Biophysical Journal, Volume *120*

Supplemental Information

Predicting Genetic Variation Severity Using Machine Learning to Interpret Molecular Simulations

Matthew D. McCoy, John Hamre III, Dmitri K. Klimov, and M. Saleet Jafri

Supplement

The supplemental materials provide the following information to help clarify topics described in the main manuscript:

- Calmodulin physiology
- Supplemental Figure S1. Pairs plot of the calmodulin
- Supplemental Figure S2. Random forest output for calmodulin
- Aβ Physiology
- Supplemental Figure S3. Pairs plot of the Aβ
- Supplemental Figure S4. Random forest output for Aβ
- Supplemental Figure S5. Confusion matrices of MDPPM for Cam and Aβ

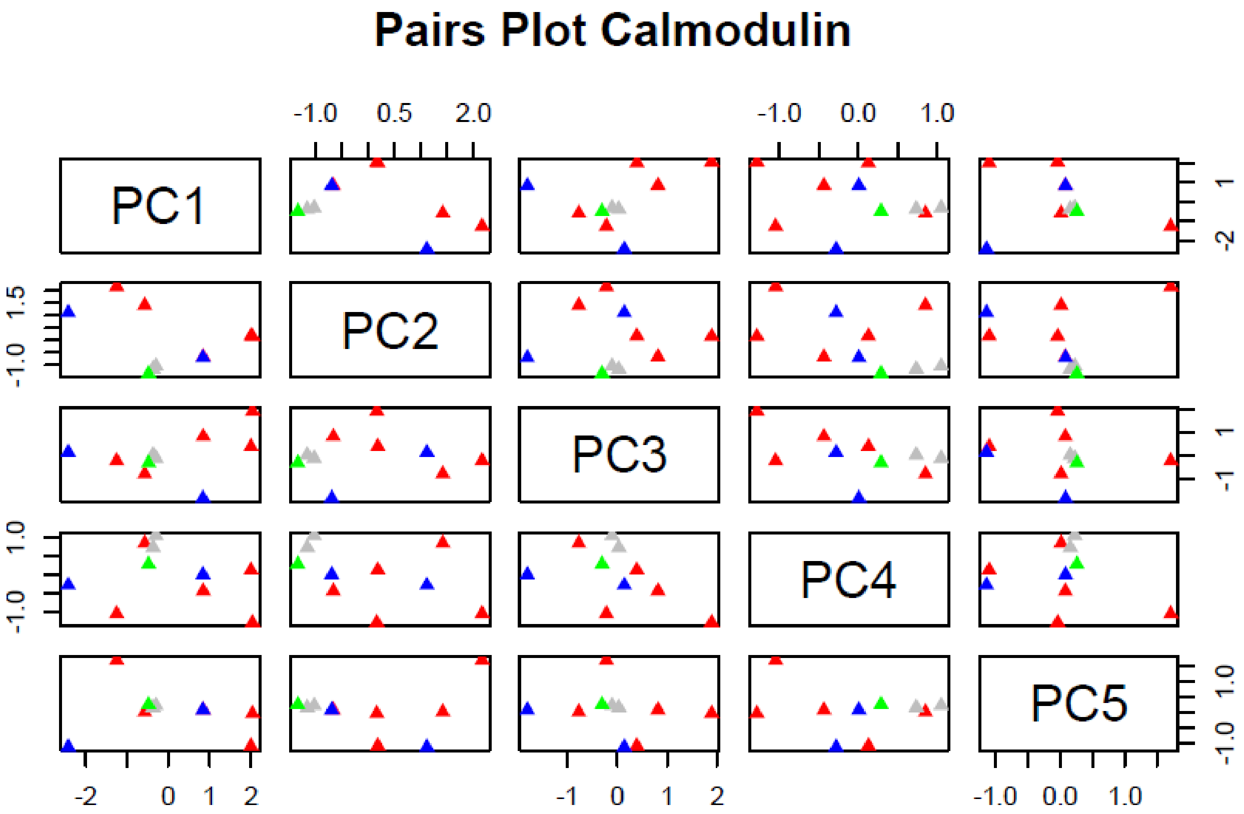
Calmodulin Physiology

CaM is a 149 amino acid, cytoplasmic protein which regulates many calcium dependent processes in the cell. Four Ca^{2+} binding, EF-hand domains are evenly distributed into two globular lobes joined by a flexible linker helix. The structural organization allows for enormous conformational variability, as the two lobes are able to twist with respect to each other and collapse of the central linker allows the two lobes to form a compact structure. The conformational dynamics are regulated by binding calcium ions and interacting protein targets and allow for a wide range of calcium dependent regulatory functions. With hundreds of protein regulatory targets and cooperative calcium binding affinities across a wide range of physiological Ca^{2+} concentrations, there are numerous mechanisms where CaM binding contributes to stabilizing functional conformations of protein interaction partners (46).

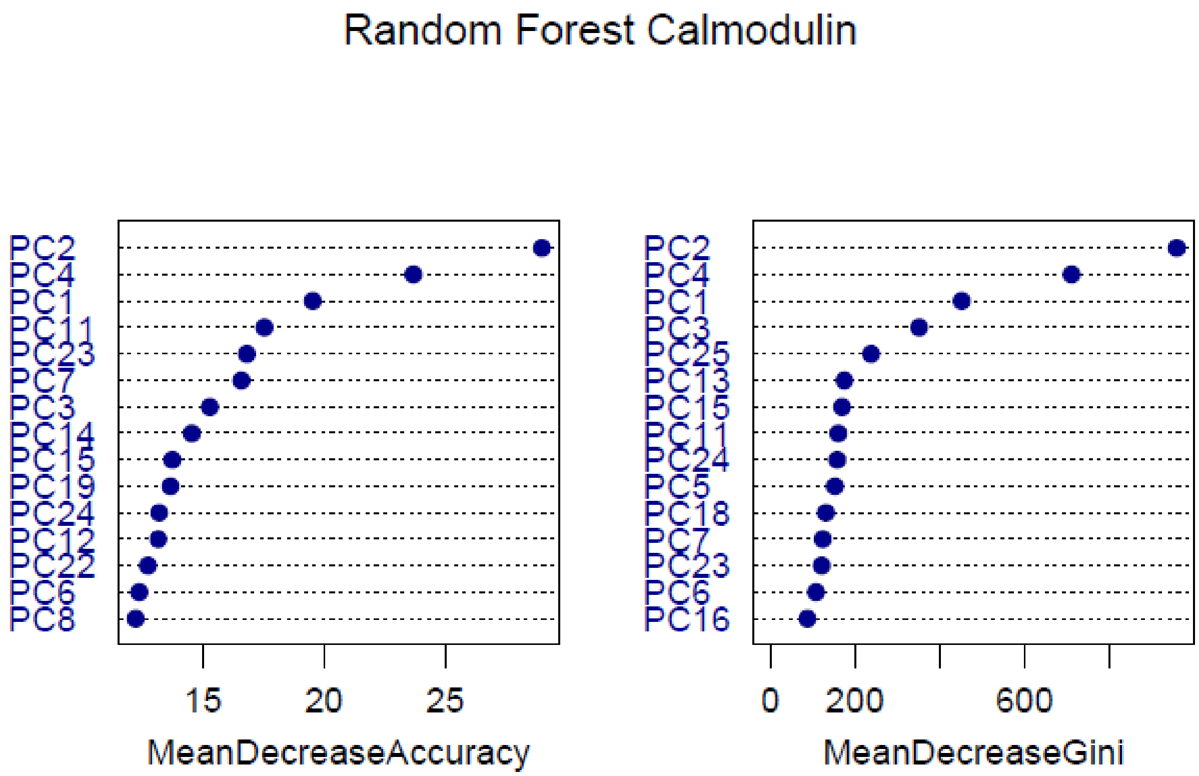
In the cardiac myocyte, CaM plays a crucial role in regulating the coupling of cell excitation to contraction at the dyad, a specialized subcellular structure where L-type Ca^{2+} channels (Cav1.2) on the cell membrane are closely apposed to arrays of calcium sensitive ryanodine receptor 2 (RYR2) on the sarcoplasmic reticulum. Depolarization opens Cav1.2, allowing Ca^{2+} influx to trigger RYR2 channel opening thereby amplifying the calcium signal to activate the contractile machinery. The spike in Ca^{2+} also regulates the CaM dependent inactivation (CDI) of Cav1.2, attenuating Ca^{2+} entry. The functional impairment of these proteins will increase the risk of arrhythmia, as is the case in Long QT Syndrome (LQTS) and Catecholaminergic Polymorphic Ventricular Tachycardia (CPVT). Novel, single missense mutations in CaM have been associated with LQTS and CPVT phenotypes, highlighting the sensitivity of CaM regulatory function to changes in the sequence.

Structurally, CaM binds to a cytoplasmic IQ domain on Cav1.2 even at low concentrations. Upon Ca^{2+} binding, conformational changes in CaM induce channel closing at increased rates compared to thermodynamic or transmembrane voltage-mediated processes. A leading mechanistic hypothesis of CaM dependent inactivation involves dissociation of the CaM N-lobe from the IQ domain upon increasing Ca^{2+} concentrations and the subsequent binding to the cytoplasmic Cav1.2 NSCaTE domain. Presumably, the linking of these two domains by Ca^{2+} bound CaM imparts a mechanistic force on the channel to promote a closed conformation (47).

Some CaM variants disrupt calcium binding and alter CDI of the L-Type Ca^{2+} channel, leading to the LQTS phenotype. Other variants have been observed to affect CaM interactions with other important proteins, including binding to RYR2 and sodium channels, the former resulting in altered cardiac myocyte functions associated with CPVT. Predicting the potential for a specific variant to have a physiological consequence is a major challenge in the age of high throughput sequencing, especially in diseases that arise from subtle changes to molecular interactions. Experimental evaluation of all potential variants is intractable, but as our work shows, advanced computational analysis may enable the *a priori* prediction of the functional impact of genomic variation.



Supplemental Figure S1 - Pairs plot of Calmodulin PC's. Pairs of PCs are derived from either the second PCA analysis (shown here) or the aggreagate centroid from the first PCA, and chosen visually to resolve which are effective at separating the data. Resulting PC's are used for the KNN analysis (iii) to determine variant pathogenicity or the phenotype in a "yes/no" verdict. (CPVT red; LQTS blue; negative control grey; wild-type green).



Supplemental Figure S2. Random forest output for the first PCA analysis for Calmodulin. The PC's are used in the KNN analysis for two-out accuracy measurements (ii) on the entire simulation trajectory to derive confusion matrices. A "yes/no" verdict is made on every PDB file, as opposed to the single, averaged variant used to determine pathogenicity and phenotype. The left panel refers to the accuracy during the calculation phase and the gini (right panel) refers to how homogenous the data is, measuring each PC contribution to variance

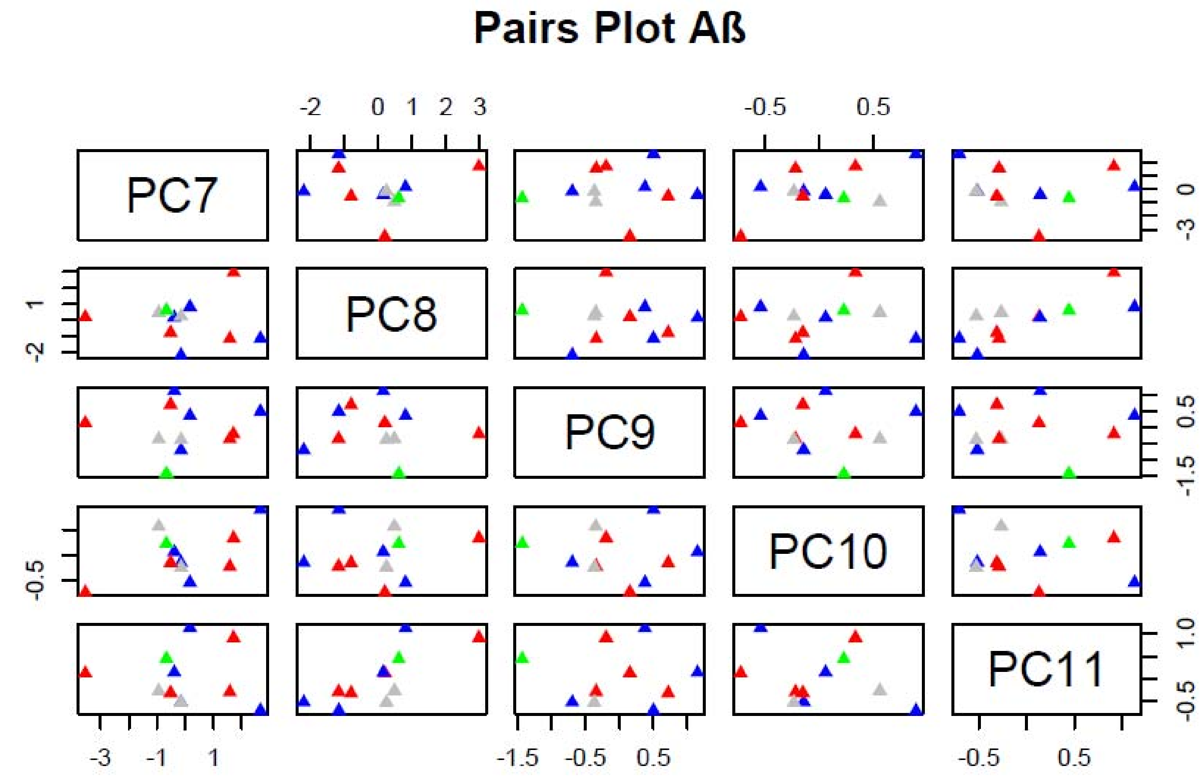
Aβ Physiology

Missense variants in amyloid precursor protein (APP) have been associated with two neurodegenerative diseases: Familial Alzheimer's disease (AD1) and Cerebral Amyloid Angiopathy (CAA). Currently, there are more than 30 common mutations in the APP gene that have been described, and approximately 25 of those are associated with the pathogenesis of Alzheimer's disease (AD) (50). Aβ is derived from the β-amyloid precursor protein (APP) which is cleaved by β and γ secretase to yield the Aβ peptides. Aβ is typically 36 to 43 amino acids in length as a monomer and exists as a random coil, however, aggregated forms found in the brain routinely exhibit stable beta sheet confirmations. Mechanisms whereby the molecules transition from random coil to beta-sheet are not fully understood. Two primary forms of Aβ occur in humans, Aβ40 and Aβ42 (51), though other varieties have been detected in-vivo, such as Aβ39, Aβ37, small fragment isolates from AD brains, Aβ(1-5), Aβ42(6-16) (51), and Aβ 25-35 (52)]. AD is associated with the accumulation of amyloid plaques primarily consisting of aggregated amyloid beta peptide (Aβ40 and Aβ42). CAA has been associated with the accumulation of Aβ peptides (Aβ40) in cerebral arteries.

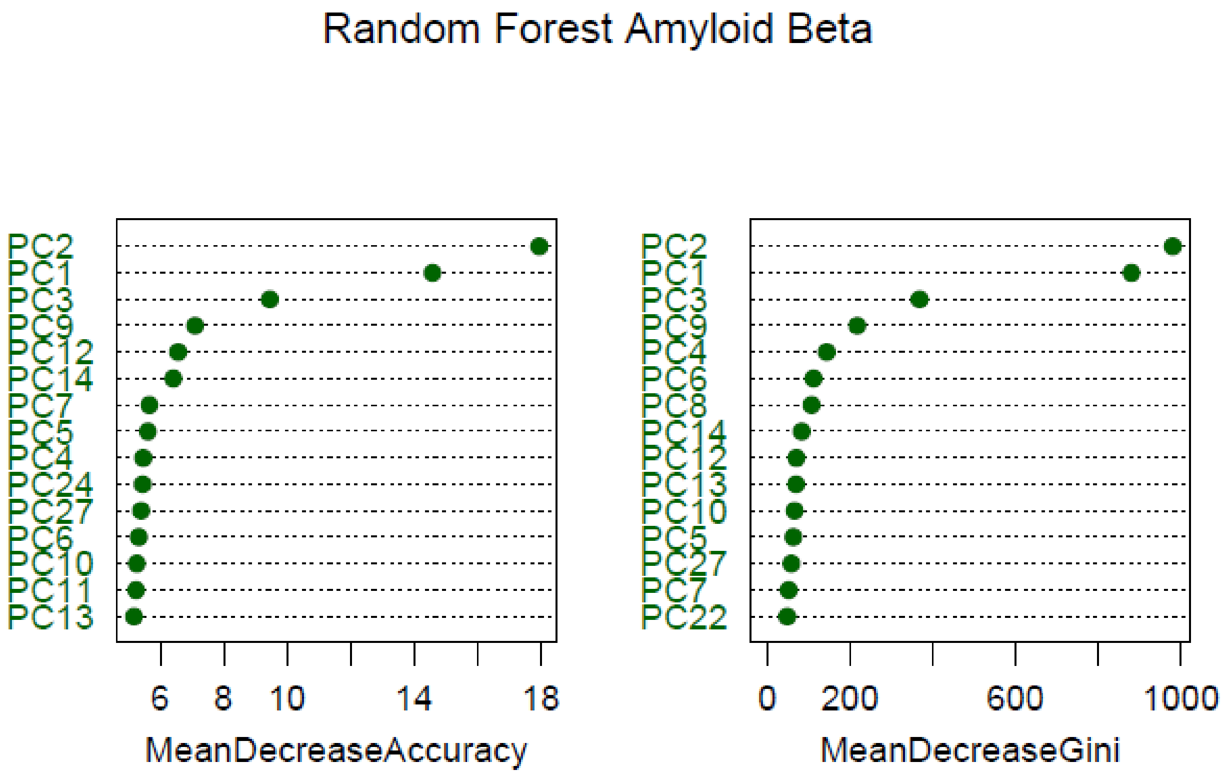
General overview of AD1 variants. The Arctic (E22G) mutation has been well studied and was initially classified from a four-generation family from northern Sweden. Affected individuals with Arctic phenotype characteristically express clinical symptoms early with and rapid cognitive decline (54). The Arctic variant has specifically been shown to form protofibrils at an increased propensity and at a faster rate in contrast to wild-type, and was one of several pathogenic APP mutations found to present resistance to neprilysin-catalyzed proteolysis (54, 55). Neprilysin is thought to be a fundamental component of Aβ proteolysis and therefore Alzheimer's preclusion in the normal brain. Neprilysin-deficient knockout mice convey Alzheimer's-like behavioral impairment and present with Aβ deposition in the brain (56).

Structurally and kinetically, the Tottori mutation (D7N) has been found not to accelerate the production of Aβ peptides but does in fact increase rate of fibril formation (57). The Flemish (A21G), and Arctic (E22G) mutations have been reported to precede implementation of distinct aggregate assemblies (58), and A42T has been shown to aggregate potently with significant cytotoxicity (59).

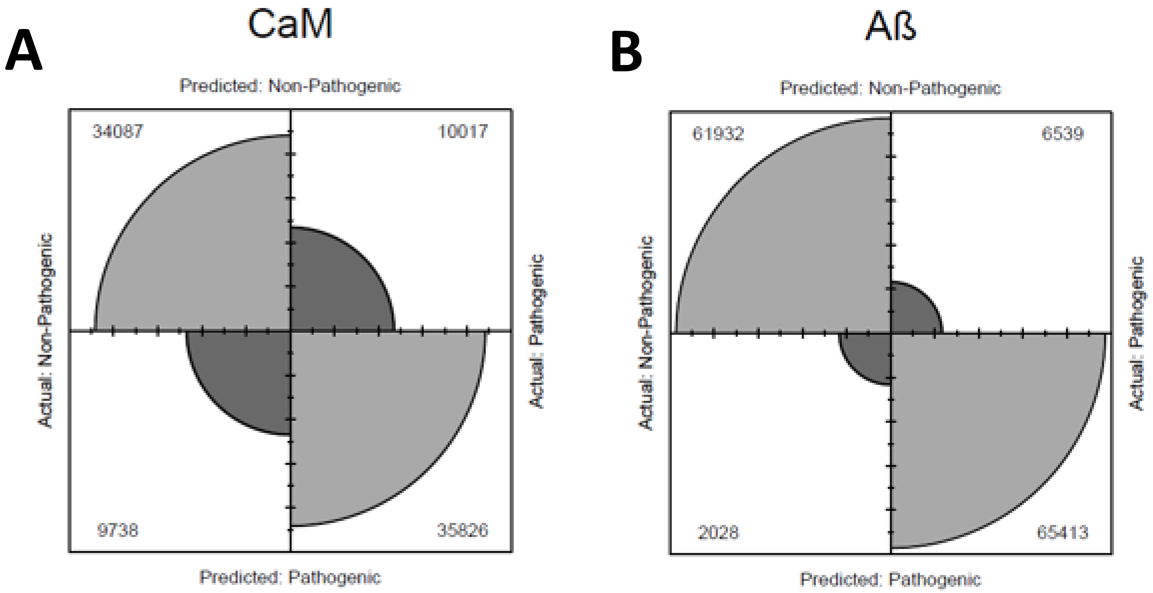
Variants of Aβ *used in this study leading to CAA.* The Italian (E22K), Dutch (E22Q), Iowa (D23N), and recently described Piedmont (L34V) variants all have been linked to CAA. Although the Piedmont L34V variant is located outside the classic Aβ residue positions 21-23 associated with CAA, it shows similar hemorrhagic phenotype, albeit less aggressive than other widely studied variants (62). The Dutch (E22Q) mutation is a vasculotropic variant of Aβ associated with hereditary cerebral hemorrhage with amyloidosis (63). The Italian (E22K) mutation is believed to exert pathogenic effects by inducing the formation of Aβ stable oligomers and protofibrils (58). Recent experimental studies have indicated that the Iowa (D23N) variant increases the risk for the pathogeny of CAA (64). Fibrils of the Iowa variant can form both parallel and antiparallel structures, however, parallel assemblies are most associated with CAA pathogenesis.



Supplemental Figure S3 - Pairs plot of the Amyloid Beta. Pairs of PCs are derived from either the second PCA analysis (shown here) or the aggreagate centroid from the first PCA, and chosen visually to resolve which are effective at separating the data. Resulting PC's are used for the KNN analysis (iii) to determine variant pathogenicity or the phenotype in a "yes/no" verdict. (AD1 red; CAA blue; negative control grey; wild-type green).



Supplemental Figure S4. Random forest output for the first PCA analysis for Amyloid Beta. Random forest output for the first PCA analysis for Calmodulin. The PC's are used in the KNN analysis for two-out accuracy measurements (ii) on the entire simulation trajectory to derive confusion matrices. A "yes/no" verdict is made on every PDB file, as opposed to the single, averaged variant used to determine pathogenicity and phenotype. The left panel refers to the accuracy during the calculation phase and the gini (right panel) refers to how homogenous the data is, measuring each PC contribution to variance



C

	CaM	Aβ
Kappa	**0.5592**	**0.8740**
Sensitivity	0.7778	0.9683
Specificity	0.7815	0.9091
Precision	0.7729	0.9045
Balanced Accuracy	0.7796	0.9387

Figure S5. A. Calmodulin confusion matrix showing results of different permutations of KNN algorithm (ii) pulling two out, training on the remainder and then plugging back with a KNN vote for class for every PDB file within the trajectory. **B.** Aβ confusion matrix also using the same method. **C.** The confusion matrix metrics for both CaM and Aβ. The Kappa statistic is a measure of instances that may have been correctly classified by chance. This is calculated by the total accuracy and the random accuracy. The higher the kappa the more percent of reliable data there is.