Doctoral Dissertation

Estimating cleavage mechanisms of γ-secretase using machine learning

Hideki Ueda

Program of Data Science
Graduate School of Science and Technology Nara
Institute of Science and Technology

Supervisor: Professor Shigehiko Kanaya
Computational Systems Biology Lab.
(Division of Information Science)

Submitted on July 10, 2023

A Doctoral Dissertation
submitted to Graduate School of Science and Technology,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degreeof
Doctor of Engineering

Hideki Ueda

Thesis Committee:
Supervisor Shigehiko Kanaya
(Professor, Division of Information Science)
Keiichi Yasumoto
(Professor, Division of Information Science)
MD.Altaf-Ul-Amin
(Associate Professor, Division of Information Science)
Naoaki Ono
(Associate Professor, Division of Information Science)
Ming Huang
(Assistant Professor, Division of Information Science)

Estimating cleavage mechanisms of γ-secretase using machine learning*

Hideki Ueda

Abstract

γ-Secretase is an intramembranous protease that generates Aβ, a pathogenic molecule in Alzheimer's disease. Although previous studies revealed that γ-secretase cleavage occurs successively and more than one hundred γ-secretase substrates were identified, the underlying mechanisms of the cleavage by γ-secretase are only poorly understood. Understanding the cleavage mechanism and specificity of γ-secretase is essential, as it may contribute to developing new drugs and therapies targeting Alzheimer's disease. In this study, we estimated the number of pockets in the active site of γ-secretase, the amino acid properties which the active site recognizes, and the preferred amino acids for each pocket. Using six pocket models, ten types of peptide properties, and 88 machine learning methods, we exhaustively examined 5,280 regression models trained by the quantitation data of γ-byproduct of Amyloid beta precursor protein cleavage. Using these models, we conducted cleavage site predictions for 35 identified cleavage sites, and obtained a model with the highest prediction accuracy of 85.7%. Notably, cleavage simulations by the best regression model reproduced characteristic cleavages of γ-secretase for APP and Notch1 substrates. Furthermore, in silico cleavage of random peptides revealed amino acid preferences in the cleavage site region of γ-secretase, which we further validated experimentally. We interpreted the model and estimated that the active site of γ-secretase consisted of seven contiguous pockets and the active site recognized amino acid properties associated with protein secondary structure. Further investigation of the model obtained in this study is expected to advance our fundamental understanding of the cleavage mechanism of γ-secretase and provide helpful information for developing γ-secretase inhibitors.

Keywords:

Alzheimer's disease, amyloid-beta, γ-secretase, cleavage site prediction

---

*Doctoral Dissertation, Graduate School of Science and Technology, Nara Institute of Science and Technology, July 10, 2023.

Contents

1. Introduction 1
1.1 Increase in dementia patients 1
1.2 Alzheimer's disease 1
1.3 γ-Secretase 2
1.4 Medications of Alzheimer's disease 4
1.5 Purpose of this study 5

2. Development of a substrate cleavage site prediction model for γ-secretase and estimation of its cleavage mechanism 7
2.1 Introduction 7
2.2 Materials and Methods 7
2.2.1 Training Dataset 7
2.2.2 Pocket model for the cleavage active site 9
2.2.3 Compression of amino acid property information by principal component analysis 9
2.2.3.1 AAindex 9
2.2.3.2 Compression of AAindex by principal component analysis 10
2.2.4 Generation of regression models of γ-secretase cleavage 12
2.2.4.1 Machine learning algorithms 12
2.2.4.2 Generation of regression models 12
2.2.4.3 RMSE 12
2.2.5 Model validation using substrate cleavage endpoint prediction 15
2.2.5.1 Validation dataset 15
2.2.5.2 Prediction of substrate cleavage termination site using regression models 17
2.2.6 Factor loading analysis of AAindex 17
2.3 Results 18
2.3.1 Creation of regression models 18
2.3.2 Validation of regression models 20
2.3.3 Factor loading analysis of PC9 29
2.4 Discussion 31
2.4.1 The most accurate regression model 31
2.4.2 PC9 32

3. Development of a substrate cleavage site prediction model for γ-secretase and estimation of its cleavage mechanism 34
3.1 Introduction 34
3.2 Materials and Methods 34
3.2.1 Visualization of amino acid preference in the cleavage site region 34
3.2.2 Biochemical experiments 36
3.2.2.1 cDNA constructs 36

3.2.2.2 Cell culture 36
3.2.2.3 Generation of APP knockout (KO) cells using CRISPR/Cas9 genome editing 36
3.2.2.4 Immunoblot analysis 37
3.2.2.5 Cell-free γ-secretase assay for the analysis of de novo AICD generation 38
3.2.2.6 MALDI-TOF mass spectrometry (MS) analysis 38
3.3 Results 39
3.4 Discussion 44

4. Conclusion 47

Acknowledgements 48

References 49

List of Figures

1 Two pathways of APP cleavage. 3
2 Pocket model. 9
3 The schema for creating regression models based on γ-byproducts 14
4 RMSE of regression models. 19
5 Percentage of correct predictions of terminal cleavage of regression models . . 21
6 Visualization of *in silico* successive cleavage 24
7 Probability plot of γ-secretase cleavage of APP variant substrates 25
8 Probability plot of γ-secretase cleavage of substrates. 26
9 Prediction and experimental validation of γ-secretase cleavage of rat APLP1. . 28
10 The method of analyzing frequencies of amino acids appearing at each pocket. 35
11 Sequence feature of γ-secretase cleavage extracted from the best model. 41
12 Unpreferable mutations in P1 and P1' of APP cleavage shifted γ-secretase cleavage site. 42
13 Model of substrate cleavage site selection in the catalytic site of γ-secretase. . 46

List of Tables

1 Levels of γ-byproducts used as the training data 8
2 The principal component scores for each PC for the 20 amino acids. 11
3 Machine learning algorithms used in the present study. 13
4 Summary of reported γ-cleavage sites. 16
5 Summary of reported and predicted γ-cleavage sites 22
6 Observed m/z and theoretical molecular mass of rat APLP1β species 29
7 Top 10 positive and negative amino acid indices by the factor analysis of PC9 30
8 Observed m/z and theoretical molecular mass of AICDs 43
9 Observed m/z and theoretical molecular mass of AICDs derived from APP LV49/50IW 43

1. Introduction

1.1 Increase in dementia patients

As the population ages, the number of dementia patients is projected to increase. According to one study, the number of dementia patients estimated at 57.4 million worldwide in 2019 is projected to triple to 152.8 million by 2050 [1]. A study conducted in Japan predicted an increase in the number of patients with dementia. This study projected that the prevalence of dementia in the population aged 65 years and older will exceed 20% by 2030 and 25% by 2035 in 42 prefectures. By 2045, the prevalence of dementia will exceed 25% in all prefectures except Tokyo [2].

The increase in the number of dementia patients will significantly affect not only the patients themselves but also their families. Caring for patients is time-consuming and expensive, and the financial burden is significant. In addition, changes in the patient's self-perception and behavior can significantly stress family relationships. In response to this critical situation, there is a need to establish methods for preventing and treating dementia.

While there are many causes of dementia, the most significant cause is Alzheimer's disease (AD), estimated to be responsible for 60-80% of patients diagnosed with dementia [3]. Therefore, AD is considered a central research target for dementia treatment, and research on its prevention and treatment is underway. Establishing preventive and therapeutic methods for AD is a significant challenge before the further aging of the population.

1.2 Alzheimer's disease

AD is a progressive neurodegenerative disorder of the brain [3]. The exact cause of AD remains unclear. However, various hypotheses regarding its onset have been proposed based on numerous experimental and observational results, and research into the treatment of Alzheimer's has been conducted based on these hypotheses. Among these, the widely accepted mechanism for the onset of AD is the amyloid cascade hypothesis [4]. According to this hypothesis, the pathogenesis of AD is explained as follows. First, Amyloid Beta (Aβ), produced by the cleavage of amyloid precursor protein (APP) on the surface of the brain's nerve cell membranes, accumulates between the brain nerve cells, forming senile plaques. This is followed by the accumulation and phosphorylation of the tau protein, leading to neurofibrillary changes. As a result, neuronal cell death is induced, leading to the onset of AD.

According to this amyloid cascade hypothesis, the fundamental cause of the onset of Alzheimer's is believed to be the accumulation of Aβ. Therefore, inhibiting the accumulation of Aβ may suppress the onset and progression of AD. Drug discovery research based on this idea is currently being conducted, and γ-secretase, the enzyme that cleaves APP to produce Aβ, is the target of inhibitor development.

1.3 γ-Secretase

γ-Secretase is an enzyme that localizes to the plasma membrane as a complex of four subunits: presenilin, Nicastrin, APH-1 (anterior pharynx-defective 1), and PEN-2 (presenilin enhancer 2) [5]. Among these, presenilin is responsible for protein cleavage activity. γ-Secretase cleaves single-pass membrane proteins, and cleavage occurs in the transmembrane region.

γ-Secretase generates Aβ by cleaving the single-pass transmembrane protein APP within the cell membrane. In the process of generating Aβ, γ-secretase performs consecutive cleavages mainly based on a tripeptide cleavage (Fig. 1) [6]. This consecutive cleavage occurs primarily via two pathways. One begins with the ε49 site cleavage and repeats a tripeptide cleavage three times to produce Aβ40 (Aβ40 Line: Aβ49 → Aβ46 → Aβ43 → Aβ40), and the other starts with the ε48 site cleavage and repeats a tripeptide cleavage twice to produce Aβ42 (Aβ42 Line: Aβ48 → Aβ45 → Aβ42). Of these two pathways, the most common is the former Aβ40 line. Cleavage by γ-secretase is not limited to tripeptides; cleavage of tetrapeptides and pentapeptides has been confirmed, and it is known that there are cleavage pathways that produce other Aβs such as Aβ37 and Aβ38 by combining these cleavages [7].

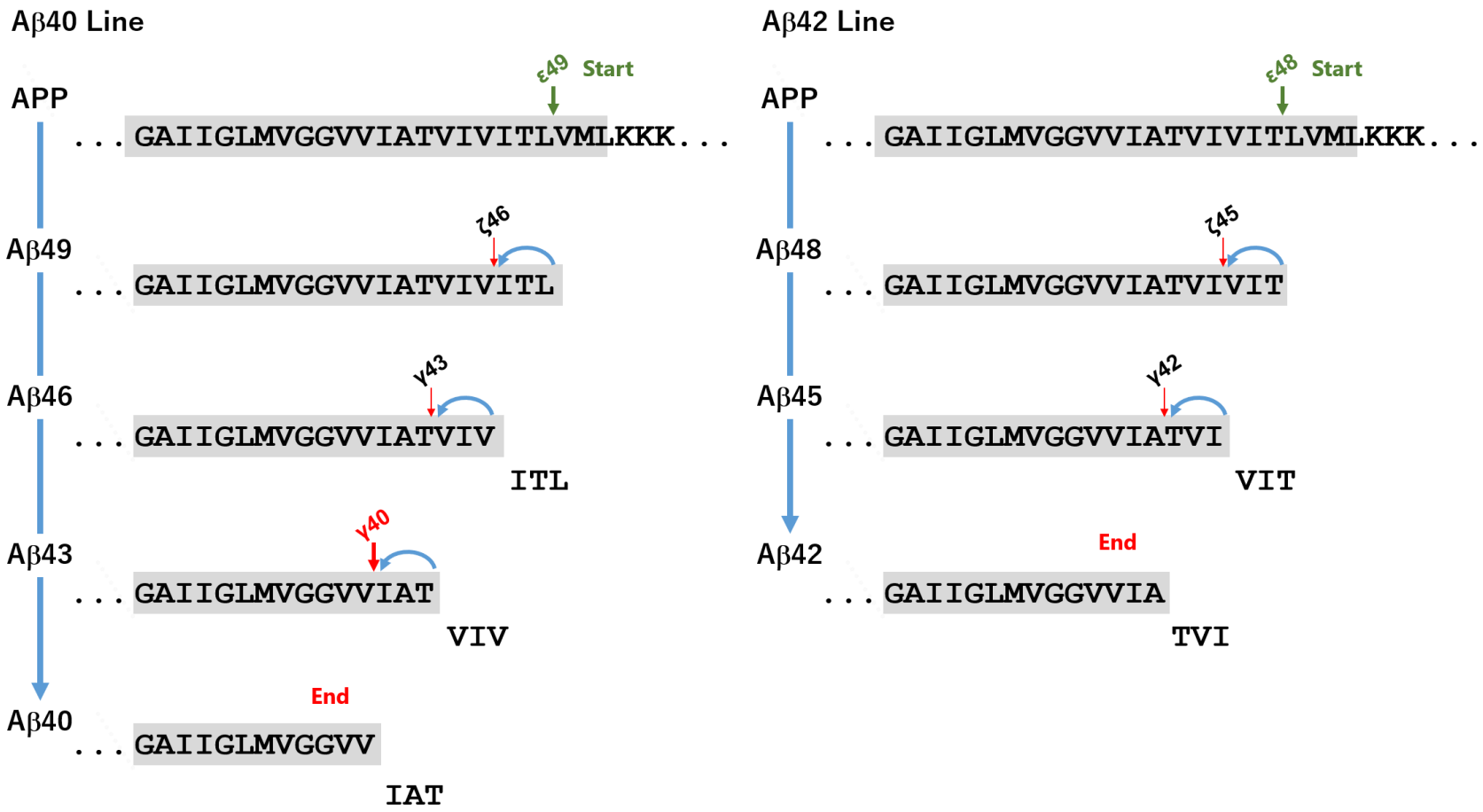

Figure 1. Two pathways of APP cleavage.

The process of Aβ produced from APP by consecutive cleavage by γ-secretase has been clarified. However, the mechanism of γ-secretase cleavage remains unclear. DM Bolduc et al. proposed a model to explain the tripeptide cleavage by γ-secretase based on experiments using several APP mutants in which amino acids in the APP sequence were replaced with aromatic amino acids [8]. According to their model, the binding site of γ-secretase consists of three consecutive pockets, and one in the middle is smaller than the others. They suggest that the size of the amino acids that bind to the pockets determines cleavage. However, cleavage can occur in sequences that do not fit their model, so the cleavage phenomenon of γ-secretase cannot be explained by the size of the amino acids alone. Since enzymatic cleavage is a chemical reaction, it can be thought that not only the size of the amino acids in the sequence but also the physicochemical properties of each amino acid are determinants of cleavage. However, it is still unclear what physicochemical properties contribute to cleavage.

Furthermore, the specificity of the enzyme, that is, what kind of sequence is more likely to be cleaved by γ-secretase, is unknown. Knowing the protein sequence that γ-secretase is likely to cleave may provide insights for designing new preventive strategies. For example, it may be useful for designing specific inhibitors or modulators of γ-secretase. Understanding the cleavage mechanism and specificity of γ-secretase is important, as it may potentially contribute to developing new drugs and therapies targeting AD.

1.4 Medications of Alzheimer's disease

Currently, there are two types of drugs used for AD: cholinesterase inhibitors and NMDA receptor antagonists. One of the factors contributing to the cognitive decline in AD patients is the deficiency of a neurotransmitter called acetylcholine in the brain. Cholinesterase inhibitors suppress the activity of acetylcholinesterase, an enzyme that breaks down acetylcholine, thereby reducing the degradation of acetylcholine in the brain and alleviating cognitive decline [9]. Another cause of cognitive decline in AD patients is neuronal damage induced by excessive stimulation caused by the excitatory neurotransmitter glutamate. NMDA receptor antagonists are used as a treatment for moderate to severe AD, and they are believed to provide neuroprotection by inhibiting the NMDA receptors, which are receptors for the neurotransmitter glutamate in the brain, thus suppressing excessive neuronal excitability [10]. These drugs primarily serve to slow down the progression of symptoms and do not constitute a cure for AD itself.

Although not yet in practical use, there have been efforts to develop drugs to inhibit γ-secretase. Semagacestat and Avagacestat are representative examples of γ-secretase inhibitors, which aim to inhibit the production of amyloid-β peptide. However,

Semagacestat caused an increased risk of skin cancer, associated with inhibition of Notch signaling, and worsened cognitive function, so phase III clinical trials were terminated [11]. Later studies showed that Semagacestat does not inhibit APP cleavage, but is a pseudo-inhibitor that causes increased gamma-byproducts and Aβ accumulation within neuronal cells. Semagacestat-induced cognitive decline may be due to synaptic dysfunction and neuronal loss caused by intra-neuronal Aβ accumulation [12]. Because Semagacestat was a pseudo-inhibitor, it would be a mistake to consider the failure of Semagacestat in clinical trials as the failure of γ-secretase inhibitors. Avagacestat was described as a Notch-sparing inhibitor but was stopped in phase II trials because of an increased risk of skin cancer and gastrointestinal problems [11]. Some have reported that Avagacestat actually lacks Notch-sparing properties, so future inhibitor-based treatment strategies are still expected to develop inhibitors with higher selectivity for APP over Notch cleavage.

Development of antibody drugs targeting Aβ has also been underway. On June 7, 2021, the U.S. Food and Drug Administration (FDA) granted the Accelerated Approval of Aducanumab for the treatment of AD. This made Aducanumab the first AD drug approved since 2003. Aducanumab is a monoclonal antibody that selectively targets aggregated Aβ, and it has been confirmed that monthly intravenous infusions of Aducanumab for one year in prodromal or mild AD patients result in a dose-dependent and time-dependent reduction of brain Aβ [13]. The FDA approved this drug based on its ability to decrease brain amyloid-β plaques; however, many experts believe that the clinical trial data do not definitively prove that Aducanumab significantly delays cognitive decline, leading to a contentious debate regarding its effectiveness [14]. Like Aducanumab, Lecanemab is a monoclonal antibody targeting Aβ and was granted the Accelerated Approval in the U.S. in January 2023, followed by full approval from the FDA in July 2023 [15]. Lecanemab reduced markers of amyloid in early AD and produced moderate reductions over placebo in measures of cognitive function at 18 months, but simultaneously, it was associated with amyloid-related imaging abnormalities (ARIA) such as brain microhemorrhages and cerebral edema [16]. When considering the use of Lekanemab, careful attention must be given to the potential risks of severe adverse events associated with ARIA.

1.5 Purpose of this study

This study was conducted to estimate the protein cleavage mechanism and sequence specificity of γ-secretase. Specifically, we aimed to estimate the following: how many pockets make up the active site of cleavage for γ-secretase, what physicochemical properties of amino acids the active site recognizes, and what kinds of amino acids,

when near the cleavage point, make cleavage more or less likely to occur. Using experimental data that measured the amount of cleaved APP fragments as training data, we utilized machine learning to model the substrate cleavage phenomenon by γ-secretase. From the models created, we selected those that reflected the characteristics of γ-secretase's consecutive cleavage. We interpreted these models to estimate the protein cleavage mechanism and sequence specificity of γ-secretase.

2. Development of a substrate cleavage site prediction model for γ-secretase and estimation of its cleavage mechanism

2.1 Introduction

In this chapter, substrate cleavage models for γ-secretase were created using training data from experimental measurements of APP cleavage by γ-secretase. We created 5280 regression models to predict the amount of substrate cleavage by γ-secretase by combining six pocket models of the active site of γ-secretase, ten physicochemical properties of amino acids, and 88 machine learning methods. To select a model that reflects the consecutive cleavage feature of γ-secretase among these models, we devised a method for predicting substrate cleavage points using regression models, targeted 35 known cleavage points for cleavage point prediction, and selected the model with the highest prediction accuracy. By interpreting the selected model, the number of pockets in the active site of γ-secretase, physicochemical properties of amino acids the active site recognizes were estimated.

2.2 Materials and Methods

2.2.1 Training Dataset

The training dataset [12] used to create regression models is listed in Table 1. This dataset contains 35 APP fragments ranging from three to six amino acids in length and their amount. In the experiment where these data were measured, no fragments with less than two or more than six sequences were detected (Okochi. M, unpublished data). Therefore, this dataset is considered to contain almost all the cleavage fragments of APP produced by the cleavage of γ-secretase. Although this is a small dataset, it is considered full of information on APP cleavage by γ-secretase.

Table 1. Levels of γ-byproducts used as the training data

fragment	peptide amount
MVG	0
VGG	0
GGV	0
GVV	36.08
GGVV	0
VVI	12.09
GVVI	0
VIA	21.89
VVIA	25.61
GVVIA	0
IAT	206.44
VIAT	0
VVIAT	26.26
ATV	0
IATV	0
VIATV	0
TVI	34.19
ATVI	0
IATVI	0
VIV	150.67
TVIV	0
ATVIV	0
IATVIV	0
IVI	6.5
VIVI	5.59
TVIVI	0
VIT	55.08
IVIT	0
VIVIT	0
ITL	206.36
VITL	26.78
IVITL	0
VIVITL	0
TLV	0
LVM	0

(fmol/dish)

2.2.2 Pocket model for the cleavage active site

When enzymatic cleavage of a substrate occurs, the substrate binds to the cleavage active site of the enzyme. Simplistically, the substrate is considered to fit into pockets aligned with the cleavage active site of the enzyme. In this study, we modeled the cleavage active site of γ-secretase as a series of pockets based on the definition by Schechter *et al*. [17]. Two integer parameters, L and R, were used to define this pocket model. Both L and R are parameters that specify the number of pockets, where L refers to the number of pockets on the N-terminal side from the cleavage site, and R refers to the number of pockets on the C-terminal side from the cleavage site. The substrate sequences these pockets recognize are denoted as PL- ... -P2 -P1-P1'-P2'- ... -PR', and the pockets as SL- ... -S2 -S1-S1'-S2'- ... -SR'. The cleavage site is located between P1 and P1', and it is at this position that cleavage of the substrate occurs. The pocket model specified by the parameters L and R is denoted as L+R pocket. As an example, a 4+3 pocket is shown in Fig. 2.

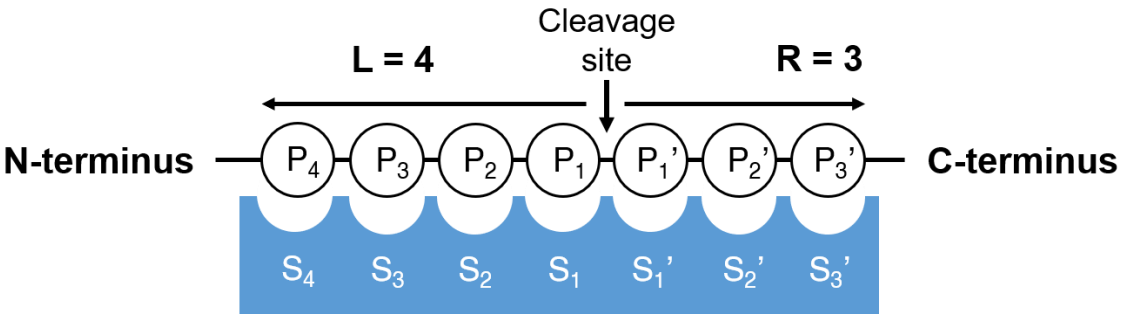

Figure 2. Pocket model.

The number of pockets in the cleavage active site of an enzyme is expected to vary from enzyme to enzyme. In this study, we estimated the pocket size of γ-secretase by using six different pocket models, where R is fixed to three and integers one, two, three, four, five, and six are specified as L, following the model of Bolduc *et al.* [8]. These pocket models were used to retrieve sequence information around the cleavage site recognized by the pocket when a given peptide fragment sequence was cut from the substrate.

2.2.3 Compression of amino acid property information by principal component analysis

2.2.3.1 AAindex

The active site of γ-secretase is thought to recognize some physicochemical properties of the amino acids in the substrate sequence to determine cleavage. To estimate the

physicochemical properties involved in the cleavage of γ-secretase, we used AAindex, a database of numerical indices representing various physicochemical properties of 20 amino acids [18]. These physicochemical properties were collected from the existing literature and contained 566 indices (release 9.2). In this study, we decided to use 553 of these 566 indices that did not contain missing values.

2.2.3.2 Compression of AAindex by principal component analysis

The 553 indices are too large to consider individually, and it is unlikely that cleavage by γ-secretase is determined by only one indexed amino acid property. Instead, it is natural to assume that multiple amino acid properties are simultaneously involved in cleavage determination. Therefore, principal component analysis was performed on the 553 indexes, and new amino acid property information was created by integrating the 553 indexes. Principal component analysis was performed after standardizing AAindex. Of the principal components obtained as a result of the principal component analysis, principal components 1 through 10 (PC1 through PC10) were used as the amino acid physical property value information in this study. The principal component scores for each PC for the 20 amino acids are shown in Table 2.

Table 2. The principal component scores for each PC for the 20 amino acids

PC	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	PV (%)
PC1	-0.23	7.97	14.84	17.89	-8.01	7.67	11.75	15.99	0.32	-20.15	-17.45	11.40	-15.77	-18.95	16.39	12.33	4.81	-17.15	-8.10	-15.55	34.39
PC2	-5.52	15.38	0.05	2.86	-7.99	8.85	11.37	-21.05	8.14	-5.72	-3.18	13.07	4.64	0.45	-10.94	-7.86	-5.86	8.57	3.21	-8.49	15.83
PC3	16.18	-1.08	-4.02	-0.04	-6.96	1.78	10.99	1.37	-4.78	2.71	11.18	5.65	3.10	-3.89	-14.19	1.33	0.18	-12.83	-12.67	5.98	11.7
PC4	-1.28	-0.41	5.58	1.96	14.58	0.34	-2.73	5.14	4.47	-3.72	-6.42	-1.68	2.24	-2.47	-18.55	3.34	1.39	0.09	-0.38	-1.49	6.966
PC5	3.27	-8.55	-1.89	5.76	10.25	1.43	8.10	-7.91	0.91	-2.69	-1.37	-5.95	6.77	-1.15	7.25	-2.19	-2.61	0.88	-7.25	-3.07	5.475
PC6	-1.91	8.07	-3.08	-5.34	8.28	2.43	-3.43	-8.99	-0.70	1.96	-1.85	1.97	-1.92	-4.03	3.69	2.81	6.09	-8.38	-1.91	6.23	4.536
PC7	3.56	3.14	-2.56	-8.80	1.77	1.11	-4.68	5.66	4.16	-3.91	0.39	4.47	6.85	-0.59	3.66	-1.75	-4.03	0.85	-4.72	-4.58	3.251
PC8	-0.65	-5.38	5.38	-1.37	-4.88	0.44	-3.22	-4.15	9.09	0.01	0.86	-1.11	3.36	2.35	0.07	3.33	2.65	-5.93	0.18	-1.05	2.548
PC9	3.46	-2.17	-0.12	-2.00	-1.27	0.45	-2.60	-3.59	-3.03	-3.81	1.59	1.01	-1.60	-1.19	-0.77	6.08	6.92	7.83	-2.61	-2.57	2.223
AASS10	-2.22	-5.70	4.67	-1.34	4.06	0.35	-1.07	-1.83	-2.92	1.22	4.69	8.69	-2.85	-1.30	0.60	0.10	-4.55	-0.75	1.39	-1.23	2.137

PV denotes the proportion of variance. The cumulative proportion calculated by the sum of PVs is 89.05

2.2.4 Generation of regression models of γ-secretase cleavage

2.2.4.1 Machine learning algorithms

We applied the caret package (short for Classification And REgression Training) to develop regression models, which contain functions to streamline the model training process for complex regression and classification problems [19]. In the present study, we compared 88 regression models in Table 3 because we did not have information for choosing suitable regression models. These are classified into non-ensemble- and ensemble-type models. The former is classified into nine subcategories, and the latter into four subcategories.

2.2.4.2 Generation of regression models

The process of creating the regression model is illustrated in Fig. 3. In the figure, an example with a 4+3 pocket is shown. First, the kth APP fragment in the training data is mapped to its original location in the APP sequence. Since the fragment is cleaved at the C- and N-terminal ends, we align the cleavage point of the pocket model to these cleavage points and extract the pocket-bound sequences (Step 1). These are combined in the order of C-terminally cleaved sequence and N-terminally cleaved sequence to create an amino acid vector (Step 2). Then, an amino acid score vector is created by replacing the elements of this amino acid vector with the corresponding scores of certain principal components (Step 3). Using machine learning, a regression model is created that uses this amino acid score vector as an input to predict the amount of cleavage y_k of the APP fragment (Step 4). We define the function $x_p^{L+R}(\cdot)$ as the transformation operation described in Steps 1 through 3 that creates a 2(L+R) amino acid vector from the substrate fragment sequence using the L+R pocket and replaces the elements of this amino acid vector with the score of the principal component p to create an amino acid score vector.

2.2.4.3 RMSE

To evaluate the goodness of fit of the regression models, we calculated the RMSE given by the equation $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2}$, where y_i is the value of the *i*-th observation (Table 1) and f_i is the predicted value of the corresponding *i*-th prediction. The RMSEs were plotted by ggplot2 3.3.5.

Table 3. Machine learning algorithms used in the present study

category	**approach name in caret**
Kernel(15)	gaussprLinear, gaussprPoly, gaussprRadial, kernelpls, krlsPoly, krlsRadial, rvmRadial, svmLinear, svmLinear2, svmLinear3, svmPoly, svmRadial, svmRadialCost, svmRadialSigma, widekernelpls
Linear(15)	bayesglm, bridge, glm, glmStepAIC, icr, leapBackward, leapForward, leapSeq, lm, lmStepAIC, pcr, pls, plsRglm, simpls, superpc
Sparse(10)	blasso, blassoAveraged, enet, glmnet, lars, lars2, lasso, penalized, ridge, spls
Neural Networks(8)	avNNet, brnn, dnn, mlpWeightDecayML, monmlp, msaenet, nnet, pcaNNet
Decision Tree(8)	ctree, ctree2, evtree, partDSA, rpart, rpart1SE, rpart2, WM
kNN(2)	knn, kknn
Spline(4)	bam, gam, gamSpline, gcvEarth
Fuzzy(4)	ANFIS, DENFIS, HYFIS, SBC
Others(3)	null, ppr, spikeslab
Ensemble Linear(3)	BstLm, glmboost, xgbLinear
Ensemble Decision Tree(11)	blackboost, bstTree, cforest, nodeHarvest, parRF, ranger, Rborist, RRFglobal, rf, treebag, xgbTree
Ensemble Spline(3)	bagEarth, bagEarthGCV, bstSm
Others(2)	gamboost, gamLoess

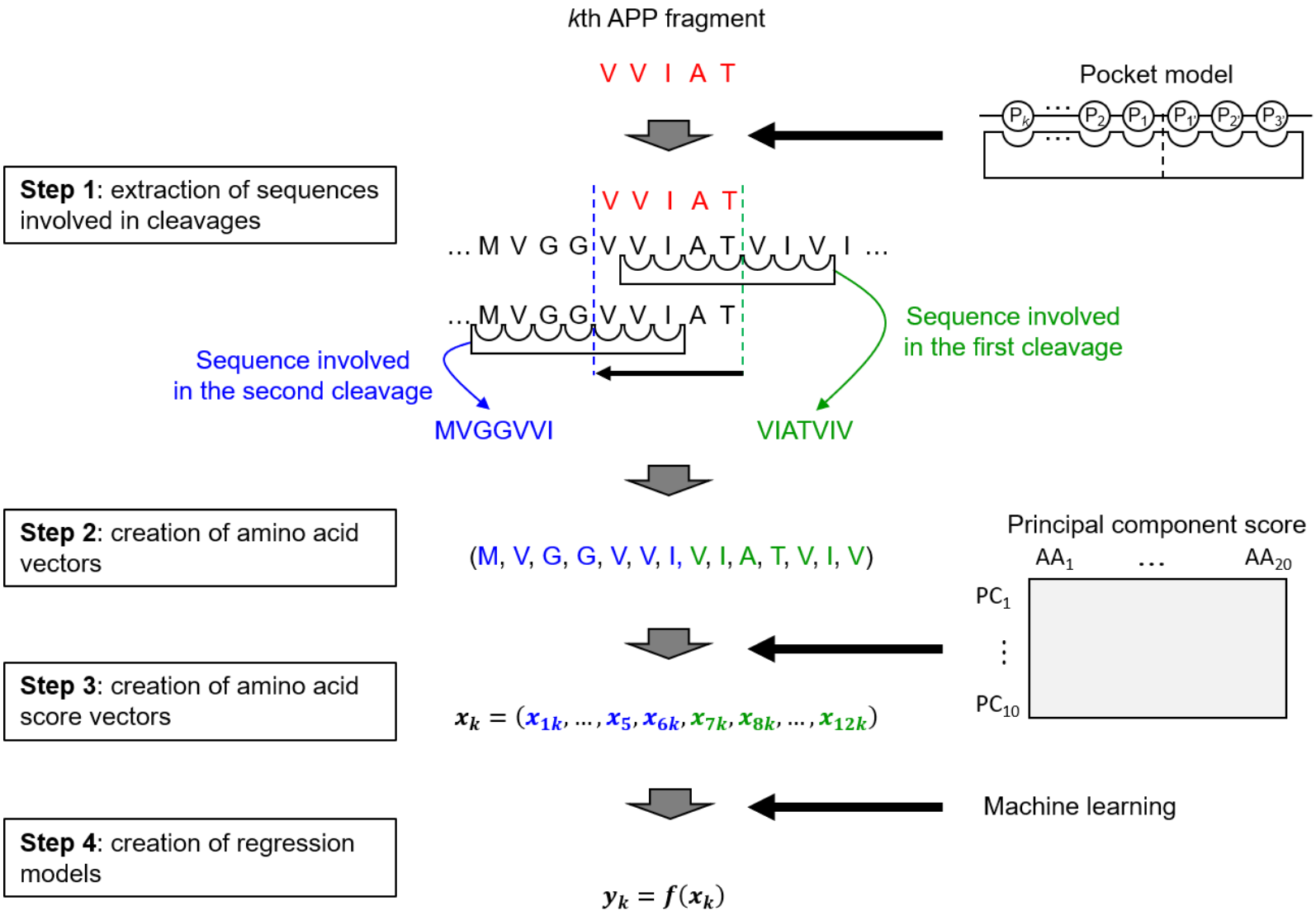

Figure 3. The schema for creating regression models based on γ-byproducts.

2.2.5 Model validation using substrate cleavage endpoint prediction

2.2.5.1 Validation dataset

To predict the substrate cleavage points of γ-secretase using the regression model, we collected 26 substrates for which both the initial (ε-like) and terminal (γ-like) cleavage sites were known. These are summarized in Table 4, with the initial cleavage sites of sequential cleavage indicated by red arrows and the terminal cleavage site indicated by blue arrows.

Table 4. Summary of reported γ-cleavage sites

Substrate	Sequence and its cleavage site (↓)	Ref.
Alcadein-α	VPSTAT ↓ VVIVVCVSFLVFMIIL ↓ GVFRI	[20]
Alcadein-β	AATLII ↓ VVCVGFLVLMVVLG ↓ LVRIH	[20]
Alcadein-γ	VPSIAT ↓ VVIIISVCMLVFVVAM ↓ GVYRV	[20]
APLP2	SALIGL ↓ LVIAVAIATVIVSL ↓ VMLRK	[21]
APP	MVGGVV ↓ IAT ↓ VIV ↓ ITL ↓ VMLKK	[22, 23, 24]
CD44	LASLLA ↓ LALILAVCI ↓ AVNSR	[25, 26]
CTFβ-GA	GVV ↓ IAT ↓ VIVG ↓ ALVML	[27]
CTFβ-GAζtoγ	VIATGA ↓ VIT ↓ LVMLK	[27]
CTFβ-GG	GVV ↓ IAT ↓ VIVGGL ↓ VMLKK	[27]
CTFβ-GGζtoγ	VIATGG ↓ VIT ↓ LVMLK	[27]
CTFβ-LL	GVV ↓ IAT ↓ VIVLLL ↓ VMLKK	[27]
CTFβ-LLγtoγ'	GVVLL ↓ T ↓ VIVIT ↓ LVMLK	[27]
CTFβ-LLζtoγ	VIATLL ↓ VIT ↓ LVMLK	[27]
CTFβ-ΔI47-L49	MVGGVV ↓ IAT ↓ VIV ↓ VMLKK	[27]
CTFβ-ΔL49	MVGGVV ↓ IATVIVI ↓ TVMLK	[27]
CTFβ-ΔL52	MVGGVV ↓ IAT ↓ VIVITL ↓ VMKKK	[27]
CTFβ-ΔM51L52	MVGGVV ↓ IAT ↓ VIV ↓ ITLVK	[27]
CTFβ-ΔT48L49	MVGGVV ↓ IATVIV ↓ IVMLK	[27]
CTFβ-ΔV50-L52	MVGGVV ↓ IATVIV ↓ ITLKK	[27]
hEpCam	VIAVIV ↓ VVVIAVVAGIV ↓ VLVIS	[28]
mEpCam	IIAVIV ↓ VVSLAVIAGIV ↓ VLVIS	[28]
Neuregulin 1	GICIAL ↓ LVVGIMC ↓ VVAYC	[29]
Notch1	LMYVAA ↓ AAFVLLFFVGCG ↓ VLLSR	[30, 31]
rNotch3	PLLVA ↓ GAVLLLVILVLGV ↓ MVARR	[32]
rNotch4	AGVIL ↓ LALGALL ↓ VLQLI	[32]
rVEGFR1	TLTCT ↓ CVAATLFWLLL ↓ TLFIR	[32]

Red and blue arrows represent the major initial and terminal cleavage sites reported, respectively.

2.2.5.2 Prediction of substrate cleavage termination site using regression models

γ-Secretase is expected to cleave substrates other than APP successively, although their γ-byproducts have not been experimentally determined. For model validation, we devised a method using a regression model to calculate the probability of cleavage occurring at each site in substrate sequences. We used this method to predict the termination point of γ-secretase successive cleavage. In the present study, we assumed that only 3-5 peptide cleavage occurred because the length distribution of APP peptides cleaved by γ-secretase was between three and five, and γ-byproducts longer than six amino acids were not observed. Firstly, the cleavage initiation site in a targeted substrate sequence is designated as position 0. The site which is q amino acids away from position 0 in the N-terminal direction of the sequence as position q. Secondly, we define P(r) as the probability of cleavage occurring at position r and calculated P(3), P(4) and P(5) as follows: $\mathrm{P}(3) = \frac{f(\mathrm{PC}_j(x_0^3))}{\sum_{i=3}^{5} f(\mathrm{PC}_j(x_0^i))}$, $\mathrm{P}(4) = \frac{f(\mathrm{PC}_j(x_0^4))}{\sum_{i=3}^{5} f(\mathrm{PC}_j(x_0^i))}$, and $\mathrm{P}(5) = \frac{f(\mathrm{PC}_j(x_0^5))}{\sum_{i=3}^{5} f(\mathrm{PC}_j(x_0^i))}$. We also define $p_s(t)$ as the probability that t amino acids are cleaved from position s and its formula as $p_s(t) = \frac{f(\mathrm{PC}_j(x_s^t))}{\sum_{i=3}^{5} f(\mathrm{PC}_j(x_s^{s+i}))}$, where x_s^u represents the amino acid vector obtained from the sequence between positions u and s. $\mathrm{PC}_j(\cdot)$ is a function that converts the amino acids in a sequence to the *j*-th principal component scores of the corresponding amino acids. P(1), P(2), $p_1(t)$, and $p_2(t)$ are defined as 0 because of the assumption that only 3-5 peptide cleavage occur. Using $p_s(t)$, P(r) is expressed as $\mathrm{P}(r) = \sum_{i=3}^{5} \mathrm{P}(r-i)p_{r-i}(i)$. For example, P(9) is obtained from relative occurrences in all paths from the initiation site to position 9, i.e., $\mathrm{P}(9) = p_0(3)p_3(3)p_6(3) + p_0(4)p_4(5) + p_0(5)p_5(4)$.

If the cleavage probability at the termination site $\mathrm{P}(\text{termination site})$ is the maximum in comparison to those for positions between termination site – 2 and termination site + 2, the termination position is correctly predicted. We assessed the performance of machine learning algorithms using 26 substrate peptides whose initiation and termination sites are reported.

2.2.6 Factor loading analysis of AAindex

To interpret the physical property information of the amino acids represented by the PCs made as compressed values of AAindex, the correlation coefficients between the principal component scores and the 553 indices were calculated

2.3 Results

2.3.1 Creation of regression models

In this study, we constructed numerous regression models to infer the γ-secretase cleavage mechanism. These models take a partial sequence from substrates as input and output a predicted amount of cleavage by γ-secretase. The process of a regression model receiving a partial sequence and predicting its cleavage amount includes extracting substrate sequence information recognized by the active site of γ-secretase using a pocket model and then replacing the amino acids in the extracted sequence information with their physicochemical properties. This operation faithfully incorporates the movement occurring when γ-secretase cuts out a partial sequence of the substrate and also embodies the idea that the active site of γ-secretase recognizes the physicochemical properties of the amino acids in the substrate sequence to determine cleavage. Therefore, it is considered that the regression models created here serve as models for the phenomena that occur when γ-secretase cleaves the substrate. If the number of pockets, physicochemical properties of the amino acids, and regression algorithm used for training are well chosen, we think it is possible to obtain a regression model that can explain the cleavage phenomenon by γ-secretase. The objective of this study, which is to estimate the cleavage mechanism of γ-secretase, is thought to be achieved by examining such regression models.

In creating a regression model, it is necessary to specify the number of pockets of the active site of γ-secretase, the physicochemical properties of the amino acids recognized by the pocket, and the functional relationship that determines the correlation between the sequence information recognized by the pockets during cleavage and the amount of cleavage. However, all of this information is unknown beforehand. Therefore, by exhaustively testing combinations of six pocket models, ten types of physical property information created by integrating AAindex, and 88 regression algorithms and training them with APP cleavage data, a total of 5,280 regression models were created.

We initially examined the goodness of fit, the difference between values of the training and predicted data, of generated models. As shown in Fig. 4A–F, the Root Mean Squared Errors (RMSEs) of regression models were plotted for the different numbers of non-prime sites and PCs. A very well fit with RMSE≤5 was observed in 244 models, suggesting that these regression models precisely reproduced APP successive cleavage. These models spread widely for different numbers of pockets and PCs. Notably, the medians of RMSE of regression models generated with PC9 were the largest, irrespective of the number of pockets.

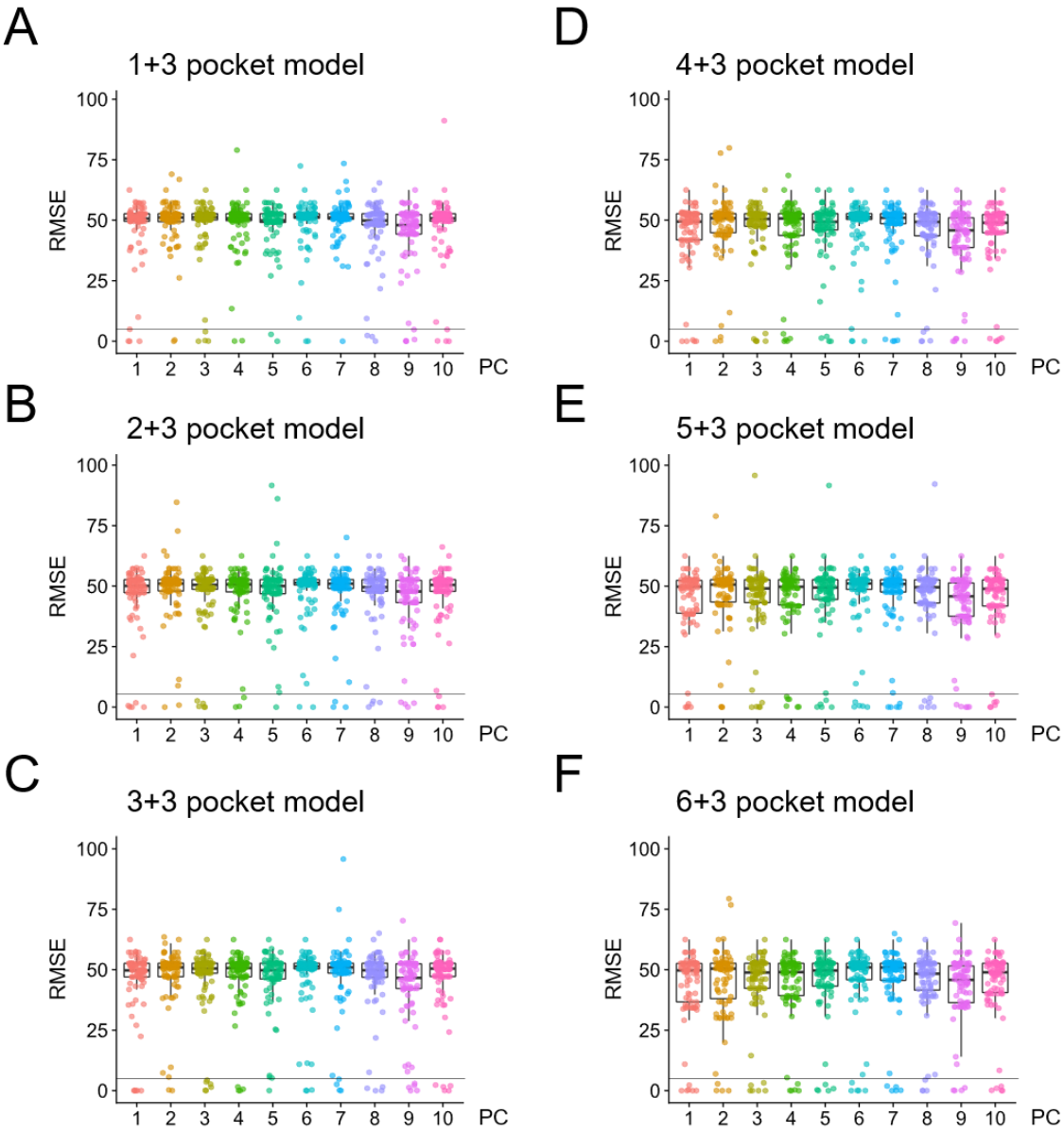

Figure 4. RMSE of regression models. RMSEs of the models were separately shown with a various number of pockets, PCs, and machine learning algorithms. Each dot represents the RMSE of one model for different parameters. L + R pocket model means the numbers of N-sided and C-sided pockets centered at cleavage site, respectively.

2.3.2 Validation of regression models

The construction of regression models revealed that well-fitted models could be developed for any number of pockets and PC conditions. However, the good fit to the training data alone does not necessarily indicate that the model successfully encapsulates the characteristics of γ-secretase. To select a regression model which reflects the characteristics of γ-secretase, we planned to examine regression models with other γ-secretase substrates. Successive cleavages of γ-secretase substrates other than APP are highly expected, although their γ-byproducts have not been experimentally determined yet. Instead, initial cleavage sites corresponding to the ε-cleavage site of APP and terminal cleavage sites corresponding to the γ-cleavage site for several substrates have been reported (major cleavage sites are summarized in Table 4). We noticed that this information could be used for evaluating the generated regression models. Because the regression models estimate the amount of 3–5 amino acid peptide cleaved at two sites, designating one cleavage site allows regression models to calculate relative cleavage probabilities at the subsequent potential cleavage sites at 3–5 amino acids away from the designated site. Moreover, cleavage probabilities at positions more than five amino acids away from the designated site could be calculated by summing up all the possible theoretical permutations of 3–5 acid-spaced cleavages since one cycle of successive cleavage occurs at 3–5 amino acids intervals. Accordingly, we calculated cleavage probabilities by assigning the reported initial cleavage site into regression models. We evaluated terminal cleavage prediction by comparing the predicted probabilities of cleavages near the reported terminal cleavage sites.

The prediction accuracies of models were calculated by the percentage of correctly predicted substrates, as shown in Fig. 5. Some models predicted as accurately as more than 80%. The best model hits 85.7% of the terminal cleavage sites, i.e., 30 over 35 cleavage sites, and is constructed with the SBC algorithm [33], the number of non-prime sites, and PC9. Detailed results of individual substrates are summarized in Table 5.

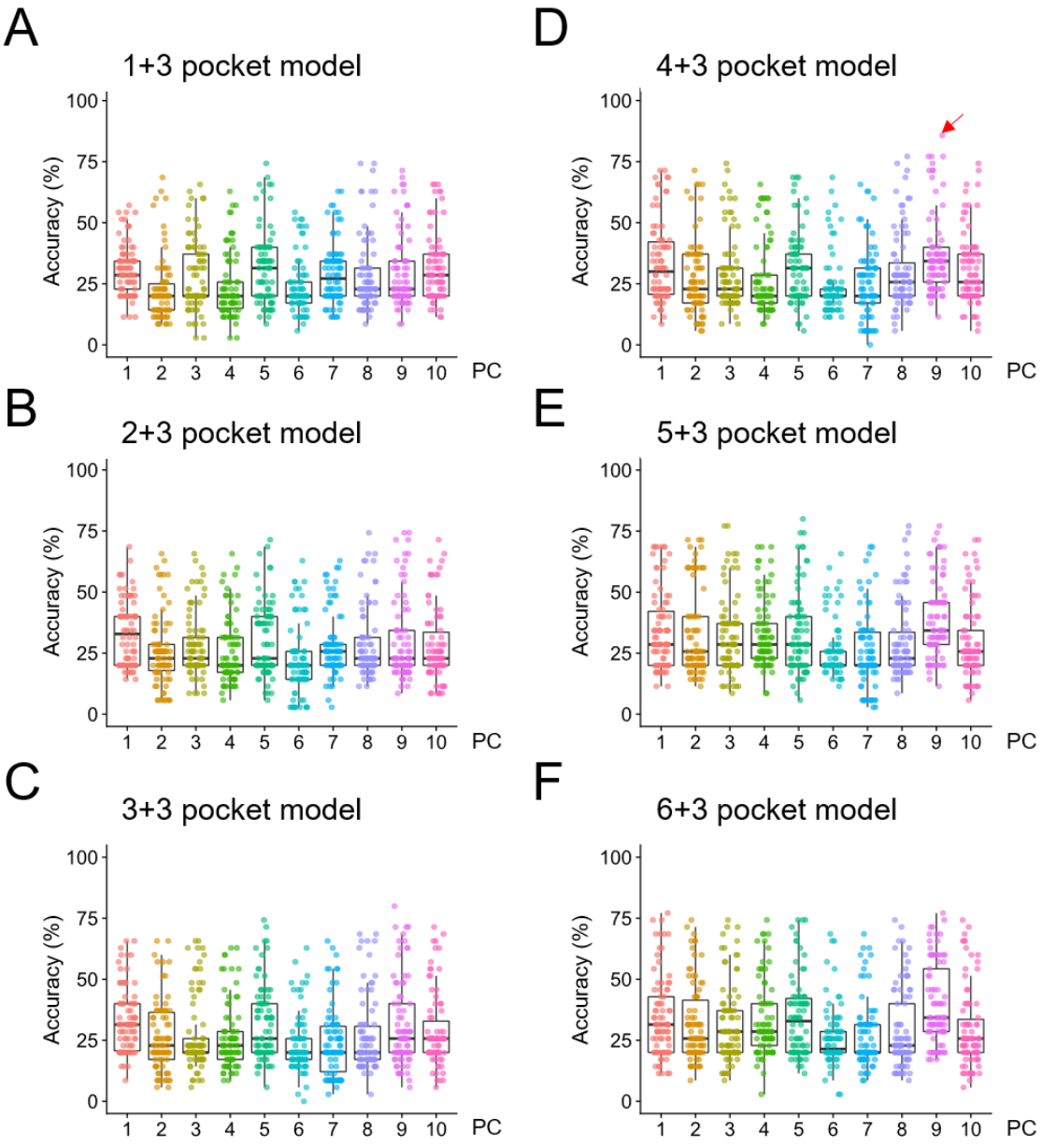

Figure 5. Percentage of correct predictions of terminal cleavage of regression models. Each dot represents the percentage of correct prediction of one model for different numbers of pockets and PC parameters. A red arrow indicates the most accurate model. L + R pocket model means the numbers of N-sided and C-sided pockets centered at cleavage site, respectively.

Table 5. Summary of reported and predicted γ-cleavage sites

Substrate	Sequence and its cleavage site (↓)	S or F
Alcadein-α	VPSTAT ↓ VVIVVCVSFLVFMIIL ↓ GVFRI	S
Alcadein-β	AATLII ↓ VVCVGFLVLMVVLG ↓ LVRIH	F
Alcadein-γ	VPSIAT ↓ VVIIISVCMLVFVVAM ↓ GVYRV	S
APLP2	SALIGL ↓ LVIAVAIATVIVSL ↓ VMLRK	S
APP	MVGGVV ↓ IAT ↓ VIV ↓ ITL ↓ VMLKK	S S S
CD44	LASLLA ↓ LALILAVCI ↓ AVNSR	S
CTFβ-GA	GVV ↓ IAT ↓ VIVG ↓ ALVML	S F
CTFβ-GAζtoγ	VIATGA ↓ VIT ↓ LVMLK	S
CTFβ-GG	GVV ↓ IAT ↓ VIVGGL ↓ VMLKK	S S
CTFβ-GGζtoγ	VIATGG ↓ VIT ↓ LVMLK	S
CTFβ-LL	GVV ↓ IAT ↓ VIVLLL ↓ VMLKK	S S
CTFβ-LLγtoγ'	GVVLL ↓ T ↓ VIVIT ↓ LVMLK	F S
CTFβ-LLζtoγ	VIATLL ↓ VIT ↓ LVMLK	S
CTFβ-ΔI47-L49	MVGGVV ↓ IAT ↓ VIV ↓ VMLKK	S S
CTFβ-ΔL49	MVGGVV ↓ IATVIVI ↓ TVMLK	S
CTFβ-ΔL52	MVGGVV ↓ IAT ↓ VIVITL ↓ VMKKK	S S
CTFβ-ΔM51L52	MVGGVV ↓ IAT ↓ VIV ↓ ITLVK	S S
CTFβ-ΔT48L49	MVGGVV ↓ IATVIV ↓ IVMLK	S
CTFβ-ΔV50-L52	MVGGVV ↓ IATVIV ↓ ITLKK	S
hEpCam	VIAVIV ↓ VVVIAVVAGIV ↓ VLVIS	S
mEpCam	IIAVIV ↓ VVSLAVIAGIV ↓ VLVIS	S
Neuregulin 1	GICIAL ↓ LVVGIMC ↓ VVAYC	S
Notch1	LMYVAA ↓ AAFVLLFFVGCG ↓ VLLSR	S
rNotch3	PLLVA ↓ GAVLLLVILVLGV ↓ MVARR	S
rNotch4	AGVIL ↓ LALGALL ↓ VLQLI	F
rVEGFR1	TLTCT ↓ CVAATLFWLLL ↓ TLFIR	F

S or F represent successful or false prediction by the best regression model, respectively.

Selection by the prediction of the terminal cleavage sites provided some regression models with better accuracy. We wondered whether these models not ostensibly but fundamentally reproduced successive cleavage of γ-secretase. Thus, we plotted cleavage probabilities at each peptide bond in substrates to evaluate the whole process of successive cleavages by the models (Fig. 6, 7, and 8). We first focused on APP cleavage, which occurs in two primary product lines with sequential cleavages at the 49/46/43/40 and 48/45/42/38 sites. As shown in Fig 6A, the probability plots starting from ε49 showed clear successive cleavages in the order: ITL, VIV, and IAT for the ε49 line. Notably, by designating the ε48 site as an initial site, the cleavage was predicted to proceed sharply in the order: VIT, TVI, and VVIA for the ε48 line as reported [6, 34]. Moreover, the model predicted that Notch1 cleavage occurred between V/G, V/L, and A/A (Fig. 6C). The cleavage at V/G is a novel unreported one, similar to the ζ-cleavage of APP, and indicating the generation of an Nβ30 as an intermediate for the subsequent cleavages at the V/L and A/A sites generating Nβ25 and Nβ21, respectively. Thus, the best regression model reproduced the characteristic terminal cleavage of Notch1 [35].

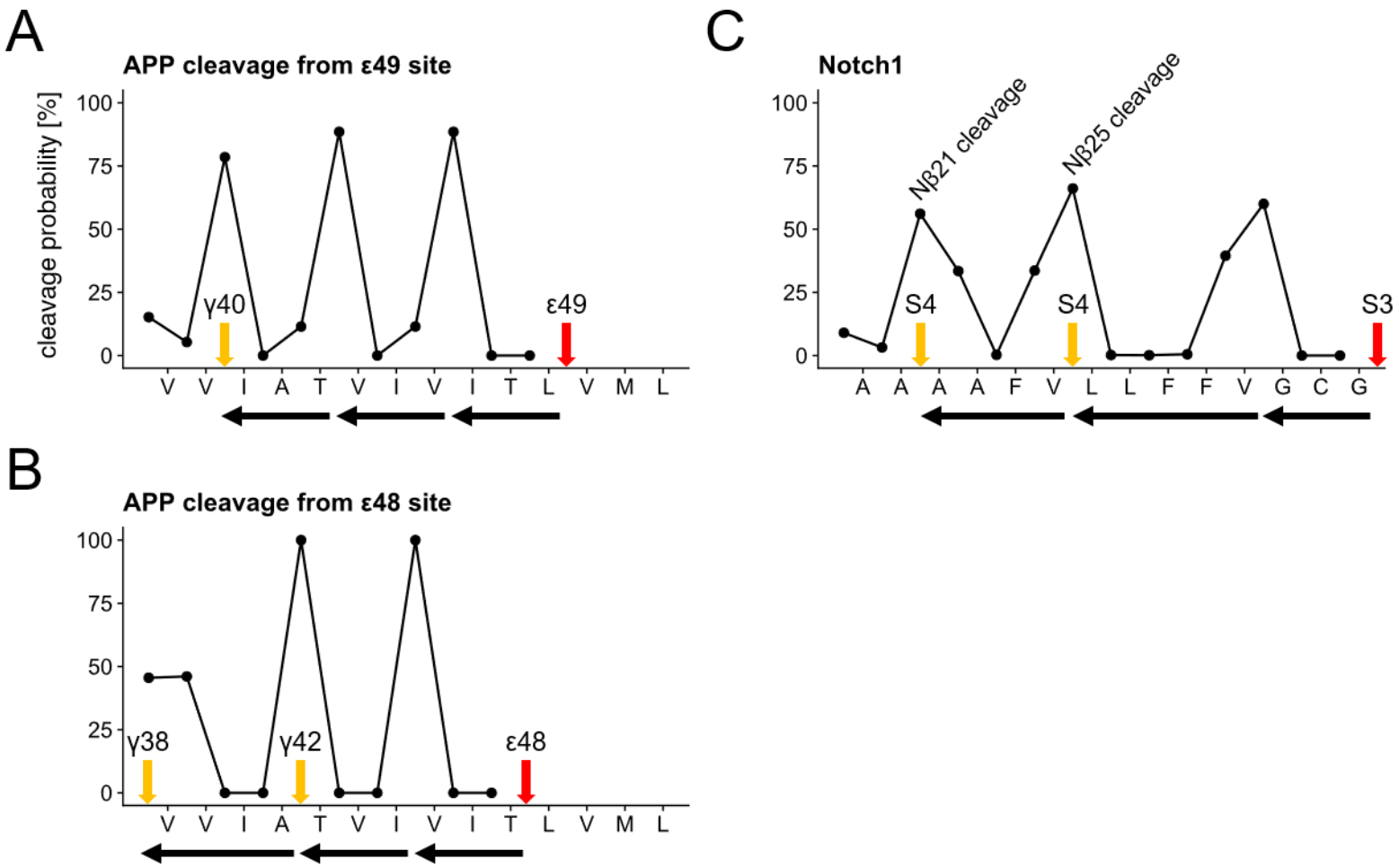

Figure 6. Visualization of *in silico* successive cleavage. The probability plots of APP cleavage starting at ε49 and ε48 by the best regression model are shown in (A) and (B), respectively. The cleavage probability plot for Notch1 is shown in (C). Red and orange arrows represent initial and terminal cleavage sites, respectively. Black arrows indicate corresponding γ-byproducts.

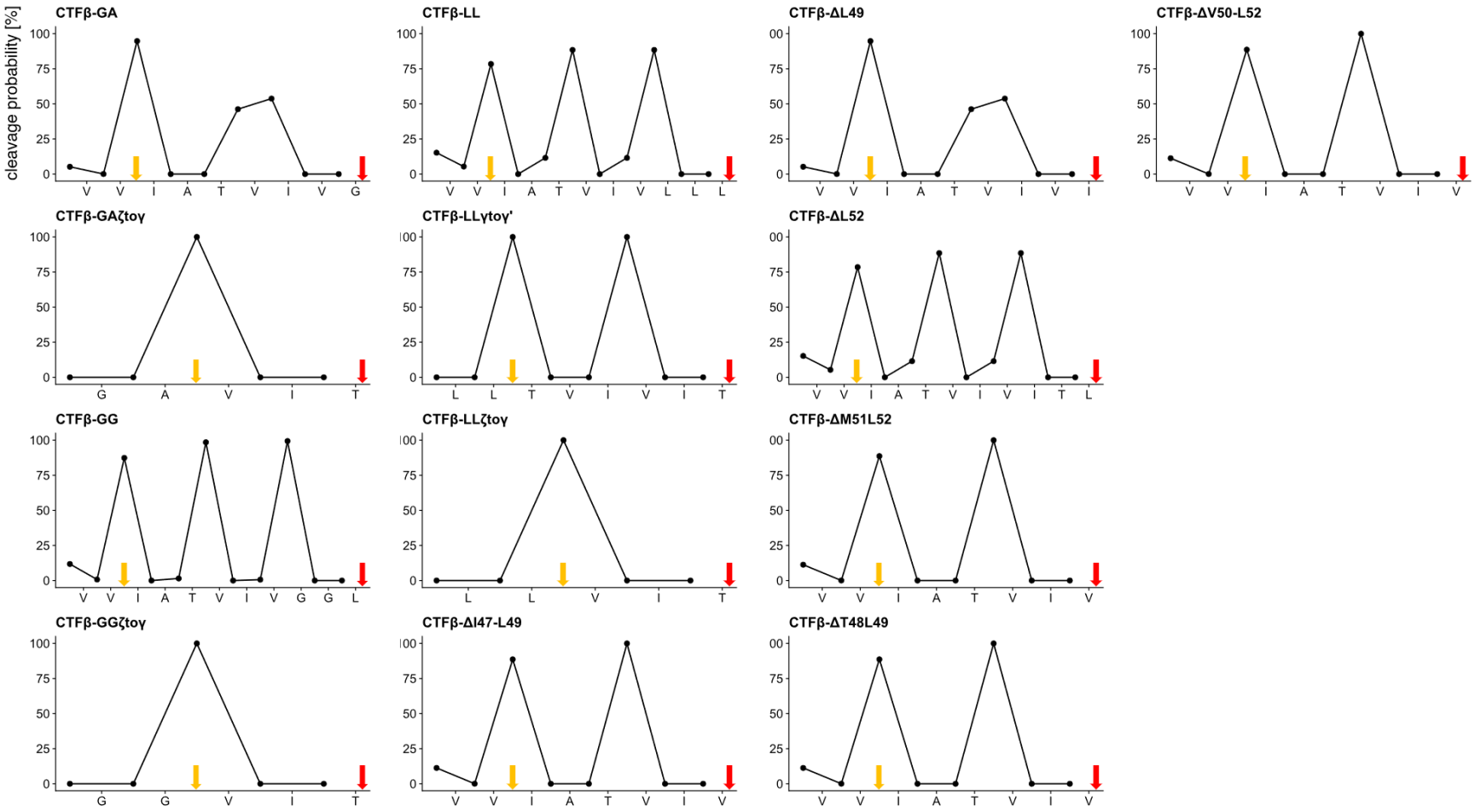

Figure 7. Probability plot of γ-secretase cleavage of APP variant substrates. The best regression model simulated the successive γ-cleavages of APP variants. Red and orange arrows represent initial and terminal cleavage sites, respectively.

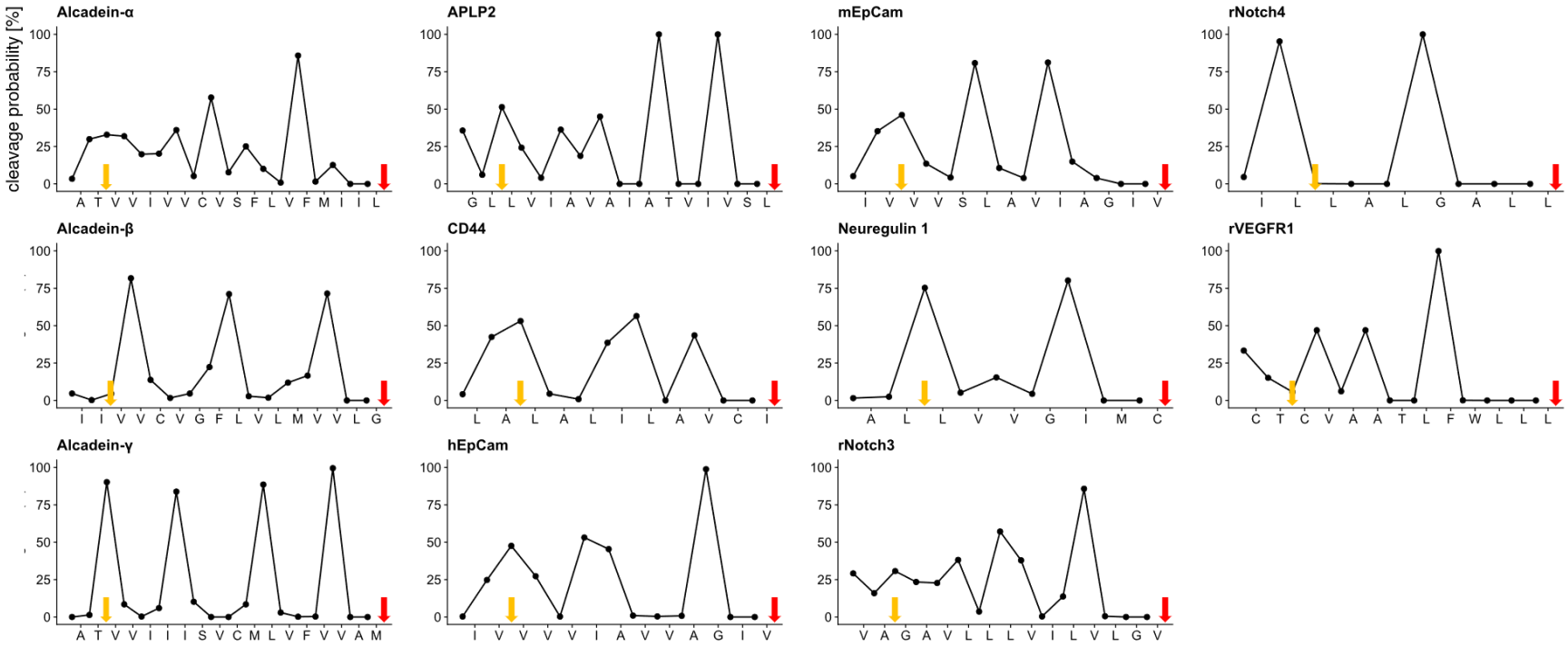

Figure 8. Probability plot of γ-secretase cleavage of substrates. The best regression model simulated the successive γ-cleavages of indicated substrates. Red and orange arrows represent initial and terminal cleavage sites, respectively.

As mentioned above, the best model well predicted successive cleavages of previously reported substrates. Therefore, we planned to test the best prediction model with a previously unanalyzed substrate. Rat APLP1 was chosen because the terminal cleavage sites of rat APLP1 were not analyzed before, although the initial cleavage site of rat APLP1 and the terminal cleavage site of human APLP1 were reported [24, 34, 36, 37] (Fig. 9A). Two amino acids are different between rat and human; rat APLP1 L571 and L589 are human APLP1 V582 and M600, respectively. We first plotted the cleavage probabilities of rat APLP1 cleavage by the best regression model by inputting the initial cleavage site at L/S. As shown in Fig. 9B, terminal cleavages of rat APLP1 were predicted to occur at L/I and G/G. Then, we experimentally determined rat APLP1 terminal cleavage sites from rat primary neurons by MALDI-TOF MS. As predicted, the major cleavage site of rat APLP1 was at L/I (Fig. 9C).

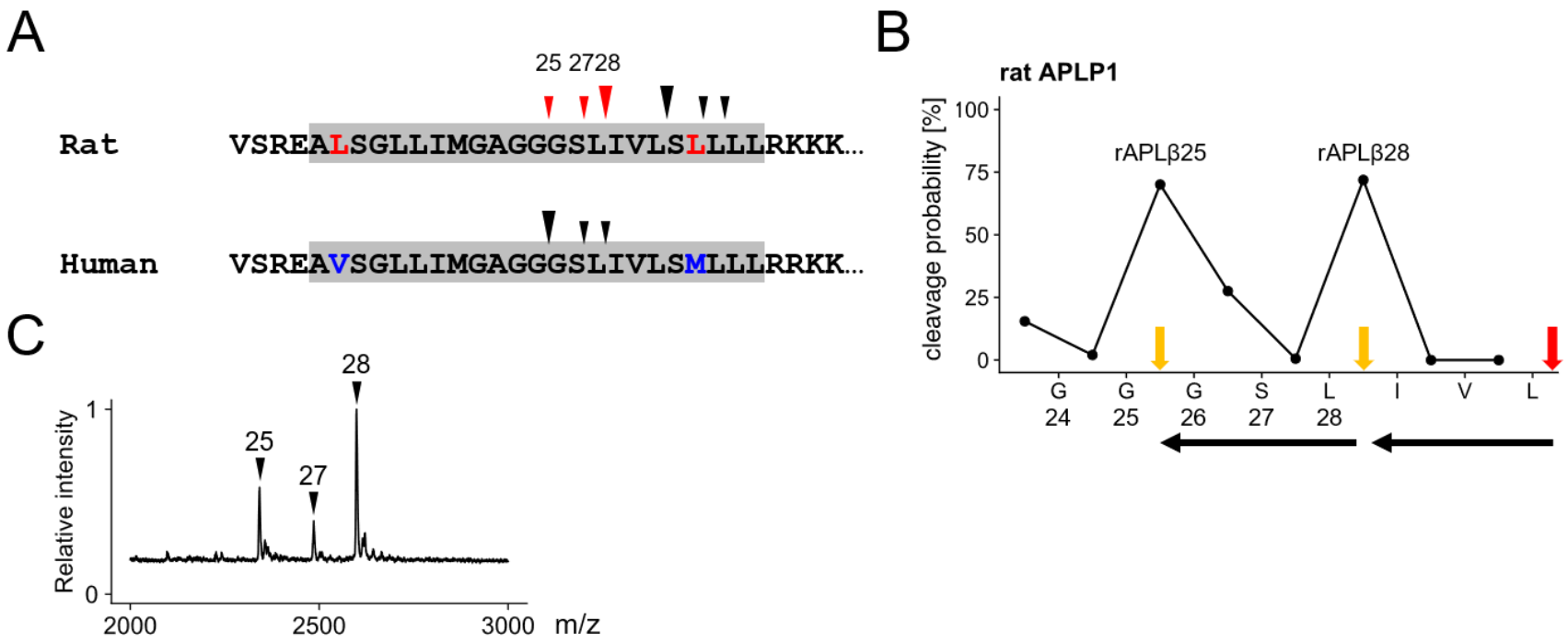

Figure 9. Prediction and experimental validation of γ-secretase cleavage of rat APLP1. The diagram shows rat and human APLP1 sequences around transmembrane domains (A). Amino acids different between rat and human APLP1 are labeled in red and blue, respectively. Black arrowheads are reported cleavage sites, while red ones are analyzed in this report. The cleavage probability plot of rat APLP1 cleavage was generated by inputting the reported initial cleavage site into the best regression model (B). The rat APLP1 terminal cleavage sites were determined by the immunoprecipitation/mass spectrometry, showing that the cleavage between amino acid 28 and 29 is the major cleavage site (C). Observed m/z and theoretical molecular mass of rat APLP1β species are shown in Table 6. The predicted cleavages for generating APLβ25 and APLβ28 were experimentally observed. The cleavage between amino acid 27 and 28 is predicted to be generated from minor initial cleavage between amino acid L33 and L34.

Table 6. Observed m/z and theoretical molecular mass of rat APLP1β species

APLP1 species	sequence	theoretical mass	observed m/z
rat APLP1β25	DELAPAGTGVSREALSGLLIMGAGG	2341.2	2343.1
rat APLP1β27	DELAPAGTGVSREALSGLLIMGAGGGS	2485.3	2486.9
rat APLP1β28	DELAPAGTGVSREALSGLLIMGAGGGSL	2598.3	2598.7

2.3.3 Factor loading analysis of PC9

The physical property information of amino acids recognized by the best model is PC9. PC9 is an integrated index combining 553 types of AAindex. In order to interpret the physical property information represented by PC9, it is necessary to investigate which AAindex is strongly related to PC9. Therefore, the factor loadings between PC9 and the 553 types of AAindex were calculated, and those with factor loadings of 0.4 or higher for positive cases and -0.4 or lower for negative cases were collected. The top 5 AAindex information with the highest factor loadings in both positive and negative cases were summarized in Table 7.

Table 7. Top 10 positive and negative amino acid indices by the factor analysis of PC9

(A) Positive top 10 amino acid indices

Accession number	Pearson correlation	Description of amino acid indices	reference
RICJ880117	0.590	Relative preference value at C"	[38]
PALJ810107	0.440	Normalized frequency of alpha-helix in all-alpha class	[39]
MAXF760106	0.425	Normalized frequency of alpha region	[40]
GARJ730101	0.408	Partition coefficient	[41]
VASM830102	0.401	Relative population of conformational state C	[42]
SNEP660104	0.392	Principal component IV	[43]
GEIM800102	0.348	Alpha-helix indices for alpha-proteins	[44]
QIAN880125	0.334	Weights for beta-sheet at the window position of 5	[45]
PRAM820103	0.328	Correlation coefficient in regression analysis	[46]
QIAN880114	0.327	Weights for beta-sheet at the window position of -6	[45]

(B) Negative Top 10 amino acid indices

Accession number	Pearson correlation	Description of amino acid indices	reference
OOBM850102	-0.511	Optimized propensity to form reverse turn	[47]
QIAN880138	-0.502	Weights for coil at the window position of 5	[45]
AURR980101	-0.494	Normalized positional residue frequency at helix termini N4'	[48]
OOBM850104	-0.467	Optimized average non-bonded energy per atom	[47]
AURR980120	-0.440	Normalized positional residue frequency at helix termini C4'	[48]
ZIMJ680101	-0.381	Hydrophobicity	[49]
CHOP780207	-0.366	Normalized frequency of C-terminal non helical region	[50]
AURR980117	-0.340	Normalized positional residue frequency at helix termini C'	[48]
VASM830101	-0.339	Relative population of conformational state A	[42]
RICJ880115	-0.333	Relative preference value at C-cap	[38]

2.4 Discussion

2.4.1 The most accurate regression model

We eventually selected the model with the highest prediction accuracy of 85.7% for the terminal cleavage site prediction of γ-secretase substrates. We examined the regression model with the prediction of terminal cleavage sites of two different substrates classes: APP mutants, intra-substrate changes (Fig. 7), and other γ-secretase substrates as inter-substrate alterations (Fig. 8). The model predicted both substrates' classes with more than 80% accuracy, indicating that it could distinguish both intra-substrate and inter-substrate differences of amino acids.

The best model reproduced well-known characteristic cleavages of the major γ-secretase substrate. First, the ε48-product-line cleavage, which starts at ε48 instead of at ε49, was reproduced (Fig. 6A and B). Second, the best regression model predicted the cleavage of Notch1, a major substrate of γ-secretase, to generate GCG, LLFFV, and AAFV peptides as byproducts. Strikingly, AAFV fits with the peptide between Nβ25 and 21 cleavages, reported two major Notch cleavages [35], further supporting the high accuracy of this model (Fig. 6C). Third, the best regression model could predict γ-secretase successive cleavage of a previously unreported substrate, rat APLP1 (Fig. 9). Thus, the best model can be utilized to predict how new substrates are successively cleaved. Moreover, the strategy with various machine learning algorithms combined with scanning cleavage-relevant parameters may be versatile for predictions of the cleavages by other intramembranous proteases with multiple cleavage sites [51, 52].

The best model was created by SBC regression. SBC is the abbreviation of Subtractive Clustering and Fuzzy c-Means Rules [33]. The SBC method employs fuzzy clustering to identify cluster centers from training data and make predictions by computing the distances between input and cluster centers. If an input is closer to a certain cluster, the output value is closer to the value of that cluster. By mapping training data into a feature space comprised of the 4+3 pocket model and PC9, SBC could find the well-cleaved and non-cleaved cluster centers. Therefore, SBC could differentiate between well-cleaved and non-cleaved input sequences, resulting in the highest percentage of accurate terminal cleavage predictions.

The model can be utilized to predict consecutive cleavage sites by γ-secretase. Predicting the consecutive cleavage sites of substrates in advance can reduce the trial-and-error process required to identify specific cleavage points in actual experiments. This not only enhances experimental efficiency but also contributes to the reduction of experimental volume and costs. For instance, to measure the cleavage fragments of APP containing 3 to 5 amino acid residues, which were used as the training data, it was necessary to prepare individual small-molecule peptides and create calibration curves for

each peptide measurement [12]. However, by performing cleavage site prediction in advance, it becomes possible to prepare peptides corresponding only to the predicted cleavage fragments, thereby achieving a reduction in experimental volume and costs.

Currently, there is a lack of comprehensive research on γ-byproducts for substrates other than APP. Leveraging this model is expected to predict consecutive cleavage points for substrates other than APP and further research on the generated γ-byproducts. The identification of consecutive cleavage points should serve as an essential foundation for considering factors governing consecutive cleavage, and elucidating what factors regulate a particular cleavage could potentially contribute to the development of AD treatments. For example, identifying factors that control specific cleavage points that produce Aβ42 and Aβ38 could help in the development of inhibitors and modulators that suppress the production of pathogenic Aβ products, leading to more effective therapeutic strategies.

2.4.2 PC9

PC9 may contain information relevant to γ-secretase proteolysis because of the following reasons. First, the best prediction model was generated using PC9. Second, the median of PC9 was the best compared to those of PCs in the goodness of fit to predict APP successive cleavages (Fig. 4). Third, the median of PC9 for predicting terminal cleavages was the best among PCs (Fig. 5). Thus, we calculated the correlation between PC9 and the original standardized amino acid index [18] and showed the top ten positive and negative amino acid indices.

Based on the factor loading analysis, the indices related to protein secondary structure correlated with PC9. Among the indices related to secondary structure, three indices, PALJ810107 [39], MAXF760106 [40], and VASM830102 [42], were positively correlated with PC9, while two indices, AURR980101 [48] and AURR980120 [48], were negatively correlated. PALJ810107 represents the frequency of amino acids appearing in regions composed solely of α-helix (all-alpha). This index can be interpreted as an indicator of the ease of α-helix formation for each amino acid. MAXF760106 represents the frequency of amino acids appearing in regions that are known to form α-helix structures when plotted using backbone dihedral angles. This index can be interpreted as the ease of α-helix formation for each amino acid. VASM830102 represents the frequency of amino acids appearing in regions known to form β-sheet structures when plotted using backbone dihedral angles. This index can be the ease of β-sheet formation for each amino acid.

Regarding negative factor loadings, AURR980120 represents the frequency of amino

acids appearing at the position C4', which is located four positions in the C direction from the C cap, the end of the C-terminal side of the α-helix structure. This index could be interpreted as an indicator of the difficulty of α-helix formation for each amino acid, as it exhibits a moderate inverse correlation with the frequency of amino acids appearing in positions within the helix, such as C3.

In summary, the interpretation of these indices suggests that the amino acid physicochemical properties represented by PC9 are related to the ease or difficulty of forming secondary structures such as α-helix and β-sheet. This may imply that PC9 contains important information on the transmembrane cleavage of γ-secretase because it cleaves its substrates in transmembrane regions, and the substrates are assumed to take α-helix conformations.

3. Development of a substrate cleavage site prediction model for γ-secretase and estimation of its cleavage mechanism

3.1 Introduction

We have estimated that the active site of γ-secretase possesses a 4+3 pocket structure, and the active site recognizes the physicochemical properties of amino acids associated with the secondary structure. However, this estimation does not provide information about the characteristics of each pocket. It is considered important to uncover the properties of each pocket as it would contribute to a deeper understanding of the cleavage mechanism of γ-secretase and potentially lead to the development of new drugs targeting γ-secretase. In this chapter, we investigated the amino acids recognized by each pocket of the model. We aimed to estimate the sequence specificity and make further estimations regarding the cleavage mechanism of γ-secretase.

3.2 Materials and Methods

3.2.1 Visualization of amino acid preference in the cleavage site region

The method of analyzing frequencies of amino acids which appear at each pocket is illustrated in Figure 10. The figure shows an example with three-sequence cleavage. First, we randomly generated a million sequences of ten amino acids (Step 1). Since cleavage occurs at two points: between a4 and a5, and between a7 and a8, we aligned the cleavage point of the pocket model to these two points and created amino acid vectors (Step 2). We converted these amino acid vectors using PC9 and input them into the model to predict the cleavage amounts of the sequences a5a6a7, then collected the sequences with cleavage amounts higher than the threshold of 20 fmol (Step 3). The collected sequences consist of two concatenated sequences from the two cleavage points. We decomposed them into two sequences and aligned them to the positions of pockets P4- ... -P1-P1'- ... -P3' (Step 4). The frequencies of amino acids in the positions were calculated. The z-normalized frequencies were visualized as heatmaps (Step 5). In the case of four-sequence cleavage, sets of 11 amino acids were generated, while for five-sequence cleavage, sets of 12 amino acids were generated. The amino acid occurrence frequency analysis at each pocket position was carried out using the same approach as in the case of three-sequence cleavage.

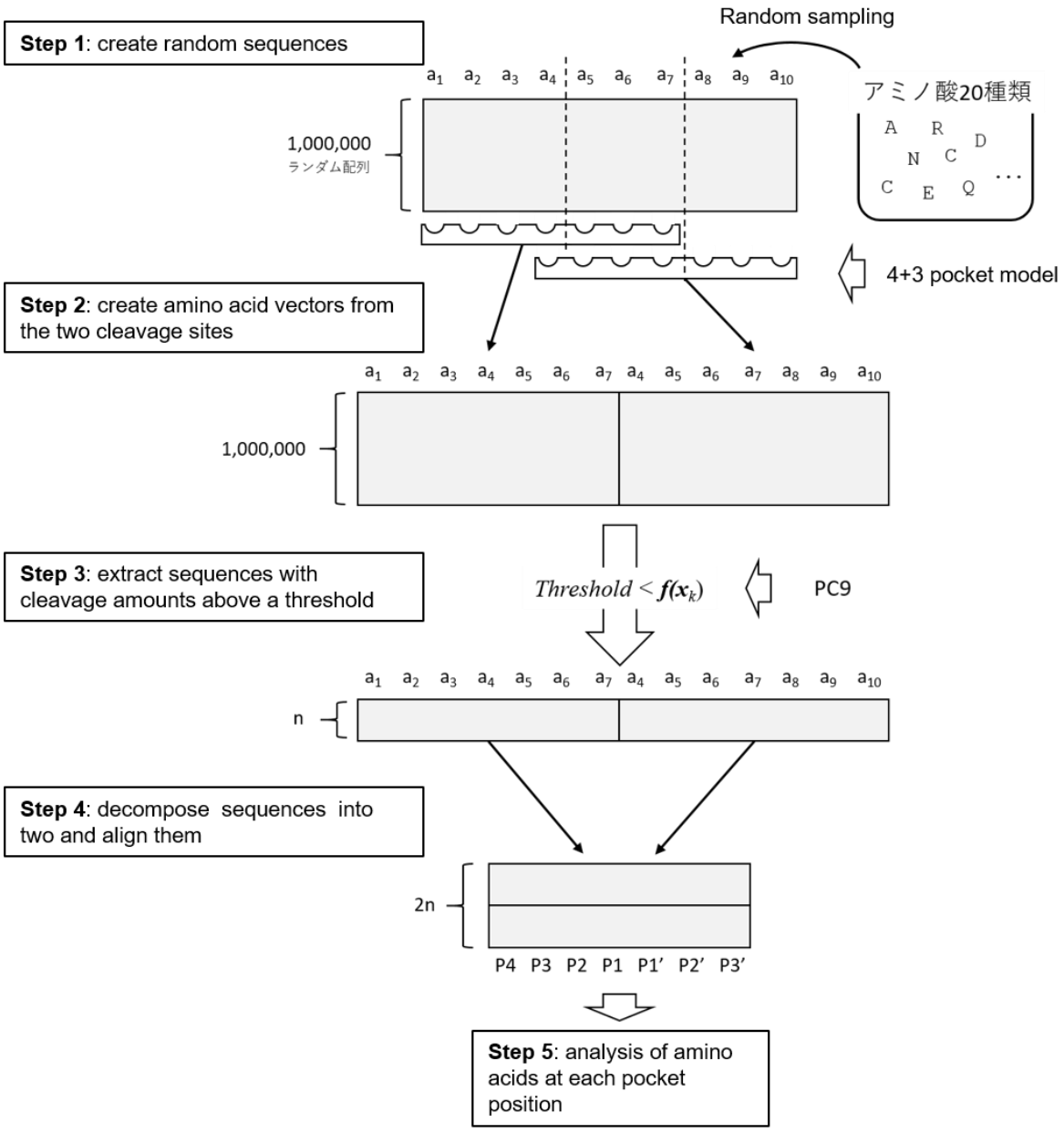

Figure 10. The method of analyzing frequencies of amino acids appearing at each pocket.

3.2.2 Biochemical experiments

Unless otherwise indicated, reagents were purchased from Sigma-Aldrich (St. Louise, MO). All experiments were performed at least 3 times and representative blots are shown in the figures.

3.2.2.1 cDNA constructs

The pcDNA3 neo::APPsw LV49/50IW mutant construct was generated by QuickChange-based mutagenesis with APPsw cDNA as a template using PrimeSTAR Max DNA Polymerase (Takara Bio, Kyoto, Japan, R045A).

3.2.2.2 Cell culture

HEK293 cells were cultured in DMEM (Nacalai tesque, Kyoto, Japan, 08458-16) containing 10% fetal bovine serum and 1% Penicillin/Streptomycin (Nacalai tesque, Kyoto, Japan, 09367-34). Transfection into HEK293 cells was performed using Lipofectamine 2000 (Invitrogen, Cleveland, CA, 11668019) according to the manufacturer's instructions.

Rat primary neuron cells (1x106, Lonza, Basel Switzerland, R-CX-500) were seeded on a poly-L-lysine coated dish (φ = 10 cm) and cultured in Neurobasal medium (Thermo Scientific, Waltham, MA, 21103049) with 2% B27 supplement (Thermo Scientific,17504044) and 2 mM GlutaMax-1 (Thermo Scientific, 35050061) for 14 days. Overnight culture supernatants of rat primary neuron cells were harvested, centrifuged at 2,500 x g for 5 minutes, and stored at -80 °C until use.

3.2.2.3 Generation of APP knockout (KO) cells using CRISPR/Cas9 genome editing

APPKO HEK293 cells were generated by CRISPR/Cas9 genome editing. The five 20-nucleotide guide sequences targeting human APP were designed using the CRISPR design tool at http://crispr.dbcls.jp/. Five guide RNA sequences for APP were cloned into pSpCas9(BB)-2A-Puro (pX459) V2.0, a gift from Feng Zhang (Addgene plasmid # 62988; http://n2t.net/addgene:62988 ; RRID:Addgene_62988) [53], using the following primers;

APP718-740f, caccgTCGGAACTTGTCAATTCCGC;

APP718-740r, aaacGCGGAATTGACAAGTTCCGAc;

APP917-939f, caccgAGAAGAAGCCGATGATGACG;

APP917-939r, aaacCGTCATCATCGGCTTCTTCTc;

APP1095-1117f, caccgCGGGAGATCATTGCTCGGCA;

APP1095-1117r, aaacTGCCGAGCAATGATCTCCCGc;

APP1162-1184f, caccgACGGCGGATGTGGCGGCAAC;

APP1162-1184r, aaacGTTGCCGCCACATCCGCCGTc;

APP2115-2137f, caccgGCAGCAGGGCGGGCATCAAC;

APP2115-2137r, aaacGTTGATGCCCGCCCTGCTGCc.

The five guide RNAs in the pX459 vector (1 μg) were co-transfected into HEK293 cells using Lipofectamine 2000 as above. Twenty-four hours post-transfection, the cells were treated with 1 μg/ml puromycin (Nacalai tesque, Kyoto, Japan, 29455-12) for another twenty-four hours. Cells were seeded on a 96-well plate by a limited dilution method to separate one cell per well. Single clones were expanded and screened for APP expression by immunoblotting analysis. Verified APPKO HEK293 cells were maintained in DMEM containing 10% fetal bovine serum and 1% Penicillin/Streptomycin.

3.2.2.4 Immunoblot analysis

APPKO HEK293 cells were transiently transfected with expression vectors carrying APPsw cDNA with or without the LV49/50IW mutation. Twenty-four hours after transfection, the medium was refreshed and further cultured for twenty-four hours. Forty-eight hours after transfection, cells were harvested with ice-cold PBS and frozen at -80 °C until use. Aliquots of stored cells were lysed in STEN-lysis buffer (50 mM Tris-HCl pH 7.6, 150 mM NaCl, 2 mM EDTA, 1% Nonidet P-40), and protein levels were normalized by protein assay (Nacalai tesque, Kyoto, Japan, 06385-00). Lysates were loaded on 10–20% Tris-Tricine gel (Thermo Fisher Scientific, MA), transferred to nitrocellulose membrane, and immunostained by APP C-terminal antibody Y188 (abcam, Cambridge, UK, ab32136) or anti-β-Actin (Santa Cruz Biotechnology, Dallas, TX, sc-47778). The immune reactive bands were detected by enhanced chemiluminescence (ECL; Cytiva, RPN2109, Boston MA), and signals were quantified by Amersham Imager 600 (Cytiva). The relative CTF was calculated by the band intensities of CTFs, the sum of α and β, normalized by the intensity of total APP. The

relative de novo AICD is calculated by the band intensity of AICD for 2 hours normalized by the CTFs band intensity in 0 hours.

3.2.2.5 Cell-free γ-secretase assay for the analysis of de novo AICD generation

Preparation of membrane fractions for cell-free γ-secretase assays was performed as described [54]. Briefly, HEK293 cell pellets harvested in PBS were resuspended in the hypotonic buffer (15 mM Na Citrate pH 6.4, 10 mM DTT, 1 mM EDTA, 1x Complete Mini Protease Inhibitor Cocktail (Roche, Basel, Switzerland, 04693124001)) and incubated on ice for 30 minutes. Samples were frozen in liquid nitrogen, subsequently thawed in the water bath at room temperature, and kept on ice. The cell suspensions were centrifuged at 2,500 x g for 20 minutes, and the resultant supernatant was supplemented with 5% glycerol and further ultracentrifuged at 100,000 x g for 1 hour. The resulting membrane pellets were stored at -80 °C until use. The cell-free assay was performed as described [36, 55]. Briefly, cellular membranes from 6 dishes (φ10 cm) were incubated in 100 µl of reaction buffer (150 mM Na Citrate pH 6.4, 5 mM 1,10-Phenanthroline, 5x Complete Mini Protease Inhibitor Cocktail) at 37 °C for 2 hours.

3.2.2.6 MALDI-TOF mass spectrometry (MS) analysis

Immunoprecipitation combined with mass spectrometry was performed as described [30, 50]. For the detection of APL1β species, 6 ml of the thawed medium was supplemented with 300 µl of 1 M Tris-HCl pH 7.4, 60 µl of 0.5 M EDTA, Protease Inhibitor Cocktail (Sigma-Aldrich, St Louis, MO p8340), 4 µl of OA858, an antibody against full-length APL1β25 [36], and 10 µl of Protein A Sepharose CL-4B (GE, Boston MA, 17-0780-01) and rotated overnight. For the detection of de novo generated AICD species, samples after the cell-free assay were sonicated 3 times on ice for 10 seconds and ultracentrifuged at 100,000 x g at 4 °C. The resultant supernatants were incubated with 20 µl Protein A Sepharose 4 Fast Flow (Cytiva, 17061801) and 2.5 µl of Y188 antibody and rotated overnight. Beads were washed three times with IP/MS wash buffer (10 mM Tris-HCl pH 8.0, 140 mM NaCl, 1 mM EDTA, 0.1% n-octylglucoside) and once with pure water. Immunoprecipitates were eluted by TWA (2.5% Trifluoroacetic acid (Nacalai tesque, 34901-21)), 50% Acetonitrile (Nacalai tesque, 00404-75) in Water) saturated with α-cyano-4-hydroxycinnamic acid (Nacalai tesque, 06700-21) and loaded on a stainless plate (MSP96 target ground steel BC, Bruker Daltonics, Bremen, Germany, 8280799) and analyzed by MALDI-TOF MS (Microflex, Bruker Daltonics). AICD peaks were measured with the reflect mode. Python 3.8.11 and matplotlib 3.4.2 were used for plotting mass spectrums.

3.3 Results

The most accurate regression model was generated by the SBC machine learning method. The characteristic point of SBC algorithm is finding the clusters, one of which can be the favorable sequence in the cleavage site region in the case of proteases. Thus, we hypothesized that the most accurate regression model would intrinsically contain some sequence features around the cleavage site. To extract such sequence features from the most accurate model, we conducted in silico cleavage of random sequence peptides (also see Materials and Methods). Three different lengths (10–12 amino acids) of random peptides, which include 4 non-prime and 3 prime sites of two cleavages generating 3–5 amino acid long γ-byproducts, were generated separately. We limited the number of random peptides to a million because it is technically not feasible to analyze all peptides (20^{10}–20^{12}) in a reasonable time. The best regression model processed random sequence peptides to obtain their amount of cleavage. Frequencies of the amino acids at individual locations Ri in random peptides that are calculated to be cleaved more than 20 fmol in silico were visualized by the heatmaps (Fig. 11A). Surprisingly, the order of the frequencies of amino acids in most locations was along with that of PC9. Notably, R6 and R7 in three-amino acid spaced cleavage, R7 and R8 in four-amino acid spaced cleavage, and R8 and R9 in five-amino acid spaced cleavage commonly showed more abundance of tryptophan (W), threonine (T), serine (S), and alanine (A), indicating a common amino acid preference near the scissile bond regardless of the number of amino acids spaced between the two cleavages. Peptide regions R4–R7 for the three-amino acid spaced cleavage, R5–R7 for the four-amino acid spaced cleavage, and R6 and R7 for the five-amino acid spaced cleavage are overlapping regions of the cleavage site regions of two cleavages (Fig. 11A). Thus, to better visualize the sequence features of the cleavage site regions, we merged two cleavage site regions. As shown in Fig. 11B, in all 3–5 amino acids spaced cleavages, P4, P1, and P3' showed more abundance of tryptophan (W), threonine (T), serine (S), and alanine (A), whereas P1' showed less emergence of these amino acids, indicating some preferences of γ-secretase cleavages in the cleavage site region.

To examine the above findings in the computational analysis, we conducted biochemical experiments whether the unpreferable amino acid sequence is dodged by γ-cleavage. The P1 and P1' amino acids at ε-49 (Aβ numbering), the major APP initial cleavage site, are L and V, respectively. The latter amino acid, V in P1' is a relatively favorable amino acid, whereas L in P1 is not an unfavorable amino acid in the computational results in Fig. 11B. Thus, to test our computational result, we prepared APP LV49/50IW, the worst preferable cleavage site mutants at P1 and P1' in the computational analysis (Fig. 12A). As expected, the γ-secretase cleavage efficiency was reduced by 75% in LV49/50IW (Fig. 12B–E), indicating that cleavage at ε-49 in LV49/50IW was indeed

not preferable. To investigate this finding further, we analyzed cleavage sites of AICD by MALDI-TOF MS. Notably, the major AICD cleavage site was not at ε-49 (Fig. 12D) revealing cleavage at ε-49 in LV49/50IW was indeed unpreferable. However, the cleavage was shifted to another site as shown by the appearance of a major peak in LV49/50IW at ε-47. This shifted ε-cleavage was however not efficient enough to compensate the strong loss in total cleavage specificity, likely because IT at P1/P1´for ε-47 was also not preferred. Taken together, these data validate the predicted preferences in the P1 and P1' sites of γ-secretase cleavage in APP.

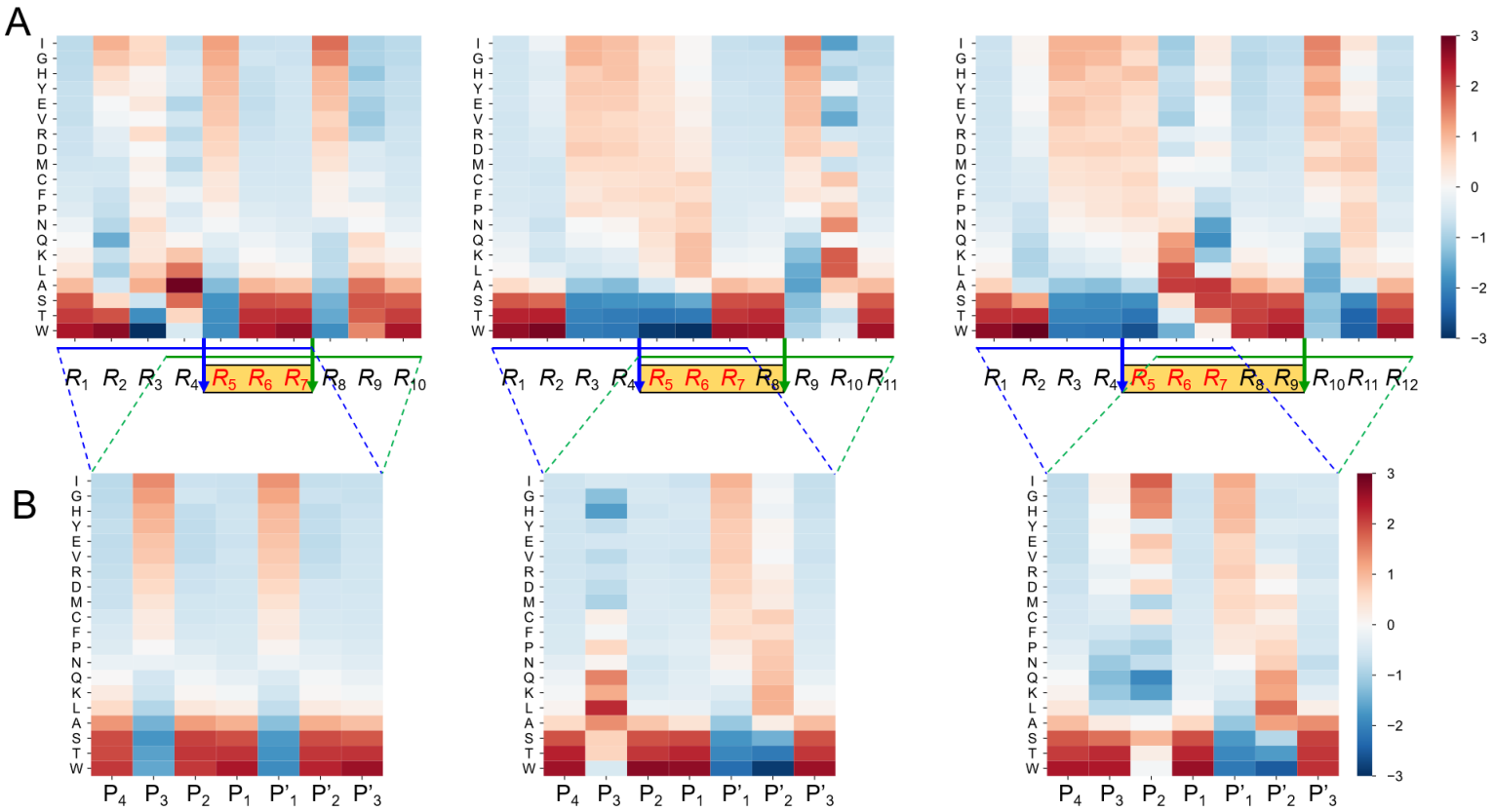

Figure 11. Sequence feature of γ-secretase cleavage extracted from the best model. Heatmap visualization of frequencies of amino acids in random peptides whose cleavage products in silico were more than 20 fmol by the best model following z-score transformation. PC9 and L = 4 are the components of the best regression model. Heat maps for cases of 3–5 amino acid spaced cleavages were separately shown from the left to the right panel. The frequencies of amino acids in the random peptides were shown in (A). The frequencies of two consecutive cleavage sites in (A) were merged in (B). Note that the common preferences in P4, P1, P1', and P3' were reproduced in additional ten independent experiments.

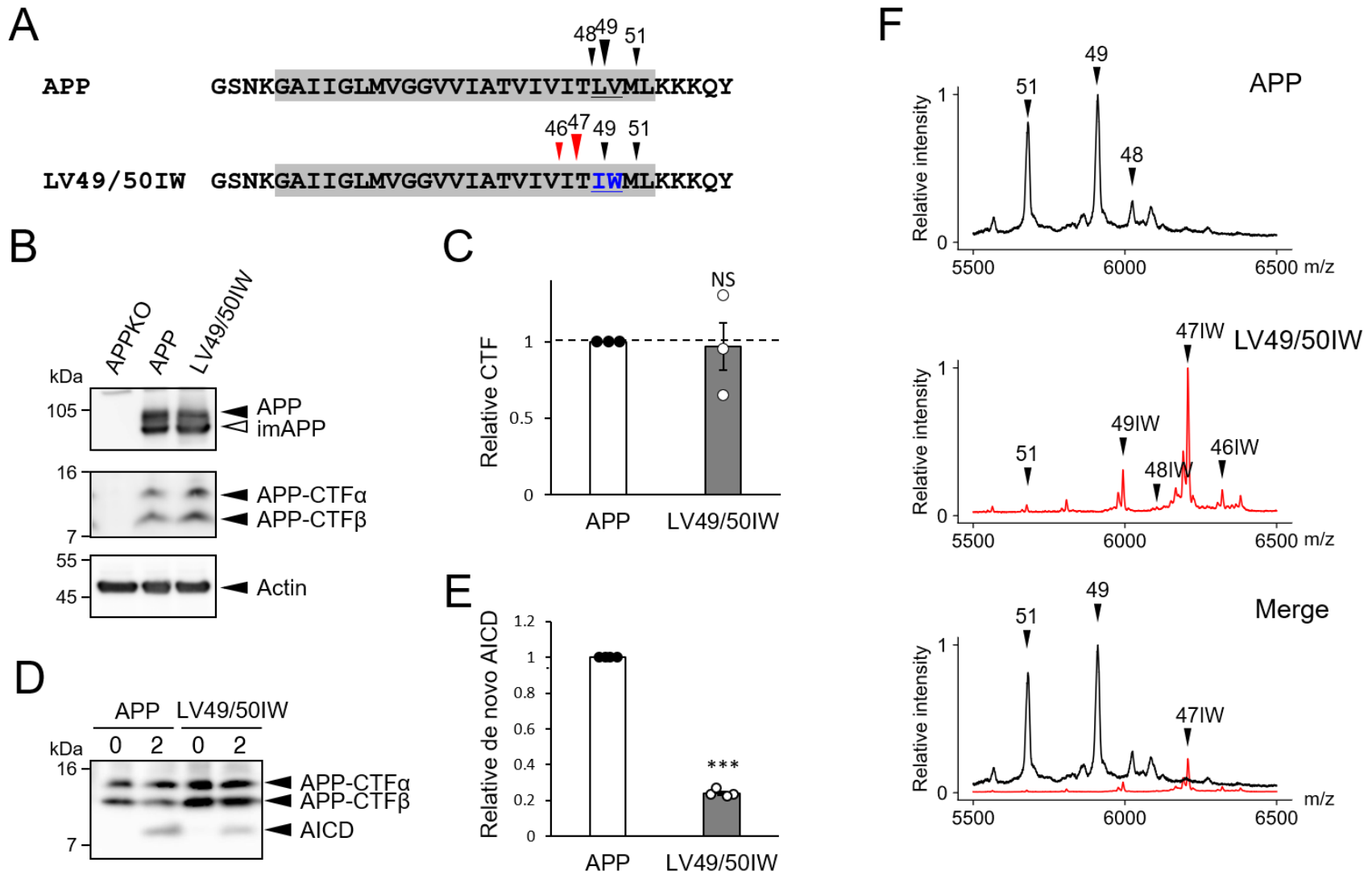
A
48 49 51
APP GSNKGAIIGLMVGGVVIATVIVITLVMLKKKQY
46 47 49 51
LV49/50IW GSNKGAIIGLMVGGVVIATVIVITIWMLKKKQY
B
APPKO
APP
LV49/50IW
kDa
105
APP
imAPP
16
APP-CTFα
APP-CTFβ
7
55
45
Actin
C
Relative CTF
NS
1
0.5
0
APP
LV49/50IW
D
APP LV49/50IW
kDa
0 2 0 2
16
APP-CTFα
APP-CTFβ
AICD
7
E
Relative de novo AICD
1.2
1
0.8
0.6
0.4
0.2
0
***
APP
LV49/50IW
F
APP
Relative intensity
51
49
48
1
0
5500
6000
6500 m/z
LV49/50IW
47IW
49IW
46IW
48IW
51
Merge
49
51
47IW

Figure 12. Unpreferable mutations in P1 and P1' of APP cleavage shifted γ-secretase cleavage site. Sequences around transmembrane domains of APP wild type and LV49/50IW were shown in (A). The mutated amino acids were underlined and shown in blue. Bigger arrowheads indicate the major cleavage sites. Red arrowheads are cleavage sites that emerged in the mutant. Effects of the LV49/50IW mutation on APP metabolism were analyzed by the immunoblot of cell lysates (B) and their quantitation (C). de novo Generated AICD species from γ-secretase cleavage assay were analyzed by immunoblot (D), their quantitation (E), and mass spectrometry (F). Quantification data are shown as mean ± SEM (n = 5). 'NS' indicates not significant, whereas asterisks indicate significant difference, ***, $p < 0.001$ (two-tailed paired t-test). Observed m/z and theoretical molecular mass of AICDs are shown in Table 8 and 9.

Table 8. Observed m/z and theoretical molecular mass of AICDs

AICD species	sequence	theoretical mass	observed m/z
AICDε48	L^{645}VMLKKKQYTSIHHGVVEVDAAVTPEERHLSKMQQNGYENPTYKFFEQMQN^{695}	6023.9	6022.9
AICDε49	V^{646}MLKKKQYTSIHHGVVEVDAAVTPEERHLSKMQQNGYENPTYKFFEQMQN^{695}	5910.7	5908.9
AICDε51	L^{648}KKKQYTSIHHGVVEVDAAVTPEERHLSKMQQNGYENPTYKFFEQMQN^{695}	5680.4	5679.0

Table 9. Observed m/z and theoretical molecular mass of AICDs derived from APP LV49/50IW

AICD species	sequence	theoretical mass	observed m/z
AICDε46 IW	I^{643}TIWMLKKKQYTSIHHGVVEVDAAVTPEERHLSKMQQNGYENPTYKFFEQMQN^{695}	6325.2	6324.0
AICDε47 IW	T^{644}IWMLKKKQYTSIHHGVVEVDAAVTPEERHLSKMQQNGYENPTYKFFEQMQN^{695}	6212.0	6210.1
AICDε49 IW	W^{646}MLKKKQYTSIHHGVVEVDAAVTPEERHLSKMQQNGYENPTYKFFEQMQN^{695}	5997.8	5995.5
AICDε50	M^{647}LKKKQYTSIHHGVVEVDAAVTPEERHLSKMQQNGYENPTYKFFEQMQN^{695}	5811.6	5809.3
AICDε51	L^{648}KKKQYTSIHHGVVEVDAAVTPEERHLSKMQQNGYENPTYKFFEQMQN^{695}	5680.4	5678.1

3.4 Discussion

in silico Cleavage analysis of random peptides revealed amino acid preferences of γ-secretase cleavage. Tryptophan (W), threonine (T), serine (S), and alanine (A) in P4, P1, and P3' positions and isoleucine (I), glycine (G), histidine (H), tyrosine (Y), glutamate (E), and valine (V) in P1' were commonly observed in 3–5 amino acids spaced cleavages (Fig. 11B). Interestingly, this relative valine preference in P1' is consistent with an earlier hypothesis of γ-secretase substrate research [57, 58]. Moreover, the cleavage preference shown in Fig. 11B may offer an explanation of the cleavage pattern of the familial AD London mutant. The reduction of ε49-cleavage of the London mutant (V46I) [59] may be attributed to the isoleucine substitution, which is the most unpreferable amino acid in P4 of the ε49-cleavage.

Thus, we envisaged biochemically examining computational findings by analyzing the worst preferable cleavage sites mutant LV49/50IW. To our knowledge, the ε-cleavage in APP has long been thought to be rather permissive to mutagenesis [23, 57, 60]. However, LV49/50IW showed a drastic shift of major cleavage from ε49 to ε47 (Fig. 12D). Both cleavages were at non-preferred P1/P1´sites, which is consistent with the observed strong drop in total cleavage efficiency of the LV49/50IW mutant. These results indicate preferences of the amino acids around the ε-cleavage site of γ-secretase cleavage. Previously Bolduc et al. proposed that the large-small-large S1'-S2'-S3' pockets in γ-secretase govern the specificity of cleavage site usage in APP [20]. One may think our finding that tryptophan residue does not preferentially occur at P1' may contradict their proposal; S1' pocket is large enough to accept bulky amino acids. Since V50W or M51W mutants allowed normal γ-secretase cleavage as assessed by unchanged Aβ formation [8], we speculate that the P1/P1' combination of I49W50 is problematic with accommodation of these side chains into the corresponding subsite pockets. Conversely, isoleucine at P1 alone also may not be the cause of this strong reduction of the cleavage because isoleucine at P1 was shown to occurs in two substrates EphB2 and CD44 [26, 61].

The estimated amino acid preferences of γ-secretase cleavage are expected to be used to gain knowledge for the development of new γ-secretase inhibitors. Several inhibitors such as E2012 and Nirogacestat have been reported to be Notch sparing [62]. By comparing these inhibitors with the estimated amino acid preferences of γ-secretase cleavage, it might be possible to infer specific properties associated with inhibitors that demonstrate Notch sparing. If the essential features for Notch sparing can be deduced, they could be utilized in the design of inhibitors with such properties. In addition, if the features necessary for Notch sparing are known, it may be possible to narrow down the search space, which may facilitate the screening of inhibitors using a computer.

Based on our finding of the sequence preferences around the cleavage site of γ-secretase, we propose a possible mechanistic model of γ-secretase cleavage (Fig. 13). After substrates encounter γ-secretase, they undergo their extracellular domain size-selection by Nicastrin [63] and/or with the recognition by this exosite [64]. Then, substrates are translocated to the active site to accommodate amino acids around the membrane cytosolic border in the catalytic pocket, thereby undergoing one more selection with side-chain compatibility. Among the peptide bonds around the membrane cytosolic border of substrates, better preferable sites are selected for catalysis to occur.

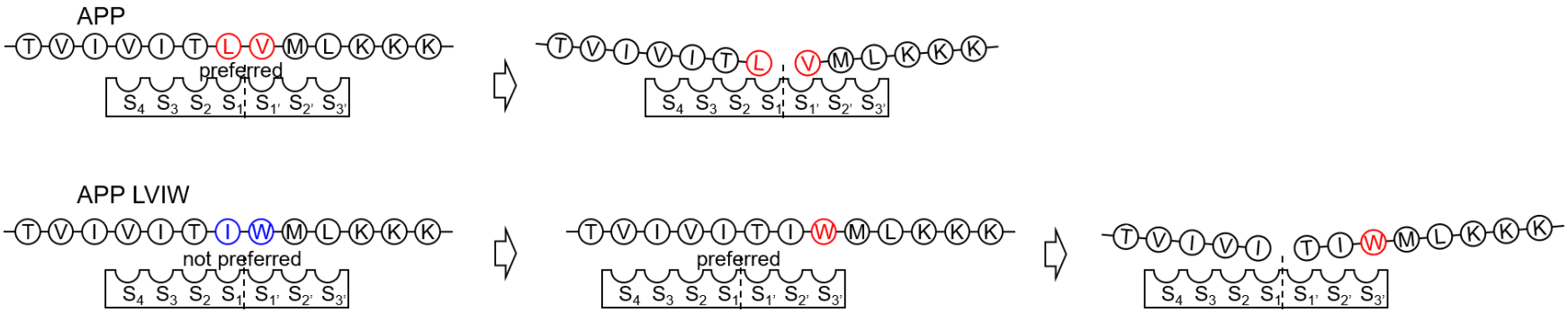

Figure 13. Model of substrate cleavage site selection in the catalytic site of γ-secretase. After the cleavage site region of APP enters the catalytic pocket of γ-secretase, a relatively preferred amino acid sequence is chosen depending on its compatibility. In the wild-type APP, L49/V50 is frequently chosen, as this amino acids pair is compatible with the catalytic site. On the other hand, I49/W50 in the APP LV49/50IW is not compatible with the active site pocket of γ-secretase, leading to a selection of I47/T48 for an alternative cleavage site. Red and blue letters indicate preferable and unpreferable amino acids in the cleavage site regions, respectively.

4. Conclusion

Elucidating the substrate cleavage mechanism of γ-secretase is important as it may lead to the development of preventive drugs for AD. Therefore, in this study, we estimated the number of pockets in the active site of γ-secretase, the physicochemical properties of the amino acids recognized by these pockets, and the amino acid preference for each pocket. By combining six pocket models, ten principal components (PC1 to PC10) made from the AAindex, and 88 regression methods, we created 5,280 regression models using the amount of cleaved APP fragments as training data. To select a model that reflects the continuous cleavage features of γ-secretase, we collected 35 cleavage sites from substrates for which the cleavage initiation and termination sites by γ-secretase were known and performed cleavage site prediction using the regression models. As a result, the 4+3 pocket-PC9-SBC model showed the highest prediction accuracy of 85.7%. Factor loading analysis between PC9 and AAindex was conducted to investigate the AAindex that strongly influenced PC9. The results revealed that indices related to protein secondary structures affect PC9. These findings estimated that the cleavage active site of γ-secretase consists of four continuous pockets on the N-terminal side and three on the C-terminal side. These pockets recognize physicochemical information related to the secondary structure of amino acids in the substrate sequence to determine the cleavage. Furthermore, in silico cleavage analysis of random peptides using the cleavage model predicted amino acid preferences for the P1 and P1', which were confirmed through biochemical experiments. Further investigation of the model is expected to promote a fundamental understanding of the cleavage mechanism of γ-secretase and provide valuable insights for developing γ-secretase inhibitors and other related research.

Acknowledgements

I could proceed with this research thanks to the guidance and support of many individuals. I would like to express my deepest gratitude to Professor Shigehiko Kanaya for providing me with detailed and thoughtful guidance and an excellent research environment throughout this study. Despite being very busy, Professor Kanaya always took the time to meeting, and I am genuinely thankful for that.

I am also profoundly grateful to Professor Akio Fukumori from Osaka Medical and Pharmaceutical University for his invaluable advice and enthusiastic guidance, which were highly beneficial to me as someone with no prior knowledge in the field.

I sincerely appreciate Professor Keiichi Yasumoto for his valuable insights during the research presentations.

I would like to express my sincere gratitude to Associate Professor MD. Altaf-Ul-Amin, Associate Professor Naoaki Ono and Assistant Professor Ming Huang for their invaluable opinions shared during laboratory seminars and regular discussions.

I sincerely thank Professor Satoshi Nakamura for giving me the opportunity to start this research. Through Professor Nakamura's introduction, I came into contact with Professor Fukumori, and that's how this research began. I am genuinely grateful for that.

I would also like to express my appreciation to everyone in the Computational Systems Biology Laboratory for their support in the research and various aspects of my daily life.

I would like to express my deep gratitude to Ayaho Shimizu, Mahiro Miyashita, Associate Professor. Kanta Yanagida from Osaka Medical and Pharmaceutical University for conducting biochemical experiments, Associate Professor Shinji Tagami and Dr. Masayasu Okochi from Osaka University for providing APP γ-byproduct quantification data.

Finally, I would like to thank my wife and parents for supporting my research activities

References

[1] E. Nichols, *et al.*, Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. *Lancet Public Heal.* **7**, e105–e125 (2022).

[2] N. Nakahori, *et al.*, Future projections of the prevalence of dementia in Japan: results from the Toyama Dementia Survey. *BMC Geriatr.* **21**, 1–10 (2021).

[3] Alzheimer's Association. 2020 Alzheimer's disease facts and figures. *Alzheimer's Dement.* **16**, 391–460 (2020).

[4] C. Reitz, Alzheimer's Disease and the Amyloid Cascade Hypothesis: A Critical Review. *Int. J. Alzheimers. Dis.* **2012**, 1–11 (2012).

[5] B. De Strooper, T. Iwatsubo, M. S. Wolfe, Presenilins and γ-secretase: structure, function, and role in Alzheimer Disease. *Cold Spring Harb. Perspect. Med.* **2**, a006304 (2012).

[6] M. Takami, *et al.*, γ-Secretase: Successive tripeptide and tetrapeptide release from the transmembrane domain of β-carboxyl terminal fragment. *J. Neurosci.* **29**, 13042–13052 (2009).

[7] N. Matsumura, *et al.*, γ-Secretase Associated with Lipid Rafts. *J. Biol. Chem.* **289**, 5109–5121 (2014).

[8] D. M. Bolduc, D. R. Montagna, M. C. Seghers, M. S. Wolfe, D. J. Selkoe, The amyloid-beta forming tripeptide cleavage mechanism of γ-secretase. *Elife* **5**, 1–21 (2016).

[9] K. Sharma, Cholinesterase inhibitors as Alzheimer's therapeutics (Review). *Mol. Med. Rep*. **20**, 1479–1487 (2019).

[10] B. Reisberg, et al., Memantine in Moderate-to-Severe Alzheimer's Disease. *N. Engl. J. Med*. **348**, 1333–1341 (2003).

[11] J. E. Luo, Y. M. Li, Turning the tide on Alzheimer's disease: modulation of γ-secretase. *Cell Biosci*. **12**, 1–12 (2022).

[12] S. Tagami, *et al.*, Semagacestat Is a Pseudo-Inhibitor of γ-Secretase. *Cell Rep.* **21**, 259–273 (2017).

[13] J. Sevigny, et al., The antibody aducanumab reduces Aβ plaques in Alzheimer's

disease. *Nature* **537**, 50–56 (2016).

[14] A. Mullard, Landmark Alzheimer's drug approval confounds research community. *Nature* **594**, 309–310 (2021).

[15] FDA Converts Novel Alzheimer's Disease Treatment to Traditional Approval. U.S. Food and Drug Administration. 2023. [Accessed on 28 September 2023]. Available from: https://www.fda.gov/news-events/press-announcements/fda-converts-novel-alzheimers-disease-treatment-traditional-approval

[16] C. H. van Dyck, et al., Lecanemab in Early Alzheimer's Disease. *N. Engl. J. Med.* **388**, 9–21 (2023).

[17] I. Schechter, A. Berger, On the size of the active site in proteases. I. Papain. *Biochem. Biophys. Res. Commun.* **27**, 157–162 (1967).

[18] S. Kawashima, H. Ogata, M. Kanehisa, AAindex: Amino Acid Index Database. *Nucleic Acids Res.* **27**, 368–369 (1999).

[19] M. Kuhn, Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **28**, 1–26 (2008).

[20] Y. Piao, *et al.*, Mechanism of Intramembrane Cleavage of Alcadeins by γ-Secretase. *PLoS One* **8**, e62431 (2013).

[21] S. Hogl, P.-H. Kuhn, A. Colombo, S. F. Lichtenthaler, Determination of the Proteolytic Cleavage Sites of the Amyloid Precursor-Like Protein 2 by the Proteases ADAM10, BACE1 and γ-Secretase. *PLoS One* **6**, e21337 (2011).

[22] Y. Qi-Takahara, *et al.*, Longer Forms of Amyloid β Protein: Implications for the Mechanism of Intramembrane Cleavage by γ-Secretase. *J. Neurosci.* **25**, 436–445 (2005).

[23] M. Sastre, *et al.*, Presenilin-dependent γ-secretase processing of β-amyloid precursor protein at a site corresponding to the S3 cleavage of Notch. *EMBO Rep.* **2**, 835–841 (2001).

[24] Y. Gu, *et al.*, Distinct Intramembrane Cleavage of the β-Amyloid Precursor Protein Family Resembling γ-Secretase-like Cleavage of Notch. *J. Biol. Chem.* **276**, 35235–35238 (2001).

[25] S. Lammich, *et al.*, Presenilin-dependent Intramembrane Proteolysis of CD44 Leads to the Liberation of Its Intracellular Domain and the Secretion of an Aβ-

like Peptide. *J. Biol. Chem.* **277**, 44754–44759 (2002).

[26] I. Okamoto, *et al.*, Proteolytic release of CD44 intracellular domain and its role in the CD44 signaling pathway. *J. Cell Biol.* **155**, 755–762 (2001).

[27] M. A. Fernandez, *et al.*, Transmembrane Substrate Determinants for γ-Secretase Processing of APP CTFβ. *Biochemistry* **55**, 5675–5688 (2016).

[28] T. Tsaktanis, *et al.*, Cleavage and Cell Adhesion Properties of Human Epithelial Cell Adhesion Molecule (HEPCAM). *J. Biol. Chem.* **290**, 24574–24591 (2015).

[29] D. Fleck, *et al.*, Proteolytic Processing of Neuregulin 1 Type III by Three Intramembrane-cleaving Proteases. *J. Biol. Chem.* **291**, 318–333 (2016).

[30] M. Okochi, *et al.*, Presenilins mediate a dual intramembranous γ-secretase cleavage of Notch-1. *EMBO J.* **21**, 5408–5416 (2002).

[31] S. Tagami, *et al.*, Regulation of Notch Signaling by Dynamic Changes in the Precision of S3 Cleavage of Notch-1. *Mol. Cell. Biol.* **28**, 165–176 (2008).

[32] Y. Ran, *et al.*, γ-Secretase inhibitors in cancer clinical trials are pharmacologically and functionally distinct. *EMBO Mol. Med.* **9**, 950–966 (2017).

[33] L. S. Riza, C. Bergmeir, F. Herrera, J. M. Benítez, frbs : Fuzzy Rule-Based Systems for Classification and Regression in R. *J. Stat. Softw.* **65**, 1–30 (2015).

[34] M. Morishima-Kawashima, Molecular mechanism of the intramembrane cleavage of the β-carboxyl terminal fragment of amyloid precursor protein by γ-secretase. *Front. Physiol.* **5**, 463 (2014).

[35] M. Okochi, *et al.*, Secretion of the Notch-1 Aβ-like Peptide during Notch Signaling. *J. Biol. Chem.* **281**, 7890–7898 (2006).

[36] K. Yanagida, *et al.*, The 28-amino acid form of an APLPl-derived Aβ-like peptide is a surrogate marker for Aβ42 production in the central nervous system. *EMBO Mol. Med.* **1**, 223–235 (2009).

[37] L. Schauenburg, *et al.*, APLP1 is endoproteolytically cleaved by γ-secretase without previous ectodomain shedding. *Sci. Rep.* **8**, 1916 (2018).

[38] J. S. Richardson, D. C. Richardson, Amino acid preferences for specific locations at the ends of alpha helices. *Science (80-.).* **240**, 1648–1652 (1988).

[39] J. Palau, P. Argos, P. Puigdomenech, Protein secondary structure. Studies on the

limits of prediction accuracy. *Int. J. Pept. Protein Res.* **19**, 394–401 (1982).

[40] F. R. Maxfield, H. A. Scheraga, Status of empirical methods for the prediction of protein backbone topography. *Biochemistry* **15**, 5138–5153 (1976).

[41] J. P. Garel, D. Filliol, P. Mandel, Coefficients de partage d'aminoacides, nucléobases, nucléosides et nucléotides dans un système solvant salin. *J. Chromatogr. A* **78**, 381–391 (1973).

[42] M. Vasquez, G. Nemethy, H. A. Scheraga, Computed conformational states of the 20 naturally occurring amino acid residues and of the prototype residue α-aminobutyric acid. *Macromolecules* **16**, 1043–1049 (1983).

[43] P. H. A. Sneath, Relations between chemical structure and biological activity in peptides. *J. Theor. Biol.* **12**, 157–195 (1966).

[44] M. J. Geisow, R. D. B. Roberts, Amino acid preferences for secondary structure vary with protein class. *Int. J. Biol. Macromol.* **2**, 387–389 (1980).

[45] N. Qian, T. J. Sejnowski, Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* **202**, 865–884 (1988).

[46] M. Prabhakaran, P. K. Ponnuswamy, Shape and Surface Features of Globular Proteins. *Macromolecules* **15**, 314–320 (1982).

[47] M. Oobatake, Y. Kubota, T. Ooi, Optimization of Amino Acid Parameters for Correspondence of Sequence to Tertiary Structures of Proteins. *Bull. Inst. Chem. Res., Kyoto Univ* **63**, 82–94 (1985).

[48] R. Aurora, G. D. Rose, Helix capping. *Protein Sci.* **7**, 21–38 (1998).

[49] J. M. Zimmerman, N. Eliezer, R. Simha, The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.* **21**, 170–201 (1968).

[50] P. Y. Chou, G. D. Fasman, "Prediction of the secondary structure of proteins from their amino acid sequence" in *Advances in Enzymology and Related Areas of Molecular Biology*, A. Meister, Ed. (John Wiley & Sons, Inc., 1978), pp. 45–148.

[51] R. Fluhrer, *et al.*, A γ-secretase-like intramembrane cleavage of TNFα by the GxGD aspartyl protease SPPL2b. *Nat. Cell Biol.* **8**, 894–896 (2006).

[52] K. Matsuhisa, *et al.*, Production of BBF2H7-derived small peptide fragments via endoplasmic reticulum stress-dependent regulated intramembrane proteolysis.

FASEB J. **34**, 865–880 (2020).

[53] F. A. Ran, *et al.*, Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).

[54] E. Winkler, *et al.*, Purification, pharmacological modulation, and biochemical characterization of interactors of endogenous human γ-secretase. *Biochemistry* **48**, 1183–1197 (2009).

[55] A. Fukumori, *et al.*, Presenilin-dependent gamma-secretase on plasma membrane and endosomes is functionally distinct. *Biochemistry* **45**, 4907–4914 (2006).

[56] J. Trambauer, A. Fukumori, B. Kretner, H. Steiner, Analyzing Amyloid-β Peptide Modulation Profiles and Binding Sites of γ-Secretase Modulators. *Methods Enzymol.* **584**, 157–183 (2017).

[57] A. J. Beel, C. R. Sanders, Substrate specificity of γ-secretase and other intramembrane proteases. *Cell. Mol. Life Sci.* **65**, 1311–1334 (2008).

[58] D. Y. Kim, L. A. M. K. Ingano, D. M. Kovacs, Nectin-1α, an Immunoglobulin-like Receptor Involved in the Formation of Synapses, Is a Substrate for Presenilin/γ-Secretase-like Cleavage. *J. Biol. Chem.* **277**, 49976–49981 (2002).

[59] K. Mori, *et al.*, The production ratios of AICDε51 and Aβ42 by intramembrane proteolysis of βAPP do not always change in parallel. *Psychogeriatrics* **10**, 117–123 (2010).

[60] A. Weidemann, *et al.*, A novel epsilon-cleavage within the transmembrane domain of the Alzheimer amyloid precursor protein demonstrates homology with Notch processing. *Biochemistry* **41**, 2825–35 (2002).

[61] C. Litterst, *et al.*, Ligand Binding and Calcium Influx Induce Distinct Ectodomain/γ-Secretase-processing Pathways of EphB2 Receptor. *J. Biol. Chem.* **282**, 16155–16163 (2007).

[62] F. Panza, et al., γ-secretase inhibitors for the treatment of Alzheimer's disease: The current state. *CNS Neurosci. Ther.* **16**, 272–284 (2010).

[63] D. M. Bolduc, D. R. Montagna, Y. Gu, D. J. Selkoe, M. S. Wolfe, Nicastrin functions to sterically hinder γ-secretase–substrate interactions driven by substrate transmembrane domain. *Proc. Natl. Acad. Sci.* **113**, E509–E518 (2016).

[64] A. Fukumori, H. Steiner, Substrate recruitment of γ-secretase and mechanism of clinical presenilin mutations revealed by photoaffinity mapping. *EMBO J.* **35**, 1628–1643 (2016).