Comments on behavioral data published by the Pratico lab

This is an independent assessment on behavioral data reported in a number of publications from Dr. Domenico Pratico's lab at Temple University. Although our limited analysis is by no means exhaustive, we feel strongly that the analysis in this document should raise enough concerns to warrant further investigation.

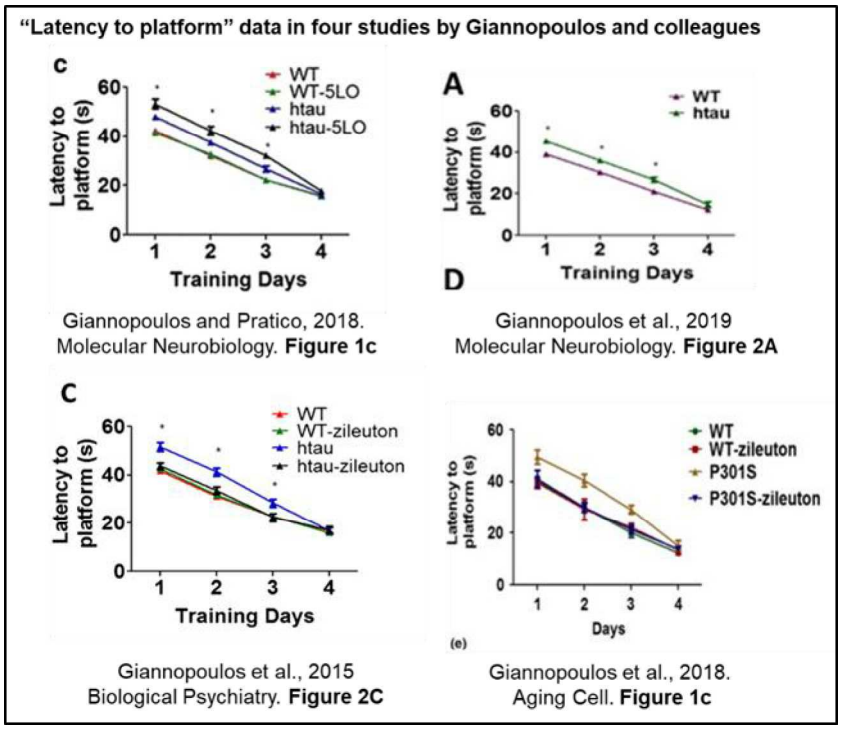
This assessment is provided by three professional behavioral neuroscientists [redacted] who have no personal or professional conflicts with Dr. Pratico. Signed statement echoing similar concerns from Dr. Richard Morris, who invented the Morris water maze in 1980s, is attached herein.

As will be expounded below, we strongly recommend that: 1) the original water maze and Y data (e.g. ANY-maze tracking raw data, as well as formatted worksheets used for data analysis) be reviewed again by experts in behavioral and experts in data science, 2) full statistical analysis on behavioral data (heretofore inadequately reported) be thoroughly reviewed against the raw data, 3) the authors should clarify the mouse strain background (inbred strain names and crosses, including generation if possible) used in all experiments for transgenic mice and controls, especially for their htau and P301s mouse models.

NOTE: Quoted text are italicized and highlighted in purple below.

Unrealistically small error bars

Morris water maze is a stressful test (Harrison et al., 2009), especially on day 1, when mice tend to react to being placed in the water very differently. This is why the variability of "latency to reach the platform" is usually large on day 1. As training proceeds, variability tends to decrease, but seldom disappears. As Dr. Richard Morris, who invented the Morris Water Maze in 1980s (Morris, 1984), stated in the accompanying letter *"I know of no previous study in the watermaze in which this has been observed - there is always variability on day 1 of spatial training even when, as in this case, preceded by visible platform training."*



The mouse strain background used for transgenic and for genetic control mice was inadequately noted in or missing from these papers. The statements such as these "*The animals were*

backcrossed ten times on the same genetic background of C57BL6/SJL. The wild-type (WT) mice are aged-matched C57BL6/SJL controls." (Giannopoulos et al., 2019). *"The animals were backcrossed 10 times on the same genetic background of C57BL6/SJL. The wild-type (WT) mice were aged-matched C57BL6/SJL control mice."* (Giannopoulos et al., 2015) inadequately described the mouse strain background used for transgenic and for genetic control mice. From these statements, it is not clear if the authors were creating a congenic strain, or if they meant "The animals were backcrossed 10 times to (C57BL/6 x SJL) F1 hybrid mice and maintained this way". This missing information essential for estimating % of mice that may be blind, and also would provide assurance that the genetic background was as evenly mixed as possible, albeit still segregating. In general, the use of genetically complex background and the use of both sexes (Giannopoulos et al., 2018, 2019; Giannopoulos et al., 2015; Giannopoulos and Pratico, 2018) are common sources of variability in the water maze and certain other tests. Furthermore, strain backgrounds that segregate certain known genetic variants inherited from parental strains – such as *Pde6b*^{rd1} the recessive blindness gene present in many common strains including C3H and SJL (Brown and Wong, 2007) and which may have been inherited in the subject mice here – are particularly problematic unless subject mice had been fully backcrossed to a sighted inbred strain or true F_1 hybrids were used. Without genotyping for this mutation – not noted in the paper – and especially with random breeding it is impossible to know whether some of the test mice were blind. If so, there is almost nil chance there would have been such low variability as shown in the water maze "latency to platform" results show in Figure 2C in (Giannopoulos et al., 2015), Figure 1c in (Giannopoulos et al., 2018), Figure 1c in (Giannopoulos and Pratico, 2018), and Figure 2A in (Giannopoulos et al., 2019). This is notwithstanding the likelihood that more variability would have been expected even if the test mice were all homozygous wildtype for *Pde6b*^{rd1}.

Inadequate statistical analysis and reporting

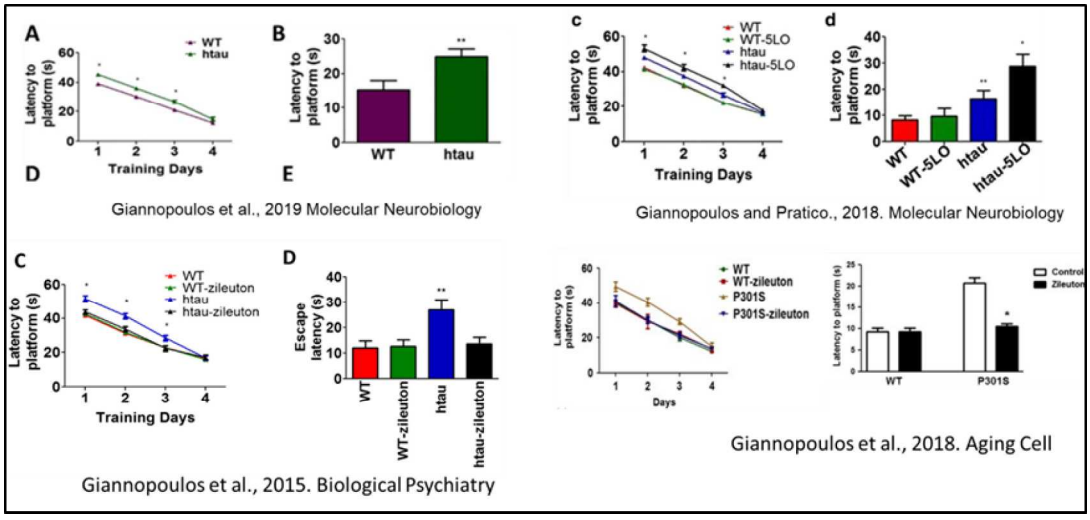
The unusual small variability noted above calls for a critical review of effect size. Unfortunately, F values are either entirely missing or inadequately reported in the papers that we have reviewed. In (Giannopoulos et al., 2018, 2019; Giannopoulos et al., 2015; Giannopoulos and Pratico, 2018; Vagnozzi et al., 2019), it was stated that one-way ANOVA and t-test were the primary tests for statistical analysis. Take (Vagnozzi et al., 2019) for example, it is unclear how the authors analyzed four days of water maze acquisition data (repeated testing) of two genotypes and two treatments with one-way ANOVA or t-test to achieve the result of *"During the training phase over 4 consecutive days, we observed no differences in performance among the different groups (Fig. 6e)"* (No F values were given). It is also unclear why in Figure 6 caption for f-h, one F value was given, followed by three p values. *"f–h During the probe trial, the following paradigms were measured: number of platform crosses for each group (F (3,28) =4.42 *p=0.0430, **p=0.0058, ****p<0.0001), latency to platform, and time spent in platform zone for the four groups. Values are expressed as mean±SEM*". Unlike this Vagnozzi paper, F values were entirely missing from the Giannopoulus et al papers (Giannopoulos et al., 2018, 2019; Giannopoulos et al., 2015; Giannopoulos and Pratico, 2018), all of which involve testing two between-subject factors (2 genotypes, 2 treatments) and four days of repeated testing in the water maze. Since these papers did not state what the p values are ---- from what statistical test and for which comparison---- it is puzzling what the asterisks mean in the water maze "latency to platform" results show in Figure 2C in (Giannopoulos et al., 2015), Figure 1c in (Giannopoulos et al., 2018), Figure 1c in (Giannopoulos and Pratico, 2018), and Figure 2A in (Giannopoulos et al., 2019). Take

(Giannopoulos et al., 2018) as another example, the statement in Results text "*In the training phase, although the P301S mice took more time to reach the platform by day four, all of them reached the training criterion (Figure 1c). However, compared with the treated group, P301S mice receiving vehicle performed significantly worse."* was not substantiated by F values or other formal description of statistics anywhere. It is unclear how "more time", "reached criterion", and "worse" were determined statistically.

Unusually straight learning curves and overlapping lines.

Notably, almost perfect linear learning progress (straight lines) were shown in Figure 2C in (Giannopoulos et al., 2015), Figure 1c in (Giannopoulos et al., 2018), Figure 1c in (Giannopoulos and Pratico, 2018), and Figure 2A in (Giannopoulos et al., 2019). Mice, including different inbred strains, differ in "search strategy" in the water maze (Brody and Holtzman, 2006; Nicolle et al., 2003; Rogers et al., 2017; Stavnezer et al., 2002). Different search strategies have been identified in non-AD mice such as C57BL/6J (Stavnezer et al., 2002), as well as in AD models such as PDAPP and APP mice (Brody and Holtzman, 2006; Janus, 2004). Differences in search strategy, response to water stress, and physical characteristics (such as seen different sexes) contribute to the commonly observed non-straight learning curves in virtually all water maze experiments that we have supervised in our labs. Dr. Richard Morris echoed a similar concern in his accompanying letter *"The strikingly straight learning curves are very unusual"*. Due to commonly recognized differences in learning and search strategies and other environmental factors and individual differences (sex, complex background, physically ability, vision etc.), even without considering treatment effects, it is extremely rare for two groups of mice to have identical "latency to platform" data throughout all four days of acquisition training, as shown in Figure 2C in (Giannopoulos et al., 2015) and Figure 1c in (Giannopoulos and Pratico, 2018).

Questionable probe trial latency data



As described by the authors, in the water maze test, mice go through 4 days of training with a hidden platform in the same location. 24 h after the last training trial, mice are tested in the "Probe trial" to see their behaviors in the absence of the previous learned platform. The latency-to-platform-region in the probe trial is expected to be close to the last acquisition latency data. In the four studies by Giannopoulos et al., all groups had almost the exact latency to platform data on day 4, with very

small error bars, indicating minimum within-group and between-group variabilities (Figure 2C in (Giannopoulos et al., 2015), Figure 1c in (Giannopoulos et al., 2018), Figure 1c in (Giannopoulos and Pratico, 2018), and Figure 2A in (Giannopoulos et al., 2019). Based on the near-perfect progression (in all groups) from day 1 to day 4, one would expect that all groups would continue to improve, and continue to have minimum within-group and between-group variabilities on day 5 (probe trial). However, the probe trial data ((Figure 2D in (Giannopoulos et al., 2015), Figure 1d in (Giannopoulos et al., 2018), Figure 1f in (Giannopoulos and Pratico, 2018), and Figure 2B in (Giannopoulos et al., 2019).) revealed substantial between-group differences. While it is possible that the first of four trials of each day is slower and more variable --- but in that case one would expect to see notable variabilities in the acquisition trials, and not just in the probe trial. To put it simply, from the straight learning curves, minuscule error bars, and identical day 4 data in the acquisition data (line graphs), the probe trial latency data are highly unexpected. Dr. Richard Morris commented on this unusual pattern repeatedly in his accompanying letter.

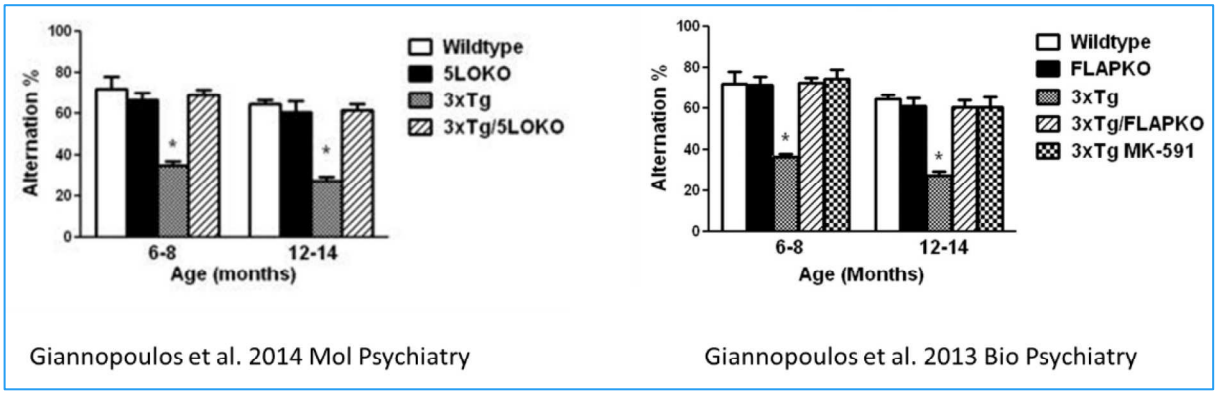
Almost identical wildtype data in different studies

The wildtype latency to platform data seem remarkably similar in Figure 2 in (Giannopoulos et al., 2015), Figure 1 in (Giannopoulos and Pratico, 2018), and Figure 1 in (Giannopoulos et al., 2018). This is notable, given that 10 months old mice were said to be the subjects in (Giannopoulos et al., 2018; Giannopoulos et al., 2015), whereas 11-12 months old mice were said to be the subjects in (Giannopoulos and Pratico, 2018). A similar concern is raised by Dr. Richard Morris in his letter attached herein.

Y Maze Spontaneous Alternation data

Duplicated data in different papers

Figure 1b (Giannopoulos et al., 2014) and figure 1B (Giannopoulos et al., 2013) appear to contain the same data: data of 6-8 month old and12-14 month old WT and 3xTg appear identical in the two papers.



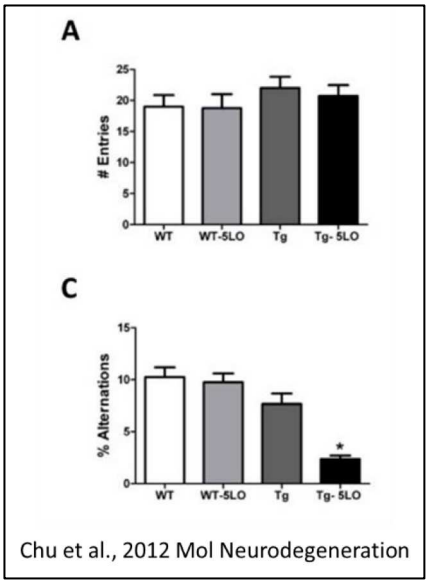
Giannopoulos et al. 2014 Mol Psychiatry

Giannopoulos et al. 2013 Bio Psychiatry

Unusually large inconsistency in Y maze data across studies of the 3xTg mouse model

We noticed an unusually large inconsistency in Y maze spontaneous alternation data from the Pratico lab. Given that results of this test, together with water maze data, are often the key behavioral readouts in their studies, the accuracy of the data is obvious. In (Li et al., 2019; Li et al., 2018), the Y maze spontaneous alternation % data for WT controls are quite similar (~60-70%), but the 3xTg data are about 30% in the 2019 paper (Figure 1B) and ~55% in the 2018 paper (Figure 1B) ---- a stunning difference. This within-model difference became more salient in a recent paper from the same authors, published in 2020. In (Li et al., 2020), the WT control average is at a surprisingly low level of 30% (more poorly than 3xTG mice in the aforementioned 2018 paper) and the 3xTg's average is only 10% (Figure 1B). This within-model variability is puzzling, given that the studies have the same first author, are about the same 3xTg mouse model tested at the same age (11-13 months) and tested following the same methods. Curiously, the error bars in each study seem to be in the normal range for this test, indicating normal variability in each study. This kind of large discrepancy can also be found in another papers on the 3xTg models from the lab. In (Chu et al., 2015) the 3xTg Y maze alternation average is only ~10% (Figure 1a), substantially lower than the average of ~40% in older 3xTg mice in an earlier paper with the same first author (Chu et al., 2012a) (Figure 1B). A review of other published literature on the spontaneous alternation behavior of the 3xTg mouse strain relative to its WT control including initial characterization of the strain by the La Ferla lab who developed the 3xTg line do not report the level of alternation deficits as described by the Pratico lab. In fact, many labs report no deficits or greater % alternation relative to controls (Stover et al., 2015). The large degree of inconsistency in Y maze alternation data reported by the Pratico lab seems highly unusual and poses questions on outcome reliability.

Curiously, seemingly impeccably consistent Y maze data were reported from the same lab, also in the past few years. As shown in Figure 2B of (Giannopoulos et al., 2015), Figure 1B of (Giannopoulos and Pratico, 2018) and Figure 1B of (Giannopoulos et al., 2019), WT mice on C57/SJL background were found to have a strikingly consistent average of 80% alternation score with very small error bars, despite the use of both sexes and the aforementioned concern on sight in these animals. Given that Li, Chu, and Giannopoulos are often co-authors, it would be surprising if they had used drastically different Y maze procedures, and indeed their first-authored papers state the same Y maze procedure. As such, original tracking data from ANY-maze, as well as the calculation results from each study, should be reviewed.



Chu et al., 2012 Mol Neurodegeneration

Mathematically unsound data The necessity for reviewing raw data is further highlighted by (Chu et al., 2012b). In Figure 1 of this paper (right), an extremely low average of ~2.5% alternation score was reported for the Tg-5LO group, with a very small error bar. With an average # entries of ~ 20 (Figure 1A), the average number of alternation is therefore less than 1, which means some mice probably had zero alternation and some had 1. At 1 alternation out of 20 entries, the percentage is 5.5%. With the score of being either 5.5% or zero, to get the group average of ~2.5% shown in Figure C, by estimation, half the mice would have a 5.5% score and the half would have zero alternation. With N=10 (as stated in the paper), this would

yield SEM=0.901 which will appear much larger than the error bar of the Tg-5LO group in Figure 1C. It is also salient that the WT alternation score in this paper is an abysmally low 10%.

In sum, the concerns that we laid out above are in such nature that we feel a strong sense of obligation, as professional behavioral scientists, to call for a thorough review of raw data, original tracking data, and statistical analysis. Note that the authors are bound by data sharing plans of their NIH grants.

Please feel free to contact us if there is anything we can be helpful with.

Sincerely,

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

References:

Brody, D.L., and Holtzman, D.M. (2006). Morris water maze search strategy analysis in PDAPP mice before and after experimental traumatic brain injury. Exp Neurol *197*, 330-340.

Brown, R.E., and Wong, A.A. (2007). The influence of visual ability on learning and memory performance in 13 strains of mice. Learn Mem *14*, 134-144.

Chu, J., Giannopoulos, P.F., Ceballos-Diaz, C., Golde, T.E., and Pratico, D. (2012a). 5-Lipoxygenase gene transfer worsens memory, amyloid, and tau brain pathologies in a mouse model of Alzheimer disease. Ann Neurol *72*, 442-454.

Chu, J., Giannopoulos, P.F., Ceballos-Diaz, C., Golde, T.E., and Pratico, D. (2012b). Adeno-associated virus-mediated brain delivery of 5-lipoxygenase modulates the AD-like phenotype of APP mice. Mol Neurodegener *7*, 1.

Chu, J., Li, J.G., Giannopoulos, P.F., Blass, B.E., Childers, W., Abou-Gharbia, M., and Pratico, D. (2015). Pharmacologic blockade of 12/15-lipoxygenase ameliorates memory deficits, Abeta and tau neuropathology in the triple-transgenic mice. Mol Psychiatry *20*, 1329-1338.

Giannopoulos, P.F., Chiu, J., and Pratico, D. (2018). Antileukotriene therapy by reducing tau phosphorylation improves synaptic integrity and cognition of P301S transgenic mice. Aging Cell *17*, e12759.

Giannopoulos, P.F., Chiu, J., and Pratico, D. (2019). Learning Impairments, Memory Deficits, and Neuropathology in Aged Tau Transgenic Mice Are Dependent on Leukotrienes Biosynthesis: Role of the cdk5 Kinase Pathway. Mol Neurobiol *56*, 1211-1220.

Giannopoulos, P.F., Chu, J., Joshi, Y.B., Sperow, M., Li, J.G., Kirby, L.G., and Pratico, D. (2013). 5-lipoxygenase activating protein reduction ameliorates cognitive deficit, synaptic dysfunction, and neuropathology in a mouse model of Alzheimer's disease. Biol Psychiatry *74*, 348-356.

Giannopoulos, P.F., Chu, J., Joshi, Y.B., Sperow, M., Li, J.G., Kirby, L.G., and Pratico, D. (2014). Gene knockout of 5-lipoxygenase rescues synaptic dysfunction and improves memory in the triple-transgenic model of Alzheimer's disease. Mol Psychiatry *19*, 511-518.

Giannopoulos, P.F., Chu, J., Sperow, M., Li, J.G., Yu, W.H., Kirby, L.G., Abood, M., and Pratico, D. (2015). Pharmacologic inhibition of 5-lipoxygenase improves memory, rescues synaptic dysfunction, and ameliorates tau pathology in a transgenic model of tauopathy. Biol Psychiatry *78*, 693-701.

Giannopoulos, P.F., and Pratico, D. (2018). Overexpression of 5-Lipoxygenase Worsens the Phenotype of a Mouse Model of Tauopathy. Mol Neurobiol *55*, 5926-5936.

Harrison, F.E., Hosseini, A.H., and McDonald, M.P. (2009). Endogenous anxiety and stress responses in water maze and Barnes maze spatial memory tasks. Behav Brain Res *198*, 247-251.

Janus, C. (2004). Search strategies used by APP transgenic mice during navigation in the Morris water maze. Learn Mem *11*, 337-346.

Li, J.G., Chiu, J., and Pratico, D. (2019). Full recovery of the Alzheimer's disease phenotype by gain of function of vacuolar protein sorting 35. Mol Psychiatry.

Li, J.G., Chiu, J., Ramanjulu, M., Blass, B.E., and Pratico, D. (2020). A pharmacological chaperone improves memory by reducing Abeta and tau neuropathology in a mouse model with plaques and tangles. Mol Neurodegener *15*, 1.

Li, J.G., Chu, J., and Pratico, D. (2018). Downregulation of autophagy by 12/15Lipoxygenase worsens the phenotype of an Alzheimer's disease mouse model with plaques, tangles, and memory impairments. Mol Psychiatry.

Morris, R. (1984). Developments of a water-maze procedure for studying spatial learning in the rat. J Neurosci Methods *11*, 47-60.

Nicolle, M.M., Prescott, S., and Bizon, J.L. (2003). Emergence of a cue strategy preference on the water maze task in aged C57B6 x SJL F1 hybrid mice. Learn Mem *10*, 520-524.

Rogers, J., Churilov, L., Hannan, A.J., and Renoir, T. (2017). Search strategy selection in the Morris water maze indicates allocentric map formation during learning that underpins spatial memory formation. Neurobiol Learn Mem *139*, 37-49.

Stavnezer, A.J., Hyde, L.A., Bimonte, H.A., Armstrong, C.M., and Denenberg, V.H. (2002). Differential learning strategies in spatial and nonspatial versions of the Morris water maze in the C57BL/6J inbred mouse strain. Behav Brain Res *133*, 261-270.

Stover, K.R., Campbell, M.A., Van Winssen, C.M., and Brown, R.E. (2015). Early detection of cognitive deficits in the 3xTg-AD mouse model of Alzheimer's disease. Behav Brain Res *289*, 29-38.

Vagnozzi, A.N., Li, J.G., Chiu, J., Razmpour, R., Warfield, R., Ramirez, S.H., and Pratico, D. (2019). VPS35 regulates tau phosphorylation and neuropathology in tauopathy. Mol Psychiatry.