, R.E., Crowley Ross Walker Wei Zhang Kenneth Merz Bing Wang Seth Hayik Adrian Roitberg Gustavo Seabra István Kolossváry Budapest, M.M., Shaw Kim Wong, D.F., Paesani, F., Vanicek Xiongwu Wu Scott Brozell Thomas Steinbrecher Holger Gohlke Lijiang Yang Chunhu Tan John Mongan Viktor Hornak Guanglei Cui David H Mathews Matthew G Seetin Celeste Sagui Volodymyr Babin Peter A Kollman, J.R., Pearlman Robert V Stanton Jed Pitara Irina Massova Ailan Cheng James J Vincent Paul Beroza Vickie Tsui Christian Schafmeister Wilson S Ross Randall Radmer George L Seibel James W Caldwell U Chandra Singh Paul Weiner, D.A., and Cieplak Yong Duan Rob Woods Karl Kirschner Sarah Tschampel Alexey Onufriev Christopher Bayly Wendy Cornell Scott Weiner Austin Yongye Matthew Tessier, P.M. (2008). *Amber 10 Users' Manual* Principal contributors to the current codes: Additional key contributors to earlier versions: Additional key people involved in force field development.

103. Fernandez-Leiro, R., and Scheres, S.H.W. (2016). Unravelling biological macromolecules with cryo-electron microscopy. *Nature* 537, 339–346.

104. Rost, B., Radivojac, P., and Bromberg, Y. (2016). Protein function in precision medicine: deep understanding with machine learning. *FEBS Lett.* 590, 2327–2341.

105. Mahlich, Y., Reeb, J., Hecht, M., Schelling, M., De Beer, T.A.P., Bromberg, Y., and Rost, B. (2017). Common sequence variants affect molecular function more than rare variants? *Sci.*

Rep. 7,.

106. Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: A new style of protein science. *Nat. Methods* 11, 801–807.
107. Araya, C.L., and Fowler, D.M. (2011). Deep mutational scanning: Assessing protein function on a massive scale. *Trends Biotechnol.* 29, 435–442.
108. Majithia, A.R., Tsuda, B., Agostini, M., Gnanapradeepan, K., Rice, R., Peloso, G., Patel, K.A., Zhang, X., Broekema, M.F., Patterson, N., et al. (2016). Prospective functional classification of all possible missense variants in PPARG. *Nat. Genet.* 48, 1570–1575.
109. Narayanan, K.K., and Procko, E. (2021). Deep Mutational Scanning of Viral Glycoproteins and Their Host Receptors. *Front. Mol. Biosci.* 8, 228.
110. Reeb, J., Wirth, T., and Rost, B. (2020). Variant effect predictions capture some aspects of deep mutational scanning experiments. *BMC Bioinformatics* 21,.
111. Araya, C.L., and Fowler, D.M. (2011). Deep mutational scanning: Assessing protein function on a massive scale. *Trends Biotechnol.* 29, 435–442.
112. Matreyek, K.A., Starita, L.M., Stephany, J.J., Martin, B., Chiasson, M.A., Gray, V.E., Kircher, M., Khechaduri, A., Dines, J.N., Hause, R.J., et al. (2018). Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* 50, 874–882.
113. Hoskins, R.A., Repo, S., Barsky, D., Andreoletti, G., Moulton, J., and Brenner, S.E. (2017). Reports from CAGI: The Critical Assessment of Genome Interpretation. *Hum. Mutat.* 38, 1039–1041.
114. Eyre-Walker, A., and Keightley, P.D. (2007). The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* 8, 610–618.
115. Song, Y., Dimaio, F., Wang, R.Y.R., Kim, D., Miles, C., Brunette, T., Thompson, J., and

- Baker, D. (2013). High-resolution comparative modeling with RosettaCM. *Structure* 21, 1735–1742.
116. Jaccard, P. (1912). THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1. *New Phytol.* 11, 37–50.
117. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361.
118. Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al. (2018). The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* 46, D649–D655.
119. Milacic, M., Haw, R., Rothfels, K., Wu, G., Croft, D., Hermjakob, H., D’Eustachio, P., and Stein, L. (2012). Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers (Basel).* 4, 1180–1211.
120. Köhler, S., Vasilevsky, N.A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S.M., Boerkoel, C.F., Boycott, K.M., et al. (2017). The Human Phenotype Ontology in 2017. *Nucleic Acids Res.* 45, D865–D876.
121. Okamura, Y., Aoki, Y., Obayashi, T., Tadaka, S., Ito, S., Narise, T., and Kinoshita, K. (2015). COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. *Nucleic Acids Res.* 43, D82–D86.
122. Poon, H., Quirk, C., DeZiel, C., and Heckerman, D. (2014). Literome: PubMed-scale genomic knowledge base in the cloud. *Bioinformatics* 30, 2840–2842.
123. Park, Y., and Marcotte, E.M. (2012). Flaws in evaluation schemes for pair-input computational predictions. *Nat. Methods* 9, 1134–1136.

124. Costanzo, M., VanderSluis, B., Koch, E.N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S.D., et al. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science* (80-. ). 353,.
125. Strobl, C., Boulesteix, A.L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* 8, 1–21.
126. Schrauwen, I., Chakchouk, I., Acharya, A., Liaqat, K., Nickerson, D.A., Bamshad, M.J., Shah, K., Ahmad, W., Leal, S.M., Anderson, P., et al. (2018). Novel digenic inheritance of PCDH15 and USH1G underlies profound non-syndromic hearing impairment. *BMC Med. Genet.* 19,.
127. Ameratunga, R., Koopmans, W., Woon, S.-T., Leung, E., Lehnert, K., Slade, C.A., Tempany, J.C., Enders, A., Steele, R., Browett, P., et al. (2017). Epistatic interactions between mutations of TACI (TNFRSF13B) and TCF3 result in a severe primary immunodeficiency disorder and systemic lupus erythematosus. *Clin. Transl. Immunol.* 6, 159.
128. Hoyos-Bachiloglu, R., Alzahrani, M., and Geha, R.S. (2017). A digenic human immunodeficiency characterized by IFNAR1 and IFNGR2 mutations *The Journal of Clinical Investigation.* *J Clin Invest* 127,.
129. Nozu, K., Inagaki, T., Fu, X.J., Nozu, Y., Kaito, H., Kanda, K., Sekine, T., Igarashi, T., Nakanishi, K., Yoshikawa, N., et al. (2008). Molecular analysis of digenic inheritance in Bartter syndrome with sensorineural deafness. *J. Med. Genet.* 45, 182–186.
130. Abdallah, A.M., Carlus, S.J., Al-Mazroea, A.H., Alluqmani, M., Almohammadi, Y., Bhuiyan, Z.A., Al-Harbi, K.M., Abdallah, A.M., Carlus, S.J., Al-Mazroea, A.H., et al. (2019). Digenic Inheritance of LAMA4 and MYH7 Mutations in Patient with Infantile Dilated

Cardiomyopathy. *Medicina (B. Aires)*. 55, 17.

131. Nieto-Marín, P., Jiménez-Jáimez, J., Tinaquero, D., Alfayate, S., Utrilla, R.G., Rodríguez Vázquez del Rey, M. del M., Perin, F., Sarquella-Brugada, G., Monserrat, L., Brugada, J., et al. (2019). Digenic Heterozygosity in SCN5A and CACNA1C Explains the Variable Expressivity of the Long QT Phenotype in a Spanish Family. *Rev. Española Cardiol. (English Ed.)* 72, 324–332.

132. Stone, S.I., Wegner, D.J., Wambach, J.A., Cole, F.S., Ornitz, D.M., and Urano, F. (2019). 26-OR: Digenic FGFR1/KLB Variants Associated with Endocrine Specific FGF-21 Signaling Defects and Extreme Insulin Resistance. *Diabetes* 68, 26-OR.

133. Kong, Y., Xu, K., Yuan, K., Zhu, J., Gu, W., Liang, L., and Wang, C. (2019). Digenetic inheritance of SLC12A3 and CLCNKB genes in a Chinese girl with Gitelman syndrome. *BMC Pediatr.* 19, 114.

134. Yang, Y., Ye, W., Guo, J., Zhao, L., Tu, M., Zheng, Y., and Li, L. (2018). CLCN7 and TCIRG1 mutations in a single family: Evidence for digenic inheritance of osteopetrosis. *Mol. Med. Rep.* 19, 595–600.

135. Heida, A., Van Der Does, L.J.M.E., Ragab, A.A.Y., and De Groot, N.M.S. (2019). A Rare Case of the Digenic Inheritance of Long QT Syndrome Type 2 and Type 6. *Case Rep. Med.* 2019, 1–4.

136. Zastrow, D.B., Zornio, P.A., Dries, A., Kohler, J., Fernandez, L., Waggott, D., Walkiewicz, M., Eng, C.M., Manning, M.A., Farrelly, E., et al. (2017). Heterozygous Exome sequencing identifies de novo pathogenic variants in FBN1 and TRPS1 in a patient with a complex connective tissue phenotype. *Cold Spring Harb. Mol. Case Stud.* 3, a001388.

137. Schaefer, J., Lehne, M., Schepers, J., Prasser, F., and Thun, S. (2020). The use of machine learning in rare diseases: A scoping review. *Orphanet J. Rare Dis.* 15,.

138. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443.
139. Vucic, S., Rothstein, J.D., and Kiernan, M.C. Advances in treating amyotrophic lateral sclerosis: insights from pathophysiological studies.
140. Mikhael, S., Dugar, S., Morton, M., Chorich, L.P., Tam, K.B., Lossie, A.C., Kim, H.G., Knight, J., Taylor, H.S., Mukherjee, S., et al. (2021). Genetics of agenesis/hypoplasia of the uterus and vagina: narrowing down the number of candidate genes for Mayer–Rokitansky–Küster–Hauser Syndrome. *Hum. Genet.* 140,.
141. Farg, M.A., Soo, K.Y., Warraich, S.T., Sundaramoorthy, V., Blair, I.P., and Atkin, J.D. (2020). Erratum: Ataxin-2 interacts with FUS and intermediate-length polyglutamine expansions enhance FUS-related pathology in amyotrophic lateral sclerosis (*Human Molecular Genetics* (2013) 22:4 (717–728) DOI: 10.1093/hmg/dds479). *Hum. Mol. Genet.* 29, 703–704.
142. Ostrowski, L.A., Hall, A.C., and Mekhail, K. Ataxin-2: From RNA Control to Human Health and Disease.
143. Mikhael, S., Dugar, S., Morton, M., Chorich, L.P., Tam, K.B., Lossie, A.C., Kim, H.-G., Knight, J., Taylor, H.S., Mukherjee, S., et al. (2021). Genetics of agenesis/hypoplasia of the uterus and vagina: narrowing down the number of candidate genes for Mayer–Rokitansky–Küster–Hauser Syndrome. *Hum. Genet.*
144. Morcel, K., Camborieux, L., and Guerrier, D. (2007). Mayer-Rokitansky-Küster-Hauser (MRKH) syndrome. *Orphanet J. Rare Dis.* 2,.
145. Patnaik, S.S., Brazile, B., Dandolu, V., Ryan, P.L., and Liao, J. (2015). Mayer-Rokitansky-Küster-Hauser (MRKH) syndrome: A historical perspective. *Gene* 555, 33–40.

146. Jenkins, M.H., Alrowaished, S.S., Goody, M.F., Crawford, B.D., and Henry, C.A. (2016). Laminin and Matrix metalloproteinase 11 regulate Fibronectin levels in the zebrafish myotendinous junction. *Skelet. Muscle* 6,.
147. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* 43, D789-98.
148. McKusick, V.A. (2007). Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.* 80, 588–604.
149. Amberger, J., Bocchini, C.A., Scott, A.F., and Hamosh, A. (2008). McKusick's Online Mendelian Inheritance in Man (OMIM Õ ). *Nucleic Acids Res.* 37, 793–796.
150. Yujian, L., and Bo, L. (2007). A normalized Levenshtein distance metric. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 1091–1095.
151. Soboloff, J., Spassova, M.A., Tang, X.D., Hewavitharana, T., Xu, W., and Gill, D.L. (2006). Orai1 and STIM reconstitute store-operated calcium channel function. *J. Biol. Chem.* 281, 20661–20665.
152. Lewis, R.S. (2001). CALCIUM SIGNALING MECHANISMS IN T LYMPHOCYTES.
153. Partisetis, M., Le Deist, F., Hivroz8, C., Fischeri, M., Korn, H., and Choquets, D. (1994). THE JOURNIU. OF BIOLOGICAL CHEMISTRY The Calcium Current Activated by T Cell Receptor and Store Depletion in Human Lymphocytes Is Absent in a Primary Immunodeficiency\*.
154. Lioudyno, M.I., Kozak, J.A., Penna, A., Safrina, O., Zhang, S.L., Sen, D., Roos, J., Stauderman, K.A., Cahalan, M.D., and Tsien, R.Y. (2008). Orai1 and STIM1 move to the immunological synapse and are up-regulated during T cell activation.

155. Lacruz, R.S., and Feske, S. (2015). Diseases caused by mutations in *ORAI1* and *STIM1*. *Ann. N. Y. Acad. Sci.* *1356*, 45–79.
156. McCarl, C.A., Picard, C., Khalil, S., Kawasaki, T., Röther, J., Papolos, A., Kutok, J., Hivroz, C., LeDeist, F., Plogmann, K., et al. (2009). ORAI1 deficiency and lack of store-operated Ca<sup>2+</sup> entry cause immunodeficiency, myopathy, and ectodermal dysplasia. *J. Allergy Clin. Immunol.* *124*,.
157. Kucuk, Z.Y., Blesing, J.J., Marsh, R., Zhang, K., Davies, S., and Filipovich, A.H. (2016). A challenging undertaking: Stem cell transplantation for immune dysregulation, polyendocrinopathy, enteropathy, X-linked (IPEX) syndrome. *J. Allergy Clin. Immunol.* *137*, 953-955.e4.
158. Picard, C., McCarl, C.-A., Papolos, A., Khalil, S., Lüthy, K., Hivroz, C., Ledest, F., Rieux-Laucat, F., Rechavi, G., Rao, A., et al. (2009). STIM1 Mutation Associated with a Syndrome of Immunodeficiency and Autoimmunity.
159. Feske, S., Gwack, Y., Prakriya, M., Srikanth, S., Puppel, S.-H., Tanasa, B., Hogan, P.G., Lewis, R.S., Daly, M., and Rao, A. (2006). A mutation in *Orai1* causes immune deficiency by abrogating CRAC channel function.
160. Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* *461*, 272–276.
161. Ionita-Laza, I., Makarov, V., Yoon, S., Raby, B., Buxbaum, J., Nicolae, D.L., and Lin, X. (2011). ARTICLE Finding Disease Variants in Mendelian Disorders By Using Sequence Data: Methods and Applications. *Am. J. Hum. Genet.* *89*, 701–712.
162. Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D.,



Shannon, P.T., Jabs, E.W., Nickerson, D.A., et al. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* *42*, 30–35.

163. Boycott, K.M., Rath, A., Chong, J.X., Hartley, T., Alkuraya, F.S., Baynam, G., Brookes, A.J., Brudno, M., Carracedo, A., Den Dunnen, J.T., et al. (2017). COMMENTARY International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases.

164. Chong, J.X., Buckingham, K.J., Jhangiani, S.N., Boehm, C., Sobreira, N., Smith, J.D., Harrell, T.M., Mcmillin, M.J., Wiszniewski, W., Gambin, T., et al. (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities.

165. Boycott, K.M., Hartley, T., Biesecker, L.G., Gibbs, R.A., Innes, A.M., Riess, O., Belmont, J., Dunwoodie, S.L., Jojic, N., Lassmann, T., et al. (2019). A Diagnosis for All Rare Genetic Diseases: The Horizon and the Next Frontiers. *Cell* *177*, 32–37.

166. Gahl, W.A., Wise, A.L., and Ashley, E.A. (2015). The Undiagnosed Diseases Network of the National Institutes of Health. *JAMA* *314*, 1797.

167. Niroula, A., and Vihinen, M. (2016). Variation Interpretation Predictors: Principles, Types, Performance, and Choice (John Wiley and Sons Inc.).

168. Tang, H., and Thomas, P.D. (2016). Tools for predicting the functional impact of nonsynonymous genetic variation. *Genetics* *203*, 635–647.

169. Peterson, T.A., Doughty, E., and Kann, M.G. (2013). Towards precision medicine: Advances in computational approaches for the analysis of human variants. *J. Mol. Biol.* *425*, 4047–4063.

170. Thusberg, J., Olatubosun, A., and Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.* *32*, 358–368.

171. Niroula, A., and Vihinen, M. (2017). Predicting Severity of Disease-Causing Variants.

Hum. Mutat. 38, 357–364.

172. Riera, C., Padilla, N., and de la Cruz, X. (2016). The Complementarity Between Protein-Specific and General Pathogenicity Predictors for Amino Acid Substitutions. *Hum. Mutat.* 37, 1013–1024.

173. Grimm, D.G., Azencott, C.A., Aicheler, F., Gieraths, U., Macarthur, D.G., Samocha, K.E., Cooper, D.N., Stenson, P.D., Daly, M.J., Smoller, J.W., et al. (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.* 36, 513–523.

174. Vihinen, M. (2021). Functional effects of protein variants. *Biochimie* 180, 104–120.

175. Niroula, A., and Vihinen, M. (2019). How good are pathogenicity predictors in detecting benign variants? *PLoS Comput. Biol.* 15,.

176. Delsuc, M., Vitorino, M., and Kieffer, B. (2020). Determination of Protein Structure and Dynamics by NMR. In *Structural Biology in Drug Discovery*, (Wiley), pp. 295–323.

177. Lyumkis, D. (2019). Challenges and opportunities in cryo-EM single-particle analysis. *J. Biol. Chem.* 294, 5181–5197.

178. Murata, K., and Wolf, M. (2018). Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. *Biochim. Biophys. Acta - Gen. Subj.* 1862, 324–334.

179. Herzik, M.A., Wu, M., and Lander, G.C. (2019). High-resolution structure determination of sub-100 kDa complexes using conventional cryo-EM. *Nat. Commun.* 10, 1–9.

180. Gauto, D.F., Estrozi, L.F., Schwieters, C.D., Effantin, G., Macek, P., Sounier, R., Sivertsen, A.C., Schmidt, E., Kerfah, R., Mas, G., et al. (2019). Integrated NMR and cryo-EM atomic-resolution structure determination of a half-megadalton enzyme complex. *Nat. Commun.* 10, 1–12.

181. Ikeya, T., Güntert, P., and Ito, Y. (2019). Protein structure determination in living cells. *Int. J. Mol. Sci.* *20*,
182. Kuhlman, B., and Bradley, P. (2019). Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.* *20*, 681–697.
183. Gao, W., Mahajan, S.P., Sulam, J., and Gray, J.J. (2020). Deep Learning in Protein Structural Modeling and Design. *Patterns* *1*, 100142.
184. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., De Beer, T.A.P., Rempfer, C., Bordoli, L., et al. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* *46*, W296–W303.
185. Vetri, L., Cali, F., Vinci, M., Amato, C., Roccella, M., Granata, T., Freri, E., Solazzi, R., Romano, V., and Elia, M. (2020). A de novo heterozygous mutation in KCNC2 gene implicated in severe developmental and epileptic encephalopathy. *Eur. J. Med. Genet.* *63*, 103848.
186. Berg, A.T., Levy, S.R., and Testa, F.M. (2018). Evolution and course of early life developmental encephalopathic epilepsies: Focus on Lennox-Gastaut syndrome. *Epilepsia* *59*, 2096–2105.
187. Berg, A.T., Mahida, S., and Poduri, A. (2021). KCNQ2-DEE: developmental or epileptic encephalopathy? *Ann. Clin. Transl. Neurol.* *8*, 666–676.
188. McTague, A., Howell, K.B., Cross, J.H., Kurian, M.A., and Scheffer, I.E. (2016). The genetic landscape of the epileptic encephalopathies of infancy and childhood. *Lancet Neurol.* *15*, 304–316.
189. Claes, L., Del-Favero, J., Ceulemans, B., Lagae, L., Van Broeckhoven, C., and De Jonghe, P. (2001). De novo mutations in the sodium-channel gene SCN1A cause severe myoclonic epilepsy of infancy. *Am. J. Hum. Genet.* *68*, 1327–1332.

190. Allen, N.M., Weckhuysen, S., Gorman, K., King, M.D., and Lerche, H. (2020). Genetic potassium channel-associated epilepsies: Clinical review of the Kv family. *Eur. J. Paediatr. Neurol.* *24*, 105–116.
191. Rudy, B., and McBain, C.J. (2001). Kv3 channels: Voltage-gated K<sup>+</sup> channels designed for high-frequency repetitive firing. *Trends Neurosci.* *24*, 517–526.
192. MacKinnon, R. (1995). Pore loops: An emerging theme in ion channel structure. *Neuron* *14*, 889–892.
193. Hidalgo, P., and MacKinnon, R. (1995). Revealing the architecture of a K<sup>+</sup> channel pore through mutant cycles with a peptide inhibitor. *Science* (80-. ). *268*, 307–310.
194. Pascual, J.M., Shieh, C.C., Kirsch, G.E., and Brown, A.M. (1995). Multiple residues specify external tetraethylammonium blockade in voltage-gated potassium channels. *Biophys. J.* *69*, 428–434.
195. Kim, D.M., and Nimigean, C.M. (2016). Voltage-gated potassium channels: A structural examination of selectivity and gating. *Cold Spring Harb. Perspect. Biol.* *8*, a029231.
196. Pongs, O. (1993). Shaker related K channels. *Semin. Neurosci.* *5*, 93–100.
197. Yarov-Yarovoy, V., Baker, D., and Catterall, W.A. (2006). Voltage sensor conformations in the open and closed states in ROSETTA structural models of K channels.
198. Long, S.B., Campbell, E.B., and MacKinnon, R. (2005). Voltage sensor of Kv1.2: Structural basis of electromechanical coupling. *Science* (80-. ). *309*, 903–908.
199. Long, S.B., Tao, X., Campbell, E.B., and MacKinnon, R. (2007). Atomic structure of a voltage-dependent K<sup>+</sup> channel in a lipid membrane-like environment. *Nature* *450*, 376–382.
200. Park, J., Koko, M., Hedrich, U.B.S., Hermann, A., Cremer, K., Haberlandt, E., Grimm, M., Alhaddad, B., Beck-Woedl, S., Harrer, M., et al. (2019). KCNC1-related disorders: new de

novo variants expand the phenotypic spectrum. *Ann. Clin. Transl. Neurol.* 6, 1319–1326.

201. Waters, M.F., Minassian, N.A., Stevanin, G., Figueroa, K.P., Bannister, J.P.A., Nolte, D., Mock, A.F., Evidente, V.G.H., Fee, D.B., Müller, U., et al. (2006). Mutations in voltage-gated potassium channel KCNC3 cause degenerative and developmental central nervous system phenotypes. *Nat. Genet.* 38, 447–451.

202. Muona, M., Berkovic, S.F., Dibbens, L.M., Oliver, K.L., Maljevic, S., Bayly, M.A., Joensuu, T., Canafoglia, L., Franceschetti, S., Michelucci, R., et al. (2015). A recurrent de novo mutation in KCNC1 causes progressive myoclonus epilepsy. *Nat. Genet.* 47, 39–46.

203. Rademacher, A., Schwarz, N., Seiffert, S., Pendziwiat, M., Rohr, A., Van Baalen, A., Helbig, I., Weber, Y., and Muhle, H. (2020). Whole-exome sequencing in NF1-related west syndrome leads to the identification of KCNC2 as a novel candidate gene for epilepsy. *Neuropediatrics* 51, 368–372.

204. Schwarz, N., Seiffert, S., Pendziwiat, M., Rademacher, A., Hedrich, U.B., Augustijn, P.B., Baier, H., Bayat, A., Bisulli, F., Buono, R.J., et al. (2021). Karl Martin Klein 19-21 , Ioanna Kousiappa 22 , Wolfram S. 7 Kunz 23 , Holger Lerche 1 , Laura Licchetta 9. Lejla Mulahasanovic 12, 20.

205. Imbrici, P., Grottesi, A., D’Adamo, M.C., Mannucci, R., Tucker, S.J., and Pessia, M. (2009). Contribution of the central hydrophobic residue in the PXP motif of voltage-dependent K<sup>+</sup> channels to S6 flexibility and gating properties. *Channels* 3, 39–45.

206. Rudy, B., Chow, A., Lau, D., Amarillo, Y., Ozaita, A., Saganich, M., Moreno, H., Nadal, M.S., Hernandez-Pineda, R., Hernandez-Cruz, A., et al. (1999). Contributions of Kv3 channels to neuronal excitability. In *Annals of the New York Academy of Sciences*, (New York Academy of Sciences), pp. 304–343.

207. Erisir, A., Lau, D., Rudy, B., and Leonard, C.S. (1999). Function of specific K<sup>+</sup> channels in sustained high-frequency firing of fast-spiking neocortical interneurons. *J. Neurophysiol.* *82*, 2476–2489.
208. Kirsch, G.E., and Drewe, J.A. (1993). Gating-dependent mechanism of 4-aminopyridine block in two related potassium channels. *J. Gen. Physiol.* *102*, 797–816.
209. Alviña, K., and Khodakhah, K. (2010). The therapeutic mode of action of 4-aminopyridine in cerebellar ataxia. *J. Neurosci.* *30*, 7258–7268.
210. Armstrong, C.M., and Loboda, A. (2001). A model for 4-aminopyridine action on K channels: Similarities to tetraethylammonium ion action. *Biophys. J.* *81*, 895–904.
211. Chang, K.W., Yuan, T.C., Fang, K.P., Yang, F.S., Liu, C.J., Chang, C.S., and Lin, S.C. (2003). The increase of voltage-gated potassium channel Kv3.4 mRNA expression in oral squamous cell carcinoma. *J. Oral Pathol. Med.* *32*, 606–611.
212. Choquet, D., and Korn, H. (1992). Mechanism of 4-Aminopyridine Action on Voltage-gated Potassium Channels in Lymphocytes.
213. Van Hoeymissen, E., Held, K., Freitas, A.C.N., Janssens, A., Voets, T., and Vriens, J. (2020). Gain of channel function and modified gating properties in TRPM3 mutants causing intellectual disability and epilepsy. *Elife* *9*, 1–12.
214. Kaczmarek, L.K., and Zhang, Y. (2017). Kv3 channels: Enablers of rapid firing, neurotransmitter release, and neuronal endurance. *Physiol. Rev.* *97*, 1431–1468.
215. Crawford, K., Xian, J., Helbig, K.L., Galer, P.D., Parthasarathy, S., Lewis-Smith, D., Kaufman, M.C., Fitch, E., Ganesan, S., O'Brien, M., et al. (2021). Computational analysis of 10,860 phenotypic annotations in individuals with SCN2A-related disorders. *Genet. Med.* *23*, 1263–1272.

216. Labro, A.J., Raes, A.L., Bellens, I., Ottschytsch, N., and Snyders, D.J. (2003). Gating of Shaker-type Channels Requires the Flexibility of S6 Caused by Prolines. *J. Biol. Chem.* 278, 50724–50731.
217. Yazdani, M., Jia, Z., and Chen, J. (2020). Hydrophobic dewetting in gating and regulation of transmembrane protein ion channels. *J. Chem. Phys.* 153, 110901.
218. Aryal, P., Sansom, M.S.P., and Tucker, S.J. (2015). Hydrophobic gating in ion channels. *J. Mol. Biol.* 427, 121–130.
219. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Res.* 49, D412–D419.
220. Gifford, C.A., Ranade, S.S., Samarakoon, R., Salunga, H.T., Yvanka De Soysa, T., Huang, Y., Zhou, P., Elfenbein, A., Wyman, S.K., Bui, Y.K., et al. (2019). Oligogenic inheritance of a human heart disease involving a genetic modifier. *Science* (80-. ). 364, 865–870.
221. Duerinckx, S., Jacquemin, V., Drunat, S., Vial, Y., Passemard, S., Perazzolo, C., Massart, A., Soblet, J., Racapé, J., Desmyter, L., et al. (2020). Digenic inheritance of human primary microcephaly delineates centrosomal and non-centrosomal pathways. *Hum. Mutat.* 41, 512–524.
222. Chen, Y., Barajas-Martinez, H., Zhu, D., Wang, X., Chen, C., Zhuang, R., Shi, J., Wu, X., Tao, Y., Jin, W., et al. (2017). Novel trigenic CACNA1C/DES/MYPN mutations in a family of hypertrophic cardiomyopathy with early repolarization and short QT syndrome. *J. Transl. Med.* 15,.
223. Yao, Q., Li, E., and Shen, B. (2019). Autoinflammatory disease with focus on NOD2-associated disease in the era of genomic medicine. *Autoimmunity* 52, 48–56.
224. Wallace, M.J., El Refaey, M., Mesirca, P., Hund, T.J., Mangoni, M.E., and Mohler, P.J.

- (2021). Genetic Complexity of Sinoatrial Node Dysfunction. *Front. Genet.* *12*,.
225. Monasky, M.M., Micaglio, E., Ciconte, G., and Pappone, C. (2020). Brugada Syndrome: Oligogenic or Mendelian Disease? *Int. J. Mol. Sci.* *21*, 1687.
226. Nijman, S.M.B. (2011). Synthetic lethality: general principles, utility and detection using genetic screens in human cells. *FEBS Lett.* *585*, 1–6.
227. Srivas, R., Shen, J.P., Yang, C., Aza-Blanc, P., Sobol, R.W., and Correspondence, T.I. (2016). A Network of Conserved Synthetic Lethal Interactions for Exploration of Precision Cancer Therapy. *Mol. Cell* *63*, 514–525.
228. O’Neil, N.J., Bailey, M.L., and Hieter, P. (2017). Synthetic lethality and cancer. *Nat. Rev. Genet.* *18*, 613–623.
229. Guo, J., Liu, H., and Zheng, J. (2015). SynLethDB: synthetic lethality database toward discovery of selective and sensitive anticancer drug targets. *Nucleic Acids Res.* *44*, 1011–1017.
230. Gong, X., Du, J., Parsons, S.H., Merzoug, F.F., Webster, Y., Iversen, P.W., Chio, L.-C., Van Horn, R.D., Lin, X., Blosser, W., et al. (2018). Aurora A Kinase Inhibition Is Synthetic Lethal with Loss of the RB1 Tumor Suppressor Gene.
231. Li, X., O’neil, N.J., Moshgabadi, N., and Hieter, P. (2014). Synthetic Cytotoxicity: Digenic Interactions with TEL1/ATM Mutations Reveal Sensitivity to Low Doses of Camptothecin.
232. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* *4*, 1073–1081.
233. Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. (2003). ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information (Valdar and Thornton).
234. Celniker, G., Nimrod, G., Ashkenazy, H., Glaser, F., Martz, E., Mayrose, I., Pupko, T., and



Ben-Tal, N. ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function.

235. Ashkenazy, H., Erez, E., Martz, E., Pupko, T., and Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids.

236. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. *Nat Methods* 7, 248–249.

237. Kircher, M., Witten, D.M., Jain, P., O’roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46, 310–315.

238. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 47, D886–D894.

239. Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., and Liu, X. (2014). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies.

240. Castellana, S., and Mazza, T. Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools.

241. Lev, B., Murail, S., Poitevin, F., Cromer, B.A., Baaden, M., Delarue, M., and Allen, T.W. (2017). String method solution of the gating pathways for a pentameric ligand-gated ion channel. *Proc. Natl. Acad. Sci. U. S. A.* 114, E4158–E4167.

242. Ruepp, A., Waegle, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.-W. (2010). CORUM: the comprehensive resource

- of mammalian protein complexes—2009. *Nucleic Acids Res.* 38, D497–D501.
243. Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegle, B., Schmidt, T., Doudieu, O.N., Stumpflen, V., et al. (2007). CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.* 36, D646–D650.
244. Turner, B., Razick, S., Turinsky, A.L., Vlasblom, J., Crowdy, E.K., Cho, E., Morrison, K., Donaldson, I.M., and Wodak, S.J. (2010). iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database* 2010, baq023–baq023.
245. Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., et al. (2017). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368.
246. Nédélec, Y., Sanz, J., Baharian, G., Szpiech, Z.A., Pacis, A., Dumaine, A., Grenier, J.-C., Freiman, A., Sams, A.J., Hebert, S., et al. (2016). Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. *Cell* 167, 657-669.e21.
247. Capra, J.A., Williams, A.G., and Pollard, K.S. (2012). ProteinHistorian: Tools for the Comparative Analysis of Eukaryote Protein Origin. *PLoS Comput. Biol.* 8, e1002567.
248. Chen, W.-H., Minguez, P., Lercher, M.J., and Bork, P. (2012). OGEE: an online gene essentiality database. *Nucleic Acids Res.* 40, D901–D906.
249. Fadista, J., Oskolkov, N., Hansson, O., Groop, L., and Hancock, J. (2016). LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals. *Bioinformatics* 33, btv602.
250. Chen, W.-H., Lu, G., Chen, X., Zhao, X.-M., and Bork, P. (2017). OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human

cancer cell lines. *Nucleic Acids Res.* 45, D940–D944.

251. Huang, N., Lee, I., Marcotte, E.M., and Hurles, M.E. (2010). Characterising and Predicting Haploinsufficiency in the Human Genome. *PLoS Genet.* 6, e1001154.

252. Matsuya, A., Sakate, R., Kawahara, Y., Koyanagi, K.O., Sato, Y., Fujii, Y., Yamasaki, C., Habara, T., Nakaoka, H., Todokoro, F., et al. (2007). Evola: Ortholog database of all human genes in H-InvDB with manual curation of phylogenetic trees. *Nucleic Acids Res.* 36, D787–D792.

253. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

254. Abdallah, A.M., Justin Carlus, S., Al-Mazroea, A.H., Alluqmani, M., Almohammadi, Y., Bhuiyan, Z.A., and Al-Harbi, K.M. (2019). Digenic inheritance of LAMA4 and MYH7 mutations in patient with infantile dilated cardiomyopathy. *Med.* 55, 1–10.

255. Breiman, L. (2004). Consistency for a simple model of random forests. Tech. Rep. 670. Stat. Dep. Univ. Calif. Berkeley 10.

256. Lek, M., Karczewski, K.J., Minikel, E. V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291.

257. (2016). ExAC project pins down rare gene variants. *Nature* 536, 249.

258. Sim, N.L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P.C. (2012). SIFT web server: Predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* 40,.

259. Vaser, R., Adusumalli, S., Ngak Leng, S., Sikic, M., and Ng, P.C. (2015). SIFT missense predictions for genomes. *Nat. Protoc.* 11,.

260. Kircher, M., Witten, D.M., Jain, P., O’Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* *46*, 310–315.
261. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* *20*, 110–121.
262. Liao, Y., Wang, J., Jaehnig, E.J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* *47*, 199–205.
263. Lemay, J.K., Weitzner, B.D., Lewis, S.M., Adolf-Bryfogle, J., Alam, N., Alford, R.F., Aprahamian, M., Baker, D., Barlow, K.A., Barth, P., et al. (2020). Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat. Methods* *17*, 665–680.
264. Barth, P., Wallner, B., and Baker, D. (2009). Prediction of membrane protein structures with complex topologies using limited constraints. *Proc. Natl. Acad. Sci. U. S. A.* *106*, 1409–1414.
265. DiMaio, F., Leaver-Fay, A., Bradley, P., Baker, D., and André, I. (2011). Modeling symmetric Macromolecular structures in Rosetta3. *PLoS One* *6*, 20450.
266. Wu, E.L., Cheng, X., Jo, S., Rui, H., Song, K.C., Dávila-Contreras, E.M., Qi, Y., Lee, J., Monje-Galvan, V., Venable, R.M., et al. (2014). CHARMM-GUI membrane builder toward realistic biological membrane simulations. *J. Comput. Chem.* *35*, 1997–2004.
267. Lomize, M.A., Pogozheva, I.D., Joo, H., Mosberg, H.I., and Lomize, A.L. (2012). OPM database and PPM web server: Resources for positioning of proteins in membranes. *Nucleic Acids Res.* *40*,
268. Maier, J.A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K.E., and Simmerling, C. (2015). ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters

- from ff99SB. *J. Chem. Theory Comput.* *11*, 3696–3713.
269. Case, D.A., Betz, R.M., Cerutti, D.S., Cheatham III, T.E., Darden, T.A., Duke, R.E., Giese, T.J., Gohlke, H., Goetz, A.W., Homeyer, N., et al. (2016). Principal contributors to the current codes : Amber 2016 Reference Manual. AMBER 2016, Univ. California, San Fr.
270. Ryckaert, J.-P., Ciccotti, G., and Berendsen, H.J.C. (1977). Numerical integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* *23*, 321–341.
271. Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: An  $N \cdot \log(N)$  method for Ewald sums in large systems. *J. Chem. Phys.* *98*, 10089–10092.
272. Roe, D.R., and Cheatham, T.E. (2013). PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* *9*, 3084–3095.
273. Smart, O.S., Neduelil, J.G., Wang, X., Wallace, B.A., and Sansom, M.S.P. (1996). HOLE: A program for the analysis of the pore dimensions of ion channel structural models. *J. Mol. Graph.* *14*, 354–360.
274. Fowler, P.W., and Sansom, M.S.P. (2013). The pore of voltage-gated potassium ion channels is strained when closed. *Nat. Commun.* *4*, 1–8.
275. Kumar, S., Rosenberg, J.M., Bouzida, D., Swendsen, R.H., and Kollman, P.A. (1992). THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* *13*, 1011–1021.
276. Grossfield, A. WHAM: the weighted histogram analysis method.

## LIST OF FIGURES

<b>Figure 1:</b> Network and Functional Features (NFFs) used for machine-learning-based identification of digenic disease gene pairs. _____	<b>23</b>
<b>Figure 2:</b> Additional feature sets used for machine learning classification of digenic diseases. _____	<b>24</b>
<b>Figure 3:</b> Positive and negative training sets used for classification of digenic disease genes. _____	<b>25</b>
<b>Figure 4:</b> Schematic of the protocol for training and evaluating the DiGePred digenic disease pair classifier. _____	<b>26</b>
<b>Figure 5:</b> Distribution of network features is different for digenic and non-digenic pairs; similar for matched gene pairs. _____	<b>27</b>
<b>Figure 6:</b> Random forest classifiers can accurately distinguish digenic and non-digenic gene pairs using different feature sets. _____	<b>29</b>
<b>Figure 7:</b> Classifier accurately identified digenic pairs from all non-digenic gene pairs using various feature sets; addition of features improved performance. _____	<b>30</b>
<b>Figure 8:</b> Phenotype features were most important for classifier to identify digenic pairs; evolutionary features more important for matched pairs. _____	<b>33</b>
<b>Figure 9:</b> Comparison of feature importance ranks measured using GINI and permutation OOB error approaches. _____	<b>34</b>
<b>Figure 10:</b> Classifiers accurately distinguish digenic pairs from non-digenic pairs on held-out test sets. _____	<b>36</b>
<b>Figure 11:</b> DiGePred accurately identifies novel digenic pairs from the recent literature. _____	<b>40</b>

<b>Figure 12:</b> Validation of other models of classifier using novel digenic pairs from recent literature.	41
<b>Figure 13:</b> Leaving out phenotype features reduces DiGePred performance, but it remains strong.	43
<b>Figure 14:</b> DiGePred has a low false positive rate and outperforms a recent digenic gene prediction method.	47
<b>Figure 15:</b> Low false positive rate of classifier on external negative test set of gene pairs from unaffected relatives of UDN patients; comparison with recently published variant combination pathogenicity predictor at various prediction thresholds.	49
<b>Figure 16:</b> Low false positive rate of classifier on external negative test set of gene pairs from unaffected relatives of UDN patients; comparison with recently published variant combination pathogenicity predictor using various models of training.	50
<b>Figure 17:</b> Fewer false positives for DiGePred compared to ORVAL for other models of classifier.	51
<b>Figure 18:</b> Fewer false positives for DiGePred compared to ORVAL when genes selected based on predicted deleterious variant effect.	52
<b>Figure 19:</b> Fewer false positives for DiGePred compared to ORVAL when genes selected randomly	53
<b>Figure 20:</b> Fewer false positives for other models of DiGePred compared to ORVAL when genes selected based on predicted deleterious variant effect.	54
<b>Figure 21:</b> Fewer false positives for other models of DiGePred compared to ORVAL when genes selected randomly	55

<b>Figure 22:</b> Low false positive rate of classifier on external negative test set of gene pairs from unaffected relatives of UDN patients; comparison with recently published variant combination pathogenicity predictor using various models of training, when genes selected based on predicted deleterious variant effect.	56
<b>Figure 23:</b> Low false positive rate of classifier on external negative test set of gene pairs from unaffected relatives of UDN patients; comparison with recently published variant combination pathogenicity predictor using various models of training, when genes selected randomly.	57
<b>Figure 24:</b> DiGePred predicts that UDN patients have more digenic gene pairs above high confidence thresholds than unaffected relatives.	59
<b>Figure 25:</b> ORVAL predicts that UDN patients and unaffected relatives have similar numbers of digenic gene pairs at high confidence thresholds.	60
<b>Figure 26:</b> Gene Ontology enrichment for top 100 genes with most predicted digenic pairs.	62
<b>Figure 27:</b> Gene Ontology enrichment for top 100 genes with highest average predicted value.	62
<b>Figure 28:</b> Gene Ontology enrichment for genes in the top 100 gene pairs with highest predicted value	63
<b>Figure 29:</b> Candidate pathogenic variants in KCNC2 are nearby, but have different structural contexts in Kv3.2.	77
<b>Figure 30:</b> KCNC2 variants are located in the evolutionarily conserved hinge domain and predicted to be deleterious.	79
<b>Figure 31:</b> Candidate Kv3.2 variants cause loss and gain of channel function.	79
<b>Figure 32:</b> Candidate KCNC2 variants have different effects on Kv3.2 deactivation tail kinetics.	81



<b>Figure 33:</b> 4-AP differentially blocks variant Kv3.2 channels.	82
<b>Figure 34:</b> Candidate KCNC2 variants modify Kv3.2 expression levels.	84
<b>Figure 35:</b> The channel pore of Kv3.2 V469L becomes constricted in MD simulations whereas the pore of V471L adopts a stable open conformation.	86
<b>Figure 36:</b> Distribution of $\phi$ and $\psi$ backbone angles of pore helix residues 461 to 478 sampled in MD simulation of Kv3.2 WT, V469L, and V471L.	87
<b>Figure 37:</b> Adjacent pore helices in Kv3.2 form a contact between residues 471 and 480.	88
<b>Figure 38:</b> V469L increases the energy required for K <sup>+</sup> ion transfer through the cytosolic gate of Kv3.2 compared to WT and V471L.	89

## LIST OF TABLES

<b>TABLE T1:</b> Novel digenic pairs from recent literature, not in Digenic Database (DIDA) and not used for training.	<b>38</b>
<b>TABLE T2:</b> Gene Ontology enrichment for top 100 genes with most predicted digenic pairs.	<b>64</b>
<b>TABLE T3:</b> Gene Ontology enrichment for top 100 genes with highest average predicted value.	<b>66</b>
<b>TABLE T4:</b> Gene Ontology enrichment for genes in the top 100 genes with highest average predicted value.	<b>67</b>
<b>TABLE T5:</b> Summary of $\phi$ and $\psi$ backbone angles of pore helix residues 463 to 475 sampled in MD simulation of Kv3.2 WT, V469L, and V471L.	<b>90</b>

## LIST OF VIDEOS

**Video V1:** WT Kv3.2 tetrameric protein structure side view

**Video V2:** WT Kv3.2 tetrameric protein structure bottom view

**Video V3:** p.V469L variant Kv3.2 tetrameric protein structure side view

**Video V4:** p.V469L variant Kv3.2 tetrameric protein structure bottom view

**Video V5:** p.V471L variant Kv3.2 tetrameric protein structure side view

**Video V6:** p.V471L variant Kv3.2 tetrameric protein structure bottom view

## LIST OF DATASETS

### **DATASET D1:** Held-out digenic gene pairs

DiGePred predictions on held-out digenic gene pairs from DIDA (n=28). These pairs were not used for training and used to test the trained classifier.

<https://vanderbilt.box.com/s/ufsbb48tnkkz5tkckfay23pmrkn3qfqq>

### **DATASET D2:** Novel digenic gene pairs from recent literature

DiGePred predictions on novel digenic gene pairs from recent literature (n=13). These pairs were not included in DIDA.

<https://vanderbilt.box.com/s/3cq5f51dl4h8h8w8hmnd8rmesj6in1hm>

### **DATASET D3:** Predicted digenic pairs with highest confidence from all possible gene pairs

Gene pairs predicted to be digenic by DiGePred at most confident threshold. (n=54,318)

<https://vanderbilt.box.com/s/n1nzdyj8i5fa55vultyq4xn6rsp792a7>

### **DATASET D4A:** Digenic predictions on all human gene pairs

<https://vanderbilt.box.com/s/459ethsqv339nqiarhm0j227jdjb0whq>

### **DATASET D4B:** Digenic predictions on all human gene pairs

<https://vanderbilt.box.com/s/acdqvjuihj3932c6msi5py82rvr5kam3>

### **DATASET D4C:** Digenic predictions on all human gene pairs

<https://vanderbilt.box.com/s/kb3vzubfxjctxt8x0y1vytu59x8r8no>

### **DATASET D4D:** Digenic predictions on all human gene pairs