Universidad del País Vasco
Euskal Herriko Unibertsitatea

ZIENTZIA ETA TEKNOLOGIA FAKULTATEA
FACULTAD DE CIENCIA Y TECNOLOGÍA

Bachelor's Thesis

BSc in Biochemistry and Molecular Biology

Machine learning in drug discovery: Targeting and validating PTEN interaction with PSD-95 in Alzheimer's disease

Author:
Diego Asua Corcóstegui

Director:
Shira Knafo

Leioa, 20th June 2017

CONTENTS

1 Introduction **1**

1.1 Anatomy and physiology of the hippocampus 1

1.2 Implications of PDZ-domain driven interactions in synaptic plasticity: Long-term depression 2

1.3 PTEN in Alzheimer's disease 3

2 Objectives **3**

3 Materials and methods **3**

3.1 Human brain tissue 3

3.2 Tissue processing 4

3.3 Immunohistochemistry 4

3.4 Confocal imaging and quantification 5

3.5 Thioflavin-S staining 5

3.6 Computational models 5

3.7 Data mining and machine learning 6

3.8 Network analysis 6

3.9 Statistical analysis 7

4 Results **7**

4.1 Thioflavin-S positive Aβ plaques 7

4.2 Immunohistochemistry and confocal imaging 8

4.3 Development and validation of a computational model 9

5 Discussion **11**

6 Concluding remarks **13**

7 Acknowledgments **13**

8 References **13**

9 Appendix I: Supplementary data **17**

1 INTRODUCTION

1.1 Anatomy and physiology of the hippocampus

The central nervous system receives, integrates and sends information, and among many other tasks, it stores information by creating new memory traces and updating already existing ones. In fact, memory trace formation is accomplished by specialized structures in the brain, and among them, the hippocampus (Andersen, Morris, Amaral, Bliss, & O'Keefe, 2007). This structure has been extensively studied by many neuroanatomists throughout more than a century, and among them, it is worth mentioning Santiago Ramón y Cajal's contributions. In the early twentieth century Cajal used Golgi's staining (silver staining) to get a full view of the synaptic connections within, into and outwards the hippocampus. Remarkably, he correctly pointed out directionality of information fluxes using just anatomical data (Newman et al., 2017).

We have two hippocampi, each one situated in a cerebral hemisphere. They are located deep in the human brain within the medial temporal lobe, under the allocortex. As information flow within different hippocampal subregions is largely unidirectional, it is relatively easy to visualize the whole flux. First, the hippocampus receives inputs from the pyramidal neurons of the entorhinal cortex through the perforant path, along the subiculum, which mainly synapse with neurons in the granular layer of the dentate gyrus (**Fig. 1**). The axons of these granular neurons, called mossy fibers, mainly pass on the information to CA3 pyramidal cells, where a second synapse occurs (Andersen et al., 2007). From there Schaffer collaterals (this is, axons from CA3) synapse with CA1 neurons, and then these ones project back to the entorhinal cortex, completing what is commonly known as trisynaptic loop (Andersen et al., 2007). Apart from the output to the entorhinal cortex, there are additional projections to other cortical areas, including the prefrontal cortex, which is commonly associated with decision making (Andersen et al., 2007). Of course, there are other inputs and outputs in the hippocampus, but the described above are the most relevant ones.

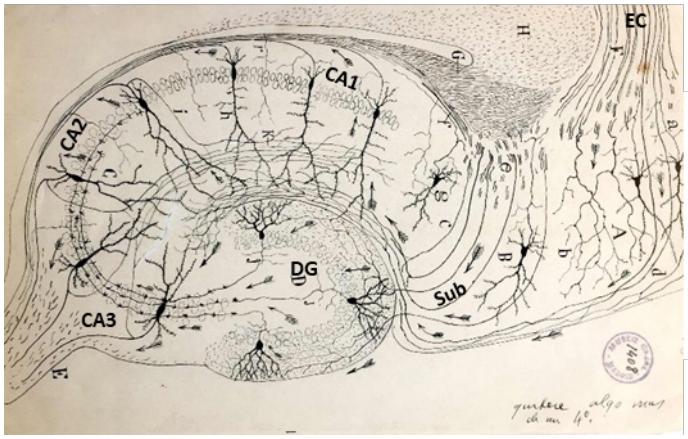

Figure 1. Santiago Ramón y Cajal's original drawing of the hippocampus of a small mammal. Cajal suggested the flow of information with arrows. DG: Dentate gyrus; Sub: subiculum; EC: Entorhinal cortex. Adapted from Newman et al., 2017.

From the functional side, the hippocampus is involved in many integrative functions regarding information processing. Moreover, it is essential for the creation of new memories concerning experienced events (episodic or autobiographical memory) and for spatial memory and navigation (Andersen et al., 2007).

1. 2 Implications of PDZ-domain driven interactions in synaptic plasticity: Long-term depression

In a chemical synapse, postsynaptic density (PSD) proteins receive, decode and propagate neuronal signals within the postsynaptic neuron as a membrane-associated superstructure specialized for postsynaptic signal transduction and processing. In fact, PSD comprises a high ordered network of dynamic components: Membrane receptors, cell-adhesion molecules, signaling proteins and their regulators, cytoskeleton proteins and cytoskeletal regulators, membrane trafficking proteins, molecular motors, scaffold proteins and protein-synthesis machinery (Feng & Zhang, 2009). One remarkable characteristic of PSD superstructure is that its complex architecture is mainly maintained by PDZ domains, which is one of the most abundant protein interaction modules in eukaryotic proteomes (Long et al., 2003). Precisely, PSD-95 is one of the most studied PDZ-domain-containing proteins and its modular PDZ structure consists of two tandem N-terminal PDZ domains connected by a short flexible linker and a third PDZ domain towards the middle of the sequence (Feng & Zhang, 2009).

Many PSD proteins are involved in synaptic plasticity, this is, activity-dependent selective strengthening or weakening of synapses over time. To note, according to the Hebbian postulate long-lasting synaptic plasticity underlies the processes of learning and memory trace maintenance (Hebb, 1949). In particular, long-term depression (LTD) is an activity-dependent reduction of the synaptic strength lasting hours or longer, and therefore is considered a Hebbian type of synaptic plasticity (Massey & Bashir, 2007). LTD can also be artificially induced by a long-lasting low frequency stimulation protocol (Jurado et al., 2010). In the hippocampus, LTD affects hippocampal CA1 pyramidal cells innerved by Schaffer collaterals, and is brought up by small, slow rises in postsynaptic calcium levels (Massey & Bashir, 2007).

At least three events are responsible for initiating diverse forms of synaptic plasticity, including LTD: Voltage-gated glutamate ionotropic receptor activation (NMDA receptor), postsynaptic calcium influx and activation of a serine/threonine phosphatase cascade (Massey & Bashir, 2007). Furthermore, during NMDA receptor-dependent LTD, a well-known lipid phosphatase, PTEN, is recruited and anchored by PSD-95 to the postsynaptic density in hippocampal dendritic spines (Jurado et al., 2010). This recruitment is based on a PDZ-dependent interaction, as it requires the C-terminal PDZ binding-motif of PTEN (Jurado et al., 2010), which interacts with the PDZ domains 1-2 of PSD-95 (Knafo et al., 2016). Enhancement of PTEN lipid phosphatase activity in

the postsynaptic terminal is able to drive depression of AMPA receptor-mediated synaptic responses, another class of glutamate ionotropic receptors (Jurado et al., 2010). This activity is specifically required for NMDA receptor-dependent LTD, and therefore contributes to synaptic plasticity.

1. 3 PTEN in Alzheimer's disease

Memory loss in Alzheimer's disease seems to begin with mild alterations to hippocampal synaptic efficacy prior to the loss of synapses and the neuronal degeneration that takes place. The amyloid-β (Aβ) interferes with the function of synapses that encode new memories, although the exact molecular mechanisms that link Aβ to memory loss have only recently started to emerge. Indeed, there is now evidence that Aβ, which itself induces synaptic depression (Cummings et al., 2015), triggers the access and accumulation of PTEN to the postsynaptic terminal in rodent hippocampal neurons (Knafo et al., 2016), therefore increasing LTD in the hippocampus. To prevent PTEN over-recruitment to the postsynaptic density, a synthetic myristoylated peptide corresponding to the PDZ binding-motif of PTEN (PTEN-PDZ peptide) was administered to mouse models of Alzheimer's disease, showing that it was able to rescue synaptic and cognitive function through reduction of LTD (Knafo et al., 2016). The proposed underlying mechanism is a competitive inhibition of the PDZ-domain-based interaction of PTEN with PSD-95, resulting in less PTEN being recruited to the synapse.

2 OBJECTIVES

Based on the hypothesis that PTEN dysregulation is implicated on the complex pathophysiology of Alzheimer's disease, this project has a dual objective. On the one hand, it aims to test through immunohistochemistry whether PTEN localization is dysregulated on post-mortem hippocampi from Alzheimer's disease patients, and whether asymptomatic and symptomatic patients present differences among them on this aspect. On the other hand, using machine learning approaches it seeks to develop a computational method for virtual screening of compounds able to restore imbalanced PTEN intracellular localization.

3 MATERIALS AND METHODS

3. 1 Human brain tissue

Brain tissue was obtained from the Institute of Neuropathology HUB-ICO-IDIBELL Biobank following the guidelines of Spanish legislation on this matter (Real Decreto de Biobancos 1716/2011) and approval of the local ethics committees. Research was conducted in compliance with the policies and principles contained in the European Policy for the Protection of Human Subjects. The pathological state and stage was defined according to a standardised protocol of the

CERAD (Consortium to Establish a Registry for Alzheimer's Disease) and the Braak and Braak criteria (Braak & Braak, 1991), and samples from patients were classified as controls (this is, showing no neuropathological findings or lesions), asymptomatic (Braak stage I/II) and symptomatic (Braak stage IV/V/VI) (**Table 1**).

Table 1. Human samples used in this study.

Group	Patient code	Gender	Age	PMI[1]	Diagnosis
Control	14/00003	Female	51	9h 35 min	No neuropathological findings
	11/00011	Male	50	17h 15 min	No neuropathological findings
	07/00084	Male	46	15h	No neuropathological findings
Asymptomatic	04/00045	Male	86	18h	Braak stage I/A
	07/00156	Male	71	5h 15 min	Braak stage II/A
	03/00163	Male	78	7h	Braak stage I/B
Symptomatic	12/00038	Male	79	4h 15 min	Braak stage IV/B
	03/00097	Male	81	3h	Braak stage IV/B
	06/00134	Female	80	2h 45 min	Braak stage IV/A

[1]PMI: Post-mortem interval. It is the time that has elapsed since the death of a patient and removal of the brain prior to cryopreservation.

3. 2 Tissue processing

Cryofixed human brain tissue (-80°C) was sliced by cryostat (Leica CM3050s, Leica Biosystems) at -20°C to obtain 20 µm coronal sections, which were subsequently placed in gelatinized frozen slides. Tissue was fixed in 4% paraformaldehyde for 10 minutes at 4°C.

3. 3 Immunohistochemistry

Fixed slices were permeabilized with 0.1% Triton X-100 for 30 minutes. Non-specific binding was avoided with 2-hour incubation at room temperature in blocking solution (5% horse serum, 0.1% Triton X-100 in PBS). Then double immunohistochemistry was performed sequentially with PSD-95 (NeuroMabs, 1:250 in blocking solution) and PTEN (Cell Signalling, 1:200 in blocking solution) primary monoclonal antibodies and their corresponding secondary antibodies: Donkey anti-mouse conjugated to Alexa Fluor 350 (Invitrogen, 1:1000 in blocking solution) and goat anti-rabbit conjugated to Alexa Fluor 594 (Life Technologies, 1:1000 in blocking solution), respectively. Sections were incubated overnight at 4°C with the primary antibody and the following day, after washing out with PBS, they were incubated for 2 hours at room temperature with the secondary antibody, washed out again with PBS and dipped in DAPI. After staining, the

sections were covered with ProLong Gold (Thermo Fisher Scientific) and coverslipped. All the samples were processed in parallel.

3. 4 Confocal microscopy and quantification

High magnification images of immunostained sections were obtained by laser-scanning confocal microscopy (Zeiss LSM 880, Carl Zeiss) in aleatory hippocampal regions using the same laser intensity for all the slices. Quantification was performed with Imaris software (version 7.2, Bitplane AG). Three channels were built for PSD-95, PTEN and DAPI. PSD-95 puncta single dots (PSD-95 puncta) were identified with the built-in spot detection algorithm in Imaris, allowing consideration of spots of different sizes and using local background subtraction. In order to follow a rigorous methodology for all samples and to avoid as much as possible artefacts in the quantification, the entire process was done stablishing the average diameter of the spot as 0.3 μm and using the same threshold levels for all the samples. The algorithm yielded the fluorescence intensity sum of PTEN in each spot, as well as individual spot sizes. Spots with extreme outlier values of size or fluorescence intensity sum of PTEN were discarded from the analysis. It is assumed that each PSD-95 puncta corresponds to a single synapse.

3. 5 Thioflavin-S staining

Adjacent slices to those immunostained were processed to visualize amyloid beta plaques and neurofibrillary tangles. For so, fixed slices were placed in 1% Thioflavin-S (Sigma-Aldrich) for 45 minutes, then differentiated in 70% ethanol for 5 minutes, rehydrated and dipped in DAPI. After staining, the sections were covered with ProLong Gold (Thermo Fisher Scientific) and coverslipped. Imaging was performed through structured illumination microscopy (Apotome 2, Carl Zeiss), which is a three-dimensional optical imaging technique that enables acquisition of high resolution sectioned images, much as a scanning confocal microscope does. After scanning stitching procedure was performed to ensure overlap of individual images. For high-magnification images the maximum intensity projection of a z-stack is shown.

3. 6 Computational models

The following generalized linear expression was derived for generating computational models suitable for virtual screening in drug discovery:

$$p(A_i|\vec{c}) = \phi_0 + \sum_{m=1}^{m\,max} \phi_m \cdot p(\mathrm{A}|c_m)\Big|_{\vec{c}} + \sum_{k=1,m=1}^{k,m\,max} \phi_{k,m} \cdot \left(D_k(i) - D_k(\mathrm{A}|c_m)\right)\Big|_{\vec{c}} \quad (1)$$

The outcome of the computational model is the probability that a certain compound i will be active (event A) under the selected values of the boundary conditions of the assay ($\vec{c}$). Such

conditions may be, for example, the type of target, or even the target itself. The different boundary conditions are denoted as c_m. The outcome value depends, on the one hand, on the *a priori* probability that an aleatorily chosen compound will be active under $\vec{c}$, denoted by $p(\mathrm{A}|c_m)$, and on the other hand, on the difference between the values of the compound's molecular descriptors, denoted by $D_k(i)$, and the expected values for the molecular descriptors of an aleatorily chosen compound under $\vec{c}$, denoted by $D_k(\mathrm{A}|c_m)$. A molecular descriptor is any variable that describes physicochemical or biological properties of compounds. Last, the phenomenological coefficients (ϕ_0, ϕ_m and $\phi_{k,m}$) are the parameters that will be determined by machine learning algorithms from a training dataset.

3. 7 Data mining and machine learning

Homologous proteins to the target (PSD-95) were selected with default-mode BLAST algorithm (Altschul, Gish, Miller, Myers, & Lipman, 1990) over the PubChem BioAssay database (Kim et al., 2016), using the sequence of the targeted domains (PDZ1 and PDZ2) as query. Then, data from previously assayed compounds against these selected proteins was obtained from the publicly available ChEMBL-EBI database (Gaulton et al., 2017). This initial dataset was manually curated, and all the different possible parameters for the computational model were calculated. Then compounds were classified as active or inactive within their assays' boundary conditions. The curated dataset was aleatorily divided into two mutually disjoint subsets, denoted as training and validation (in a 3:1 proportion, respectively). Machine learning was then performed over the training subset with Statistica software (version 5.0, Tibco) in the form of both Fisher's generalized linear discriminant analysis and artificial neural networks following the previous expression for computational models. Three built-in architectures of artificial neural networks in Statistica were used: Linear, multilayer perceptron and radial basis function neural networks. Machine learning algorithms select and weight a subset of variables from the problem in order to generate the different computational models, with the objective of fitting the input data to the outcome (similarly to a multivariate regression analysis). The obtained models were ranked according to their sensitivity and specificity when predicting the validation subset.

3. 8 Network analysis

Active compounds in real and predicted datasets and their associated boundary conditions were converted into directed graphs. A graph of "N" nodes can be mathematically formalised as an adjacency matrix (G):

$$G = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} = (a_{pq}) = \begin{cases} 1, & \exists e(N_p, N_q) \\ 0, & otherwise \end{cases} \quad (2)$$

Where $e(N_p, N_q)$ is a directed edge from the node N_p to the node N_q. Comparison of topological measures of graphs, this is, invariants of the adjacency matrix, is a way to quantitatively evaluate the quality of computational models. Values for selected boundary conditions were defined as nodes, and directed edges were stablished among them following this scheme: Organism to target, target to compound, compound to bioactivity, bioactivity to organism. Then two topological centrality measures were made using CentiBin software (version 1.4.3, Junker et al., 2006): Diameter and average distance (**Table 2**). Two other graphs with similar number of nodes were also stochastically generated with CentiBin software for comparative purposes: Erdős-Rényi (random) network and Eppstein (power law) network. For them the same parameters depicted above were also calculated. Graphs were represented using Cytoscape software (version 3.5.1, Shannon et al., 2003) in an edge-weighted spring embedded layout.

Table 2. Definition and calculation of network centrality measures.

Parameter	Formula	Definition
Diameter (D)	$G^{D+1} = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix}$; Where D$\in \mathbb{N}$ is the minimum value that satisfies this relation	The largest value among the minimum number of edges connecting any two nodes. In other words, the longest geodesic.
Average distance (μ)	$\mu(G) = \frac{1}{\binom{n}{2}} \cdot \sum_{p \neq q} d(N_p, N_q)$ Where d(N_p, N_q) is the shortest path-length between two nodes N_p and N_q (the geodesic).	The average minimum number of edges connecting any two nodes, or the average geodesic.

3. 9 Statistical analysis

Statistical significance was determined by one-way analysis of variance (ANOVA) followed by Tukey's multiple comparison post hoc test using GraphPad Prism (version 7.00, GraphPad Software). The null hypothesis was rejected at $p<0.05$.

4 RESULTS

4. 1 Thioflavin-S positive Aβ plaques

As it has been extensively reported (Casanova et al., 1993; Cole et al., 1991), Aβ plaques and neurofibrillary tangles were found in the hippocampus of Alzheimer's disease patients (**Fig. 2A-**

C). In particular, Thioflavin-S staining revealed multiple Aβ plaques in the CA1 and CA2 region around the dentate gyrus, but not in the dentate gyrus itself (**Fig. 2B**), and neurofibrillary tangles dispersedly distributed within the whole hippocampus, including the dentate gyrus (**Fig. 2C**). To note, stained Aβ plaques varied in size and fluorescence intensity.

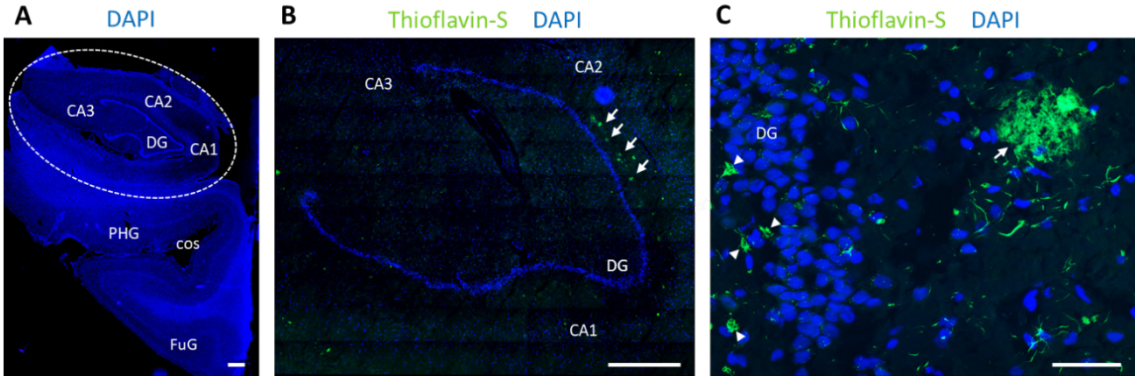

Figure 2. Alzheimer's disease correlates with Aβ plaques around the dentate gyrus. (A) Coronal section of a human brain showing the posterior hippocampus (dashed line) and part of the allocortex in the medial temporal lobe. DG, dentate gyrus; PHG, parahippocampal gyrus; cos, collateral sulcus; FuG, fusiform gyrus. (B) Aβ plaques surrounding the dentate gyrus of a human hippocampus (arrows). (C) Detail of an Aβ plaque at high magnification (arrow) and neurofibrillary tangles (arrowheads). Scale bars, 1 mm (A, B) and 50 µm (C). Structures and substructures were identified and located according to Paxino's human brain atlas (Mai, Voss, & Paxinos, 2015).

4. 2 Immunohistochemistry and confocal imaging

The following step was to test the hypothesis that PTEN might be over-recruited to synapses by PSD-95 in response to Aβ deposition during the avenue of Alzheimer's disease, and therefore contribute to its complex pathophysiology. Indeed, this interaction was previously stablished as a new pharmacological target in mouse models of the disease (Knafo et al., 2016). However, we must have into account that animal models may reproduce part of the symptomatology of a certain human disease, but they never represent the disease "as it is". Therefore, it is worth wondering whether this interaction would also be dysregulated in human patients of Alzheimer's disease. For that, we performed doubled immunohistochemistry in post-mortem human hippocampi against PTEN and PSD-95 (**Fig. 3A**) and quantified the fluorescence intensity of PTEN in the synaptic density, reflected by PSD-95 puncta (i. e. spots).

The number of PSD-95 puncta analysed was 16,904 for the control group, 23,852 for the asymptomatic group and 5,680 for the symptomatic group. Results confirmed that distributions of PTEN intensity sum in PSD-95 puncta were significantly different between the three groups (**Fig. 3B**, ANOVA, $p<0.001$). In particular, asymptomatic and symptomatic groups showed PTEN intensity distributions significantly different from the control group (**Fig. 3B**, Tukey's post hoc test, $p<0.001$ for both comparisons), but there were no significant differences between these two groups of Alzheimer's disease patients (**Fig. 3B**, Tukey's post hoc test, $p=0.2085$). As it can be

inferred from the cumulative frequency distribution data, Alzheimer's disease patients showed puncta that tended to have higher PTEN intensity values than the control group. Also, PSD-95 puncta significantly varied in size among the three groups (**Fig. 3C**, ANOVA, $p<0.001$). Interestingly, whereas asymptomatic ill patients showed more small-size puncta than controls (**Fig. 3C**, Tukey's post hoc test, $p<0.001$), the opposite was found for symptomatic patients (**Fig. 3C**, Tukey's post hoc test, $p<0.001$).

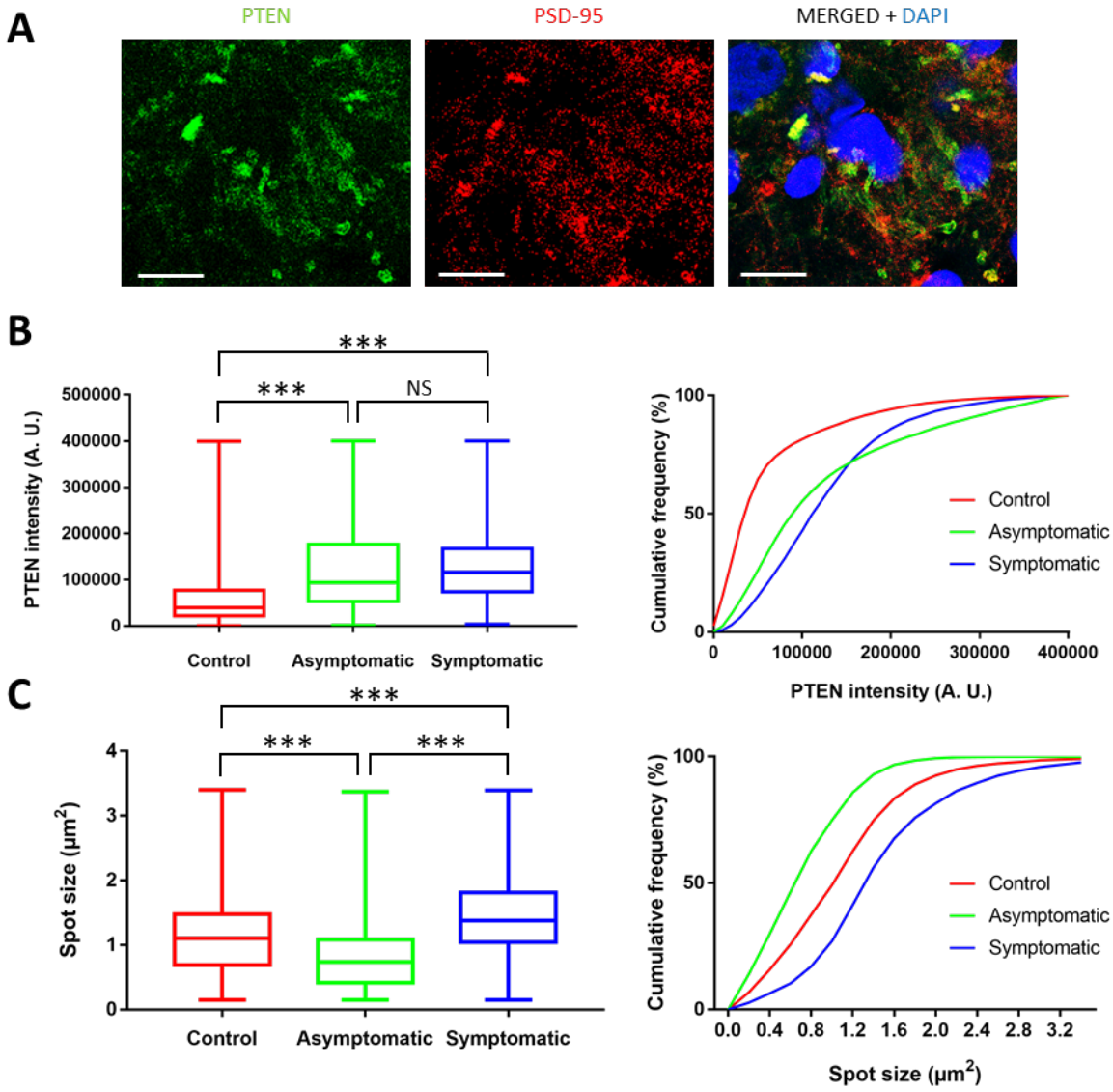

Figure 3. PTEN levels in PSD-95 puncta are altered in Alzheimer's disease. (A) Representative confocal microscopy image of an aleatory region in the hippocampus of a patient. (B) Quantification of fluorescence intensity sum of PTEN in PSD-95 puncta. (C) Quantification of PSD-95 puncta sizes. Scale bars, 10 µm. Distributions are represented as Box &Whiskers plots (left) and cumulative frequency plots (right). NS: not significant; (***$p<0.001$).

4. 3 Development and validation of a computational model

Next, based on current data of compounds assayed against proteins with PDZ domains, the aim was to develop a computational model to *in silico* screen new compounds able to disrupt the PDZ-

mediated interaction between PTEN and PSD-95. As it has been reported that different classes of PDZ domains recognise different C-terminal motifs (Songyang et al., 1997), BLAST bioinformatic tool for multiple sequence alignments (Altschul et al., 1990) was used to select a set of proteins with domains homologous to PDZ1 or PDZ2 of PSD-95. The underlying axiom was that homologous domains from different proteins would be targeted by compounds with similar physicochemical properties. Then a dataset was built from bioactivity data of compounds already tested against these proteins (i. e. 2,721 compounds comprising 4,356 assays). After classifying the compounds as active or inactive within their assays' boundary conditions, the network was navigated through Fisher's generalized linear discriminant analysis (LDA) following the classification computational model described in the materials and methods section, in which the output is the probability that a given compound will be active under certain values of the boundary conditions. The LDA model was able to correctly classify as active or inactive compounds from the validation subset with a sensitivity of 86.15% and specificity of 81.54% and included several boundary conditions for the assay (**Supplementary Table 1**). In order to improve the model, it was also tried to fit the dataset with different architectures of artificial neural networks, but there was not a big improvement in the prediction of the validation subset (**Fig. 4A**).

Then, using active compounds and their associated boundary conditions, directed networks were built for the real and predicted datasets, and also two stochastically generated networks were added to the analysis as controls (**Fig. 4B**). Comparison of their topologies through centrality measures revealed similar diameter and average distance in real and predicted networks, although the prediction overestimated the total number of nodes and edges (**Table 3**). To note, diameter and average distance differed from the ones of a randomly connected Erdős-Rényi network but were close to those from a power law distribution Eppstein network.

Table 3. Topological centrality measures of generated networks.

Network	N	L	Diameter	μ
Real	845	1784	10	4.088
Predicted	1279	2608	9	4.034
Eppstein (power law)	832	1782	11	4.8123
Erdős-Rényi (random)	820	1862	25	10.400

N: total number of nodes; L: total number of edges; μ: average distance.

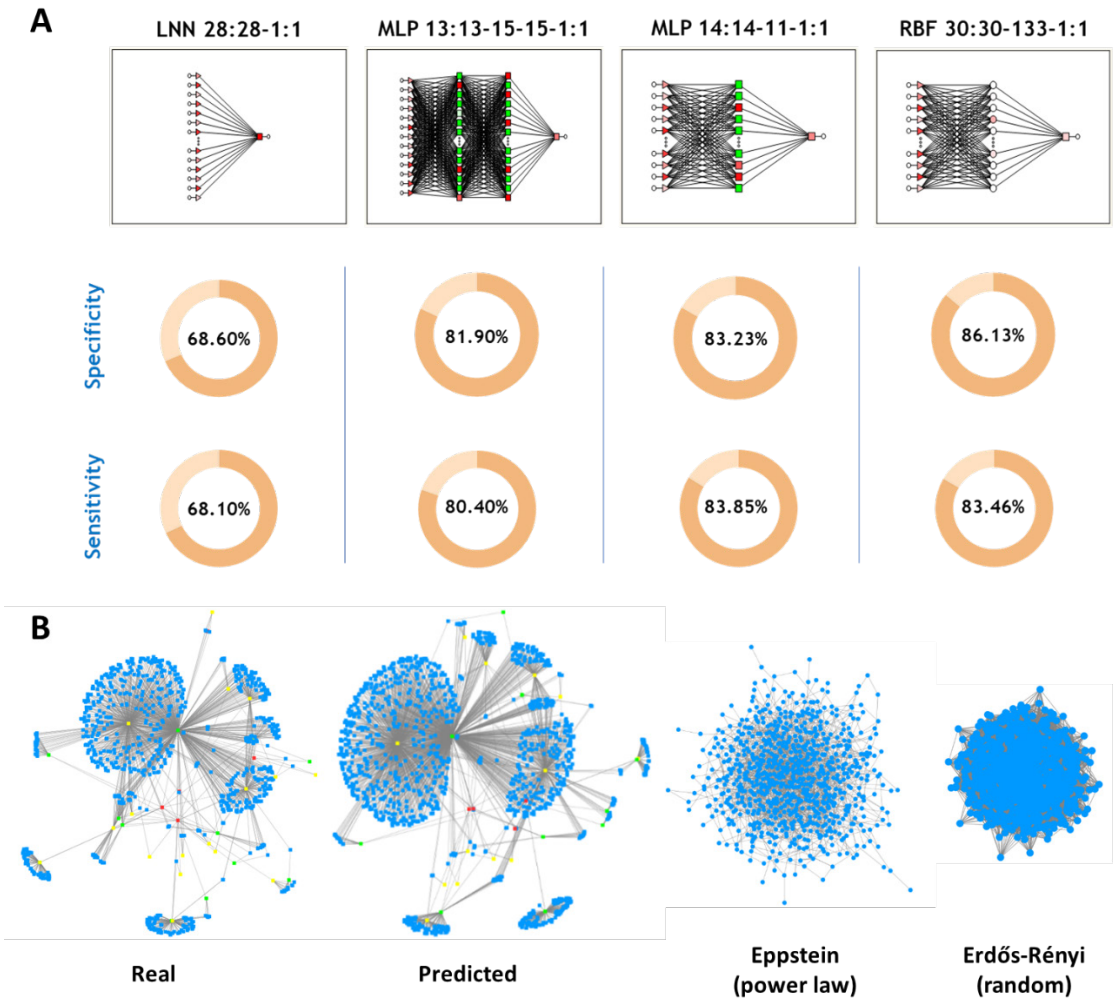

Figure 4. Development of a computational model able to predict drugs targeting PDZ scaffolding domains. (A) Different architectures of artificial neural networks used for data mining. LNN, linear neural network; MLP, multilayer perceptron; RBF, radial basis function. Shown specificity and sensitivity values correspond to the validation subset. (B) Edge-weighted spring embedded layout of generated networks. Blue: compound; green: target; yellow: bioactivity; green: organism.

5 DISCUSSION

Quantitative analysis of the immunohistochemistry experiment revealed that hippocampal postsynaptic densities of Alzheimer's disease patients tended to accumulate higher levels of PTEN with respect to healthy individuals. Indeed, PTEN was evenly distributed in both symptomatic or asymptomatic patients, suggesting that this is a common molecular signature along the avenue and evolution of the disease. This data fits with previous studies in animal models of Alzheimer's disease, where it was found that PTEN promotes LTD in response to Aβ deposition (Knafo et al., 2016). Even though in human brain samples it is not possible to directly test LTD protocols, we can hypothesize that a similar synaptic rearrangement is happening in view of the conserved molecular signature and the finding of Aβ plaques in the hippocampus. Also, these plaques appeared mainly (but not only) in the CA1 and CA2 regions around the

dentate gyrus. As Schaffer collaterals project from CA3 to CA1 crossing those precise regions (Andersen et al., 2007) and are involved in LTD induction in CA1 (Massey & Bashir, 2007), synaptic plasticity in CA1 pyramidal cells might be somehow affected by such plaques. Indeed, accumulating evidence indicates that soluble Aβ assemblies directly alter synaptic plasticity mechanisms by facilitating LTD in hippocampal neurons (Hsieh et al., 2006; Li et al., 2009).

The fact that asymptomatic patients showed more small-size puncta than healthy controls, whereas the symptomatic group showed less, might seem at first contradictory. However, two possible explanations can be proposed for this result. First, as asymptomatic patients show no pathological symptoms of the disease, but they do show Aβ plaques and neurofibrillary tangles, one might wonder whether there might be an adaptative mechanism to tackle synaptic depression induced by Aβ in these first stages of the disease. One of such possible mechanisms would be to enlarger synapses in order to retain activity over the threshold level. This would result in postsynaptic densities turning larger. Then, as the disease advances, this mechanism would get overflowed, owing to a reduction in the size of the synapse, as the one observed for symptomatic patients. The second hypothesis deals with the fact that this result might be an artefact caused by the methodology followed for quantification.

In this work we have also reported the first LDA computational model able to find compounds against PDZ motifs under different boundary conditions. This computational model could be used to discover new compounds with the potential of disrupting PTEN interaction with PSD-95, thus imitating the mechanism of the PTEN-PDZ peptide. As this peptide was able to rescue synaptic and cognitive function in mouse models of Alzheimer's disease (Knafo et al., 2016), it is expected that compounds imitating its mechanism should be strong candidates for the treatment of Alzheimer's disease. In addition recent research has linked deregulation of PTEN function to autism spectrum disorders (Lugo et al., 2014; Tilot et al., 2015), and so such compounds may also be beneficial for the treatment of this disease.

With respect to the generated model itself, similar intra-model testing statistics can be found in other classification computational models reported in the literature for a variety of pharmacological and biological problems (Jamal & Scaria, 2013; Tyagi et al., 2013), suggesting that the original dataset was well fitted. Moreover, conservation of network topology indicates that the computational model is able to fairly recreate the original dataset including the boundary conditions, and thus that it may be used for screening novel compounds with molecular signatures similar to those of the original dataset. There are several reports that suggest that biological networks tend to follow a power law distribution (Barabási & Albert, 1999), and interestingly, centrality measures of the original dataset and the prediction were quite close to those obtained with a power law distributed network. However, they were very different to those from a random

network, which suggests that real and predicted networks are not odd but contain higher complexity, and therefore contain more information.

6 CONCLUDING REMARKS

- According to immunohistochemical data, PTEN accumulated within hippocampal PSD in patients of Alzheimer's disease, suggesting that compounds based on the PTEN-PDZ peptide may be used for treating the disease.
- In this work a new LDA computational model was generated suitable for discovering new compounds targeting PDZ domains under several boundary conditions, which could be used to find compounds able to inhibit the PDZ-domain-mediated interaction of PTEN with PSD-95. Herby we provide a new opportunity for drug discovery in Alzheimer's disease based on the rescue of imbalanced hippocampal LTD.

7 ACKNOWLEDGEMENTS

First, I would like to thank the members of the molecular cognition laboratory for sharing their thoughts and for their inestimable help throughout the project, and in particular I thank Shira Knafo (MD, PhD) and Miguel Morales (PhD) for their guidance. Also, I must express my gratitude to Humberto González-Díaz (PhD) for introducing me to various computational tools, including machine learning algorithms. Finally, I express my sincere thanks to Isidro Ferrer (MD, PhD) from the University of Barcelona for providing the human brain tissue and to the research group led by Carlos Matute (MD, PhD) for kindly letting me use their cryostat.

8 REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

Andersen, P., Morris, R., Amaral, D., Bliss, T., & O'Keefe, J. (2007). *The Hippocampus Book*. New York: Oxford University Press.

Barabási, A.-L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, *286*(5439), 509–512.

Braak, H., & Braak, E. (1991). Neuropathological staging of Alzheimer-related changes. *Acta Neuropathologica*, *82*, 239–259.

Casanova, M. F., Carosella, N. W., Gold, J. M., Kleinman, J. E., Weinberger, D. R., & Powers, R. E. (1993). A topographical study of senile plaques and neurofibrillary tangles in the hippocampi of patients with Alzheimer's disease and cognitively impaired patients with

schizophrenia. *Psychiatry Research*, *49*(1), 41–62.

Cole, G. M., Masliah, E., Shelton, E. R., Chan, H. W., Terry, R. D., & Saitoh, T. (1991). Accumulation of amyloid precursor fragment in Alzheimer plaques. *Neurobiology of Aging*, *12*(2), 85–91.

Cummings, D. M., Liu, W., Portelius, E., Bayram, S., Yasvoina, M., Ho, S.-H., … Edwards, F. A. (2015). First effects of rising amyloid-β in transgenic mouse brain: synaptic transmission and gene expression. *Brain*, *138*(7), 1992–2004.

Feng, W., & Zhang, M. (2009). Organization and dynamics of PDZ-domain-related supramodules in the postsynaptic density. *Nature Reviews Neuroscience*, *10*(2), 87.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., … Leach, A. R. (2017). The ChEMBL database in 2017. *Nucleic Acids Research*, *45*(D1), D945–D954.

Hebb, D. (1949). *The Organization of Behavior*. New York: Wiley & Sons.

Hsieh, H., Boehm, J., Sato, C., Iwatsubo, T., Tomita, T., Sisodia, S., & Malinow, R. (2006). AMPAR Removal Underlies Aβ-Induced Synaptic Depression and Dendritic Spine Loss. *Neuron*, *52*(5), 831–843.

Jamal, S., & Scaria, V. (2013). Cheminformatic models based on machine learning for pyruvate kinase inhibitors of Leishmania mexicana. *BMC Bioinformatics*, *14*, 329.

Junker, B. H., Koschützki, D., & Schreiber, F. (2006). Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics*, *7*, 219.

Jurado, S., Benoist, M., Lario, A., Knafo, S., Petrok, C. N., & Esteban, J. A. (2010). PTEN is recruited to the postsynaptic terminal for NMDA receptor-dependent long-term depression. *The EMBO Journal*, *29*(16), 2827–2840.

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., … Bryant, S. H. (2016). PubChem Substance and Compound databases. *Nucleic Acids Research*, *44*(D1), D1202–D1213.

Knafo, S., Sánchez-Puelles, C., Palomer, E., Delgado, I., Draffin, J. E., Mingo, J., … Esteban, J. A. (2016). PTEN recruitment controls synaptic and cognitive function in Alzheimer's models. *Nature Neuroscience*, *19*(3), 443–453.

Li, S., Hong, S., Shepardson, N. E., Walsh, D. M., Shankar, G. M., & Selkoe, D. (2009). Soluble Oligomers of Amyloid-β; Protein Facilitate Hippocampal Long-Term Depression by Disrupting Neuronal Glutamate Uptake. *Neuron*, *62*(6), 788–801.

Long, J.-F., Tochio, H., Wang, P., Fan, J.-S., Sala, C., Niethammer, M., … Zhang, M. (2003). Supramodular structure and synergistic target binding of the N-terminal tandem PDZ

domains of PSD-95. *Journal of Molecular Biology*, *327*(1), 203–214.

Lugo, J. N., Smith, G. D., Arbuckle, E. P., White, J., Holley, A. J., Floruta, C. M., … Okonkwo, O. (2014). Deletion of PTEN produces autism-like behavioral deficits and alterations in synaptic proteins. *Frontiers in Molecular Neuroscience*.

Mai, J. K., Voss, T., & Paxinos, G. (2015). *Atlas of the human brain* (4th ed.). Amsterdam: Elsevier & Academic Press.

Massey, P. V, & Bashir, Z. I. (2007). Long-term depression: multiple forms and implications for brain function. *Trends in Neurosciences*, *30*(4), 176–184.

Newman, E. A., Araque, A., Dubinsky, J. M., Swanson, L. W., King, L., & Himmel, E. (2017). *The beautiful brain. The drawings of Santiago Ramón y Cajal*. New York: Abrams.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, *13*(11), 2498–2504.

Songyang, Z., Fanning, A. S., Fu, C., Xu, J., Marfatia, S. M., Chishti, A. H., … Cantley, L. C. (1997). Recognition of Unique Carboxyl-Terminal Motifs by Distinct PDZ Domains. *Science*, *275*(5296), 73–77.

Tilot, A. K., Bebek, G., Niazi, F., Altemus, J. B., Romigh, T., Frazier, T. W., & Eng, C. (2015). Neural transcriptome of constitutional Pten dysfunction in mice and its relevance to human idiopathic autism spectrum disorder. *Molecular Psychiatry*, *21*(1), 118–125.

Tyagi, A., Kapoor, P., Kumar, R., Chaudhary, K., Gautam, A., & Raghava, G. P. S. (2013). In Silico Models for Designing and Discovering Novel Anticancer Peptides. *Scientific Reports*, *3*, 2984.

9 APPENDIX I: SUPPLEMENTARY DATA

Supplementary table 1. Parameters of the LDA computational model.

	$D_k(\mathrm{A}\|c_m)$			
	ALogP	PSA	$\phi_{k,m}$[1]	$p(\mathrm{A}\|c_m)$
Organism			0.023	
Homo sapiens	-	84.193		-
Rattus norvegicus	-	82.622		-
Mus musculus	-	53.458		-
Bos taurus	-	188.501		-
Other	-	97.162		-
Target			0.036	
Nitric-oxide synthase, brain	-	82.823		-
Segment polarity protein dishevelled homolog DVL-1	-	83.815		-
Peripheral plasma membrane protein CASK	-	94.459		-
Protein-tyrosine phosphatase 1E	-	104.851		-
Proteinase activated receptor 4	-	63.698		-
Voltage-gated N-type calcium channel alpha-1B subunit/Amyloid beta A4 precursor protein-binding family A member 1	-	98.479		-
Glutamate receptor-interacting protein 1	-	142.770		-
Disks large homolog 4	-	205.514		-
Membrane-associated guanylate kinase-related 3	-	299.708		-
Regulator of G-protein signaling 12	-	84.950		-
Segment polarity protein dishevelled homolog DVL-3	-	83.734		-
Nitric oxide sythases; iNOS and nNOS	-	84.204		-
Nitric-oxide synthase (endothelial and brain)	-	80.199		-
Other	-	186.107		-
Bioactivity			10.832[2]	
IC50 (nM)	-	-		0.000
Ki (nM)	-	-		0.603
Inhibition (%)	-	-		0.418
Selectivity	-	-		0.140
EC50 (nM)	-	-		0.791
Activity (%)	-	-		0.406
Potency (nM)	-	-		0.516
Kd (nM)	-	-		0.897
Selectivity index	-	-		0.259
nNOS activity (%)	-	-		0.306
% max (%)	-	-		0.364
NO formation (%)	-	-		0.333
Selectivity ratio	-	-		0.083
NOHA (%)	-	-		0.444
Activity (pmol/min)	-	-		0.625
Inhibition (uM)	-	-		0.571
Km (nM)	-	-		0.800
Kcat/Km (/s/mM)	-	-		0.333
Kcat (/min)	-	-		0.500
Target type			-0.332	
Single protein	1.819	-		-

Protein-protein interaction	3.212	-		-
ADMET	-5.454	-		-
Selectivity group	1.405	-		-
Other	-0.163	-		-
Confidence score			0.484	
0	0.986	-		-
1	2.237	-		-
4	2.205	-		-
5	-1.389	-		-
8	3.089	-		-
9	1.091	-		-
Assay description			-0.043	
Scientific Literature	-	90.351		-
PubChem BioAssay	-	97.240		-
BindingDB Database	-	86.852		-
DrugMatrix	-	38.330		-

[1]The intercept of the LDA model is ϕ_0= -5.040. [2] This value refers to ϕ_m instead of $\phi_{k,m}$.