Chapter 1
Introduction

Chapter 1

Introduction

1.1. Alzheimer's disease (AD)

AD is a progressive neurodegenerative disorder and the most prevalent cause of dementia, characterised by a gradual decline in cognitive functions, including memory, executive function, language, and visuospatial abilities [1]. Clinically, AD manifests initially with episodic memory impairment, followed by deficits in other cognitive domains and behavioural disturbances as the disease advances. In its later stages, profound cognitive decline is accompanied by functional dependence and loss of autonomy. Pathologically, AD is defined by the accumulation of extracellular beta-amyloid (Aβ) plaques and intracellular neurofibrillary tangles (NFTs) composed of hyperphosphorylated tau protein [2]. The aggregates disrupt synaptic integrity and neuronal communication, ultimately leading to widespread synapse loss, neuronal degeneration, and brain atrophy, particularly in the medial temporal lobe and association cortices [3]. This progression leads to significant synaptic and neuronal loss, culminating in cognitive decline, functional impairment, and eventually death [4]. AD is broadly classified into two types: early-onset AD (EOAD), which occurs before the age of 65 and is often associated with autosomal dominant mutations in *APP*, *PSEN1*, or *PSEN2* genes; and late-onset AD (LOAD), which typically manifests after 65 years of age and is influenced by genetic risk factors such as the *APOE* ε4 allele as well as environmental and lifestyle factors [5]. Familial Alzheimer's disease (FAD), a rare subset of EOAD, accounts for less than 1% of cases [6]. The progressive nature and complex aetiology of AD make it a significant challenge in neuroscience research and clinical practice.

AD is recognized as a "global public health priority" by the World Health Organization (WHO). Approximately 55 million individuals worldwide suffer from dementia, with AD being the predominant cause of memory loss for most patients [7]. AD and dementia rank as the fourth and fifth leading causes of death globally, and AD is the primary reason for disability-adjusted life years lost in individuals over 75 years of age [8]. A mere 0.1% of cases are attributed to genetic inheritance, with symptoms typically emerging between the ages of 30 and 50. The underlying mechanisms of the disease are multifaceted, encompassing various hypotheses such as the cholinergic, tau, and amyloid hypotheses, along with factors like glutamate excitotoxicity, vitamin B5 deficiency, mitochondrial dysfunction, and neuroinflammatory processes involving biomolecules such as TDP-43 and Fyn kinase [9]. Nevertheless, the precise mechanism behind the onset and progression of AD remains unclear.

The disease is associated with a gradual compromise of the blood-brain barrier (BBB), permitting neurotoxic substances, inflammatory cells, and pathogens from the periphery to infiltrate the central nervous system (CNS). Recently, neuroinflammation has emerged as a critical component in the pathogenesis of AD. The neurodegenerative and neuroinflammatory pathways in AD are interconnected, as highlighted by a recent study from Leng *et al.*, which posits that neuroinflammation, rather than Aβ deposition, directly impacts brain network functions, leading to cognitive deficits. They propose a mechanism wherein neuroinflammation disrupts network integrity, ultimately contributing to cognitive decline [10].

1.1.1. Pathophysiology of AD

1.1.1.1. Amyloid cascade hypothesis

The accumulation of Aβ peptides in the brain is widely regarded as a hallmark of AD pathogenesis and is considered one of the earliest initiating events in the disease process.

The accumulation typically begins in the hippocampus and entorhinal cortex—regions associated with memory and spatial navigation. Concurrently, the deposition of hyperphosphorylated tau protein in the form of intracellular NFTs disrupts axonal transport and progressively alters the neuronal cytoskeleton [11].

Aβ peptides, comprising 37 to 43 amino acids, are highly resistant to proteolytic degradation. Among the various isoforms, $A\beta_{1-40}$ and $A\beta_{1-42}$ are the most prevalent, with $A\beta_{1-42}$ being more hydrophobic and neurotoxic. These peptides tend to adopt a β-pleated sheet conformation, which enhances their propensity to aggregate and form the core of amyloid plaques, which is a major pathological features of AD [12].

The amyloid cascade hypothesis, first proposed by Hardy and Higgins in 1992, positions Aβ aggregation as the central pathogenic event in AD [13]. The hypothesis initially characterized Aβ as a harmful by-product of evolution, without accounting for its potential physiological roles at normal concentrations [14]. Amyloid aggregates are typically identified *in vitro* through cross-β x-ray diffraction patterns or visualized using transmission electron microscopy or atomic force microscopy [15]. According to the hypothesis, accumulation of neurotoxic Aβ peptides in the brain leads to a cascade of events, including tau hyperphosphorylation, NFT formation, synaptic dysfunction, cognitive impairment, vascular damage, and ultimately dementia [11, 13].

1.1.1.2. Cholinergic hypothesis

The cholinergic system plays a critical role in memory, learning, cognition, and the modulation of neuronal plasticity [16]. Proposed by Peter Davies and A. J. F. Maloney in 1976, the cholinergic hypothesis was the first mechanistic theory of AD pathophysiology at the molecular level, marking a paradigm shift from descriptive neuropathology to a synaptic neurotransmission-based understanding of the disease [17].

Davies and Maloney observed significantly diminished choline acetyltransferase (ChAT) activity in the hippocampus, cortex, and amygdala of AD brains through analysis of neurotransmitter-related enzyme activity, including that of acetylcholine (ACh), γ-aminobutyric acid (GABA), dopamine (DA), noradrenaline (NA), and serotonin in 20 brain regions from AD patients and controls. Additionally, a marked reduction in acetylcholinesterase (AChE) activity was found in cortical areas exhibiting reduced ChAT activity. Notably, the regions with the greatest loss of ChAT and AChE corresponded with areas of highest NFT density [17]. ChAT catalyses ACh synthesis, utilizing choline, acetyl-CoA, and ATP as substrates [18]. The selective degeneration of cholinergic neurons in brain regions crucial for memory, attention, and conscious awareness contributes to the observed downregulation of cholinergic markers and the onset of cognitive deficits in AD [12]. Consequently, AD was, for the first time, characterized as a disorder of cholinergic failure.

According to the cholinergic hypothesis, the major pathological changes include impaired choline uptake, diminished ACh release, reduced nicotinic and muscarinic receptor expression, disrupted neurotrophin signalling, and defective axonal transport [12]. This framework underpins many therapeutic strategies aimed at enhancing synaptic ACh levels or modulating postsynaptic cholinergic receptors. Such approaches include the use of AChE inhibitors, NMDA receptor antagonists, cholinergic receptor agonists, cholinergic precursors, and allosteric modulators [12].

1.1.1.3. Glutamatergic hypothesis

Glutamate, the principal excitatory neurotransmitter in the mammalian brain, is crucial for memory, learning, cognition, and motor control. It also plays an essential role in dendritic and synaptic development and regulates synaptic strength and neural circuit activity [19]. Glutamate homeostasis is maintained via the glutamate-glutamine cycle, in which astrocytes clear excess glutamate post-synaptically to prevent excitotoxicity [20]. Glutamatergic dysfunction has been implicated in AD pathology, marked by reduced glutamate levels, diminished receptor binding, and synapse loss, particularly in the hippocampus and cortex [21]. The alterations contribute to cognitive and behavioural deficits. Aβ accumulation and NFTs further exacerbate glutamatergic dysfunction by impairing long-term potentiation and synaptic plasticity. In AD, Aβ disrupts astrocytic glutamate uptake, leading to extracellular glutamate accumulation, neuronal swelling, membrane disruption, and ultimately cell death.

Damage is especially evident in glutamatergic neurons within the cortical layers III and IV and hippocampal circuits. Pathological overactivation of NMDA receptors (particularly extrasynaptic NMDARs) results in sustained Ca^{2+} influx and neuronal toxicity [19, 21]. This underpins the therapeutic use of NMDAR antagonists, such as memantine, to mitigate excitotoxic damage. Local glutamate concentrations rise excessively near Aβ plaques, where axonal swellings may serve as ectopic release sites. During neuronal activity bursts, these concentrations may activate non-synaptic NMDA and metabotropic glutamate receptors (mGluRs), amplifying excitotoxic cascades [20].

1.1.1.4. Tau Propagation hypothesis

Tau is a microtubule-associated protein (MAP) predominantly found in neurons of the frontal, temporal, hippocampal, and entorhinal cortices. It is encoded by the MAPT gene on chromosome 17 (17q21.31) and undergoes alternative splicing at exons 2, 3, and 10 to produce six isoforms in the adult human brain [22]. The isoforms differ by the number of N-terminal inserts (0N, 1N, or 2N) and the presence of either three (3R) or four (4R) microtubule-binding repeats in the C-terminal domain. Normally, the cortex expresses nearly equal levels of 3R and 4R isoforms [23], with 50% 1N, 40% 0N, and 10% 2N tau [24]. An imbalance in 3R:4R tau isoforms has been linked to neurodegenerative tauopathies, including AD. Tau comprises four domains: an N-terminal projection domain (residues 1–150), a proline-rich region (151–243), a microtubule-binding domain, and a C-terminal tail (370–441) [23]. Tau promotes microtubule assembly and stability, ensuring efficient axoplasmic transport and maintaining synaptic plasticity [22]. It also regulates microtubule dynamics via site-specific phosphorylation [25]. The tau propagation hypothesis, proposed in 2009, posits that hyperphosphorylated tau (P-tau) disrupts microtubules, impairs neuronal signalling, and spreads between neurons in a prion-like fashion [26]. Tau pathology initially manifests in discrete regions and subsequently progresses throughout the brain. Aberrant phosphorylation, mediated by dysregulated kinases and phosphatases, can enhance tau toxicity through gain or loss-of-function effects. Additional post-translational modifications may further exacerbate tau's pathological roles. Therapeutic strategies targeting tau include kinase inhibitors, aggregation inhibitors, microtubule stabilizers, O-GlcNAcase inhibitors, and tau-directed immunotherapies [18].

1.1.2. Pathology of AD

The definitive diagnosis of AD is based on post-mortem neuropathological evaluation. Although, no single macroscopic feature is pathognomonic for AD, several structural changes, such as cortical atrophy, ventricular enlargement, and hippocampal shrinkage, are commonly observed. However, histological examination of brain tissue remains the gold standard.

Microscopic analysis of multiple brain regions using specialized staining techniques allows for the identification of hallmark lesions, i.e., amyloid plaques and neurofibrillary tangles. Diagnosis is based on the morphology, density, and anatomical distribution of these lesions [27]. These features, while not unique to AD, are central to distinguishing it from other neurodegenerative disorders. The macroscopic and microscopic features of AD are described in **Figure 1.1**.

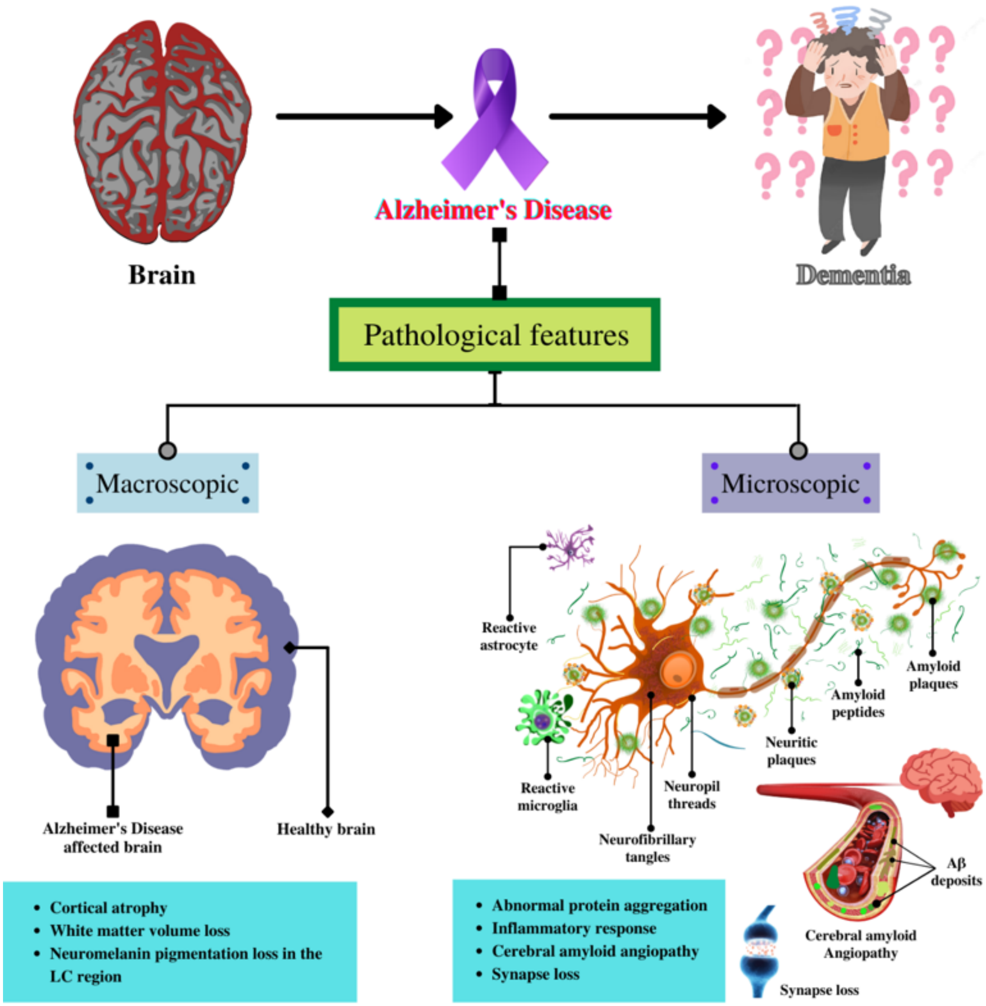

Figure 1.1 Macroscopic and microscopic features of AD

1.1.3. Stages of AD

AD progresses through a continuum of stages, primarily classified on the severity of cognitive and functional impairment. These stages reflect the pathological progression of the disease and provide a framework for clinical assessment and intervention:

a. **Preclinical stage** – This initial phase is typically asymptomatic and is marked by the early accumulation of Aβ plaques in the brain. Despite the presence of these pathological changes, individuals maintain normal cognitive function. The preclinical stage is often identified only through advanced biomarker studies or imaging techniques.

b. **Prodromal stage** – Also referred to as the stage of mild cognitive impairment (MCI) due to AD, this phase is characterized by measurable cognitive deficits that exceed age-related expectations but do not yet meet the threshold for dementia. MCI may be classified as amnestic (primarily affecting memory) or non-amnestic, and is frequently accompanied by ongoing amyloid deposition and tau pathology.

c. **Dementia stage** – This final phase is marked by significant cognitive decline that interferes with daily functioning and independence. It is subdivided into: (i) Mild AD dementia, where patients typically score above 20 on the Mini–Mental State Examination (MMSE), and (ii) Moderate to severe AD dementia, with MMSE scores below 21, reflecting profound impairments in memory, orientation, judgment, and functional abilities (**Figure 1.2**) [28].

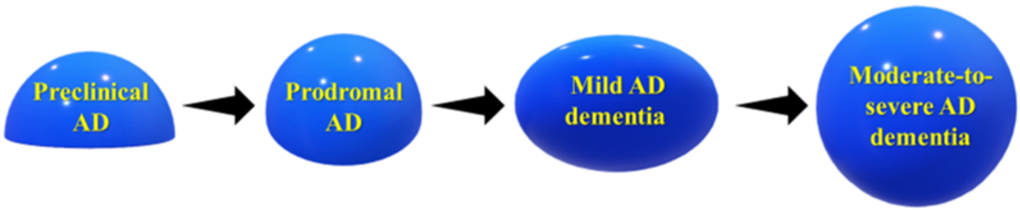

Figure 1.2 Clinical stages of AD based on the degree of cognitive impairment

Patients in the advanced stages of AD frequently exhibit pronounced disruptions in circadian rhythms and sleep-wake cycles. Notably, approximately 60% of individuals diagnosed with MCI report primary disturbances in sleep, which are recognized as significant risk factors for the subsequent development of dementia. Cronin *et al.* have attributed these circadian impairments to the early degeneration of neurons within the suprachiasmatic nucleus (SCN), the brain's central circadian pacemaker. In addition, dysregulation of peripheral circadian clocks in other brain regions has also been implicated in the initial stages of AD pathogenesis [29]. More recently, Mendonça *et al.* conducted a region-specific proteomic analysis and identified distinct molecular changes associated with tau pathology across brain areas differentially affected over the course of AD progression. Their findings revealed a complex proteomic landscape, with limited overlap but largely region-specific alterations in biological pathways, suggesting that tau-induced neurodegeneration may follow distinct molecular trajectories in different anatomical contexts [30].

1.1.4. Biomarkers of AD

AD is a multifactorial, age-associated neurodegenerative disorder characterized by multiple pathogenic events. Biomarkers from cerebrospinal fluid (CSF) and blood plasma serve as indicators of these pathological mechanisms to varying extents. Early diagnosis at the preclinical stage (prior to the emergence of observable clinical symptoms) is highly desirable, as therapeutic interventions initiated during this window are more likely to be effective. Investigating biomarkers during the early stages is crucial for drug target identification, improved clinical trial design, and enhancement of diagnostic and clinical workflows [31].

In 2011, the National Institute on Aging and Alzheimer's Association (NIA-AA) introduced diagnostic criteria for both preclinical and clinical stages of AD, including MCI and dementia [32-35]. Since then, AD has been increasingly recognized as a biological continuum rather than a series of discrete stages. Biomarkers such as Aβ-PET imaging (for amyloid plaques), tau-PET imaging, and CSF $A\beta_{42}$ or $A\beta_{42}/A\beta_{40}$ ratio have emerged as proxies for neuropathological changes. In 2018, Jack *et al.* proposed an updated, biomarker-based classification framework, called the "A/T/N" system, where "A" denotes amyloid pathology (e.g., CSF $A\beta_{42}$, amyloid PET), "T" represents tau pathology (e.g., CSF phosphorylated tau, tau PET), and "N" reflects neurodegeneration (e.g., CSF total tau, FDG-PET hypometabolism, MRI atrophy) [54]. The group also introduced a six-stage numerical clinical staging system, tracking disease progression from asymptomatic biomarker positivity to severe dementia [36].

Emerging biomarkers of neuroinflammation from CSF and serum may further enhance diagnostic accuracy, particularly in older adults [37]. Currently, CSF Aβ and tau remain the most specific fluid biomarkers for AD [38]. Promising advances in peripheral blood-based biomarkers offer less invasive alternatives for AD diagnosis [33, 39]. Comparative studies between CSF and plasma biomarkers may elucidate disease mechanisms detectable through blood-based tests [40].

1.1.5. Diagnosis of AD

The clinical diagnosis of AD primarily relies on the identification of characteristic signs and symptoms of dementia [41]. The diagnostic process typically follows a two-step approach:

- **a.** Differentiation of dementia syndromes from other conditions with overlapping clinical features, such as depression, delirium, and MCI.
- **b.** Determination of the specific dementia subtype to inform appropriate treatment strategies.

In routine clinical settings, cognitive screening tools, such as the clock-drawing test, are widely employed due to their simplicity and non-confrontational nature. Successful completion of this test is generally indicative of intact cognitive function. Evaluation of activities of daily living is also essential alongside cognitive testing. A structured clinical assessment includes:

- **a.** Detailed history-taking, incorporating information from caregivers or informants.
- **b.** Mental status examination using validated cognitive tests.

c. Physical examination, with a focus on neurological and vascular signs, supplemented by relevant investigations [42].

In recent years, novel diagnostic modalities have emerged, expanding the diagnostic landscape of AD. These include: biomarker-based diagnostics, neuroimaging techniques (e.g., PET, MRI), transcranial magnetic stimulation (TMS), electroencephalography (EEG), clinical decision support systems (CDSS), sensory assessments (e.g., olfaction, gustation, vision), salivary biomarker analysis, speech analysis, comprehensive neuropsychological testing [41].

1.1.6. Management of AD

The current therapeutic landscape for AD includes both pharmacological and non-pharmacological interventions, with a growing emphasis on disease-modifying therapies and multimodal approaches.

1.1.6.1. Pharmacological treatments

Symptomatic therapies remain the cornerstone for most patients. AChE inhibitors, i.e., donepezil [43], rivastigmine [44], and galantamine [45], are approved for mild to moderate AD and function by enhancing cholinergic neurotransmission. Memantine [46], an NMDA receptor antagonist, is indicated for moderate to severe AD and works by modulating glutamatergic signalling to reduce excitotoxicity. These agents provide modest improvements in cognition and daily functioning but do not alter disease progression (**Figure 1.3** and **Table 1.1**) [47-49].

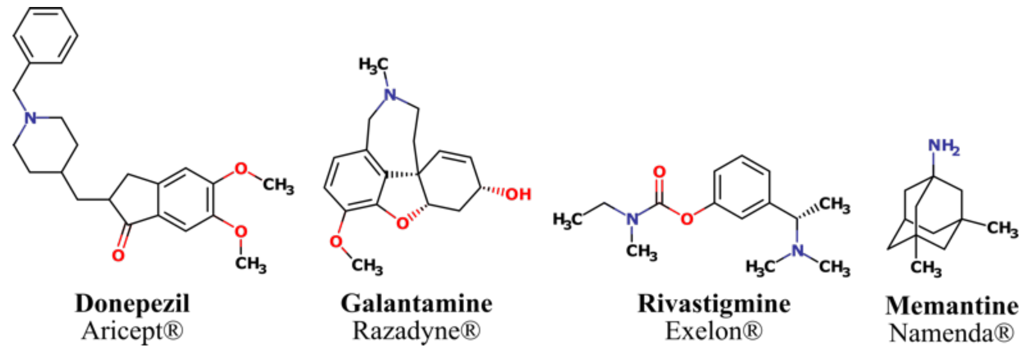

Figure 1.3 FDA approved drugs for the treatment of AD

Table 1.1 FDA approved drug treatments for AD

Compound	Chemical class	Mechanism of action	Approved for	Approval year	Side effects
Donepezil Aricept®	Piperidine derivative	Reversible AChE inhibitors	All stages	1996	Nausea, vomiting, loss of appetite, muscle cramps and increased frequency of bowel movements.
Galantamine Razadyne®	Alkaloid	Reversible AChE inhibitors	Mild to moderate	2001	Nausea, vomiting, loss of appetite and increased frequency of bowel movements.
Rivastigmine Exelon®	Carbamate	Reversible AChE inhibitors	Mild to moderate	1998	Nausea, vomiting, loss of appetite and increased frequency of bowel movements.
Memantine Namenda®	Primary aliphatic amine	Uncompetitive (open-channel) NMDA receptor antagonist	Moderate to severe	2004	Headache, constipation, confusion and dizziness.
Memantine + Donepezil Namzaric®	Primary aliphatic amine + Piperidine derivative	Uncompetitive (open-channel) NMDA receptor antagonist + Reversible AChE inhibitor	Moderate to severe	2014	Nausea, vomiting, loss of appetite, increased frequency of bowel movements, headache, constipation, confusion and dizziness.

Disease-modifying therapies have recently shifted the paradigm. Monoclonal antibodies targeting Aβ pathology, such as aducanumab (currently discontinued) [50], lecanemab [51], and donanemab [52], have received regulatory approval for the treatment of early-stage AD. These agents promote the clearance of Aβ plaques and have demonstrated efficacy in slowing cognitive decline in clinical trials, although their use is associated with risks such as amyloid-related imaging abnormalities (ARIA) and requires careful patient selection and monitoring [53-57]. Other investigational drugs are targeting tau pathology and neuroinflammation, reflecting the multifactorial nature of AD pathogenesis [49].

Drug repurposing strategies are also under investigation, evaluating agents from anti-inflammatory, anti-hypertensive, anti-diabetic, antiepileptic, and anticancer classes for their potential neuroprotective effects in AD [58].

1.1.6.2. Non-pharmacological treatments

Non-pharmacological interventions play a vital role in comprehensive AD management. Evidence-based approaches include:

- Cognitive stimulation and training: Structured activities and exercises to enhance memory, attention, and problem-solving skills [59].
- Physical exercise: Regular aerobic and resistance activities improve cognitive outcomes and overall health [60].
- Neuromodulation techniques: Transcranial direct current stimulation (tDCS) and repetitive transcranial magnetic stimulation (rTMS) have shown efficacy in improving global cognitive function [59].
- Reminiscence therapy, occupational therapy, and animal-assisted therapy: These modalities support emotional well-being, daily functioning, and quality of life [61-63].
- Multidisciplinary care: Combining pharmacological and non-pharmacological treatments, caregiver support, and lifestyle modifications yields the best outcomes [61].

AD treatment is evolving from purely symptomatic management to disease-modifying and multimodal strategies. Cholinesterase inhibitors and memantine remain standards for symptomatic relief, while monoclonal antibodies targeting amyloid-β represent a significant advance in early-stage intervention. Non-pharmacological therapies are essential adjuncts, improving cognitive and functional outcomes and patient quality of life. Ongoing research into tau-targeted therapies, neuroinflammation, and drug repurposing holds promise for future breakthroughs in AD management.

1.2. Computational chemistry

Computational chemistry refers to the application of computer modelling and simulation techniques to study, predict, and optimize the properties of drug molecules and their interactions with biological targets. This approach encompasses both quantum chemistry-based methods and empirical techniques to accelerate drug discovery and development processes. In the context of medicinal chemistry, computational methods are employed to:

- Identify and optimize lead compounds through virtual screening and structure-based drug design.
- Predict pharmacokinetic and toxicological properties of drug candidates.
- Simulate drug-target interactions and binding affinities using molecular docking and dynamics simulations.
- Develop quantitative structure-activity relationships for predicting biological activity.
- Elucidate reaction mechanisms and energetics of drug-target interactions using quantum mechanical calculations [64-66].

The computational approaches significantly reduce the time and cost associated with traditional experimental methods in drug discovery, enabling the exploration of vast chemical spaces and the rational design of novel therapeutic agents.

Computer-aided drug design (CADD) is a specialized application of computational chemistry in the field of drug discovery and development. CADD utilises computational methods and algorithms to simulate, predict, and analyse the interactions between drug molecules and their biological targets. CADD integrates the computational chemistry approaches with biological data to accelerate the drug discovery process, from target analysis and virtual screening to lead optimisation and prediction of druggability. By leveraging computational power, CADD enhances the efficiency and cost-effectiveness of drug development compared to traditional experimental methods [67].

1.2.1. Computer-aided drug design (CADD)

CADD is a sophisticated approach that leverages computational methods to streamline and accelerate the drug discovery and development process. CADD encompasses two primary strategies: structure-based drug design (SBDD) and ligand-based drug design (LBDD). SBDD utilises three-dimensional structural information of target proteins to design and optimise potential drug candidates, employing techniques such as molecular docking, de novo design, and structure-based virtual screening. LBDD, conversely, relies on known active compounds to predict and design novel ligands, utilising methods like pharmacophore modelling, quantitative structure-activity relationships (QSAR), and similarity searches.

CADD methodologies integrate various computational tools, including molecular mechanics, quantum mechanics, and statistical mechanics, to model and predict drug-target interactions, binding affinities, and pharmacokinetic properties. Advanced algorithms and machine learning techniques are increasingly employed to enhance predictive accuracy and efficiency. Virtual screening, a key component of CADD, enables the rapid evaluation of large compound libraries against specific targets, significantly reducing the time and resources required for experimental high-throughput screening.

The implementation of CADD in pharmaceutical research has demonstrated substantial benefits, including accelerated hit identification, lead optimization, and the prediction of absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties. This computational approach not only expedites the drug discovery pipeline but also contributes to reducing the overall cost and failure rates in later stages of drug development [68-70].

1.2.1.1. Structure-based drug design (SBDD)

SBDD represents a sophisticated methodology within the field of medicinal chemistry, leveraging the three-dimensional (3D) structures of biological macromolecules, primarily proteins, to facilitate the development of novel therapeutic agents. This approach has gained considerable prominence due to its capacity to streamline the drug discovery process, rendering it more efficient and targeted compared to traditional methodologies [71].

The fundamental principle underpinning SBDD is the "lock-and-key" model, wherein a drug molecule is meticulously designed to fit precisely into a specific binding site on a target protein, analogous to how a key fits into a lock. This interaction can either activate or inhibit the function of the protein, contingent upon the nature of the binding [72, 73]. The initial step in SBDD involves obtaining high-resolution 3D structures of target proteins. Techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy are commonly employed to achieve this.

Once the structure is determined, it provides critical insights into potential binding sites and the molecular environment surrounding them [74].

Key Techniques in SBDD

- **Molecular docking**:

  Molecular docking is a computational technique that predicts how small molecules interact with target proteins. By screening libraries of compounds against known binding sites, researchers can identify candidates that demonstrate favourable binding affinities. This method allows for the virtual screening of thousands of compounds, significantly reducing the time and resources required for experimental testing [75].

- **Structure-activity relationship (SAR) Analysis**:

  SAR analysis entails modifying the chemical structure of lead compounds to optimize their binding affinity and biological activity. By systematically altering functional groups and assessing their effects on activity, researchers can identify key structural features that enhance interactions with the target protein [76].

- **Virtual screening**:

  Virtual screening employs computational methods to evaluate large libraries of compounds for their potential to bind to target proteins. This technique integrates molecular docking and scoring functions to prioritize compounds for subsequent experimental validation [77].

- ***De novo* drug design**:

  In *de novo* drug design, novel chemical entities are constructed based on the 3D structure of the target protein. The approach enables the design of new molecules tailored specifically to fit into the binding site, potentially resulting in more effective drugs [78].

- **Pharmacophore modelling**:

  Pharmacophore modelling identifies the essential features required for molecular recognition between a ligand and its target. By mapping these features onto potential drug candidates, researchers can screen compound libraries more effectively [79].

1.2.1.2. Ligand-based drug design (LBDD)

LBDD represents a pivotal strategy in the realm of drug discovery, particularly in instances where the 3D structures of target proteins are not readily available. This approach capitalizes on existing knowledge of ligands (molecules that bind to biological targets) to facilitate the development of new pharmacologically active compounds [80, 81]. The underlying principle of LBDD is encapsulated in the notion that structurally similar molecules often exhibit analogous biological activities, a concept foundational to SAR.

LBDD encompasses a variety of methodologies, including QSAR and pharmacophore modelling. QSAR serves as a computational technique that quantifies the correlation between chemical structure and biological activity, thus enabling researchers to predict the efficacy of new compounds based on their structural characteristics. Pharmacophore modelling, on the other hand, identifies the essential structural features requisite for ligand-target interactions, thereby providing a framework for the design of novel molecules with desired pharmacological properties [80].

Key Techniques in LBDD

- **QSAR**:
  - **a.** QSAR models establish a correlation between chemical structure and biological activity, facilitating predictions regarding how modifications to a compound may enhance its efficacy or mitigate side effects.
  - **b.** These models take into account various physicochemical properties, including steric, electronic, and hydrophobic characteristics [82, 83].
- **Pharmacophore modelling**:
  - **a.** This technique involves the creation of a model that represents the spatial arrangement of features necessary for molecular recognition between a ligand and its target.
  - **b.** Pharmacophore models can be employed to screen extensive compound databases, identifying potential drug candidates that conform to the established criteria [79].
- **Molecular docking**:
  - **a.** While primarily associated with structure-based drug design, molecular docking can complement LBDD by predicting the binding of small molecules to receptors.
  - **b.** This technique aids in evaluating the binding affinity and interaction strength between ligands and their targets, thereby informing subsequent optimization efforts [75].

1.2.1.3. Integrated ligand-based drug design and structure-based drug design

Integrated LBDD and SBDD combine the strengths of both methodologies to enhance the drug discovery process. LBDD is particularly advantageous when the 3D structures of target proteins are unavailable, relying on existing ligand information to create pharmacophore models and predict new compounds with desired biological activity. Techniques such as QSAR and scaffold hopping are employed to optimise lead compounds based on their structural similarities to known active molecules. Conversely, SBDD utilizes detailed structural data of target proteins to inform the design of new drugs. By employing molecular docking and virtual screening, researchers can elucidate how potential ligands interact with specific binding sites, facilitating precise optimization of drug candidates.

The integration of LBDD and SBDD fosters a more comprehensive approach to drug development. For instance, pharmacophore models derived from known ligands can guide the initial stages of compound screening, while subsequent optimization can leverage structural insights gained from SBDD. This synergy not only accelerates the identification of promising drug candidates but also enhances their efficacy and specificity, ultimately leading to more effective therapeutics within a reduced timeframe [84, 85].

1.2.2. Integration of big data, AI, and ML in advanced CADD

The advancement of CADD is increasingly incorporating big data, artificial intelligence (AI), and machine learning (ML) due to their potential to significantly enhance the drug discovery process. These technologies enable the analysis of vast and diverse datasets, including genomic information, clinical trial data, and real-world evidence, to identify potential drug targets and optimize lead compounds [86]. AI and ML algorithms, particularly deep learning approaches, can process and interpret complex biological and chemical data more efficiently than traditional methods, leading to faster and more accurate predictions of drug-target interactions, pharmacokinetic properties, and potential toxicities [87]. This integration accelerates the drug discovery pipeline, reduces attrition rates, and potentially lowers the overall cost and time of bringing new therapeutics to market [88].

1.3. Machine learning: approaches, algorithms, and validation

ML is a subset of AI that focuses on the development of algorithms capable of learning from and making predictions based on data. This field has garnered significant attention across various domains, including healthcare, finance, and technology, due to its capacity to analyse extensive datasets and extract meaningful patterns. The significance of ML resides in its ability to automate decision-making processes, enhance predictive analytics, and improve operational efficiencies [89].

1.3.1. Approaches to ML

ML can be categorized into three primary approaches:

a. **Supervised learning**

   This method uses labeled training data to learn a function that maps inputs to specific outputs. It is a task-driven approach where the model is trained on a dataset with known outcomes. Common tasks include classification (separating data into categories) and regression (fitting data to a continuous output). Supervised learning is particularly useful when clear goals are identified for a given set of inputs. For example, predicting the sentiment of a product review or classifying images are typical supervised learning tasks. Algorithms such as logistic regression, support vector machines, and decision trees are prevalent in this approach.

b. **Unsupervised learning**

   This data-driven approach analyzes unlabeled datasets without human intervention. It is used to extract generative features, identify meaningful trends and structures, and explore data. Common unsupervised tasks include clustering (grouping similar data points), density estimation, feature learning, dimensionality reduction, and anomaly detection. Unsupervised learning is valuable for discovering hidden patterns in data when the desired output is not known in advance. Notable methods include k-means clustering and principal component analysis (PCA).

c. **Semi-supervised learning**

This hybrid approach combines elements of supervised and unsupervised learning, working with both labeled and unlabeled data. It is particularly useful in real-world scenarios where labeled data is scarce but unlabeled data is abundant. The goal is to improve predictive performance beyond what could be achieved using only the labeled data. Applications include machine translation, fraud detection, and text classification.

d. **Reinforcement learning**

This environment-driven approach enables software agents to automatically determine optimal behavior within a specific context to maximize performance. It is based on a system of rewards and penalties, where the agent learns to make decisions that increase rewards or minimize risks through interaction with its environment. Reinforcement learning is powerful for training AI models in complex, dynamic environments such as robotics, autonomous driving, and supply chain optimization. However, it is generally not preferred for simple or straightforward problems due to its complexity and resource requirements.

Each of the approaches has its strengths and is applied based on the nature of the available data, the complexity of the problem, and the specific goals of the machine learning task [90].

1.3.2. Machine learning algorithms

Numerous algorithms are utilized within these approaches, each suited to specific tasks:

- **Classification algorithms**: Logistic regression, support vector machines, decision trees, random forests, k-nearest neighbours.
- **Regression algorithms**: Linear regression, polynomial regression, ridge, and lasso regression.
- **Clustering algorithms**: K-means clustering, hierarchical clustering, DBSCAN.
- **Dimensionality reduction techniques**: PCA, t-distributed stochastic neighbour embedding (t-SNE)
- **Deep learning models**: Convolutional neural networks (CNNs) for image processing and recurrent neural networks (RNNs) for sequential data.

1.3.3. Validation of machine learning models

The validation of machine learning models is critical to ensuring their reliability and performance. The validation process typically involves several key steps:

a. **Data splitting**:

   Data splitting in machine learning divides datasets into training, validation, and test sets. This crucial process helps prevent overfitting and ensures model generalization. The training set is used for model learning, validation for tuning hyperparameters, and the test set for final performance evaluation. Common methods include random sampling, stratified splitting, and k-fold cross-validation [91].

b. **Performance metrics**:

   Various metrics are employed to assess model performance, depending on the specific task. For classification tasks, metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC) are commonly used. For regression tasks, metrics including mean absolute error (MAE), mean squared error (MSE), and R-squared are employed [92].

c. **Cross-validation**:

Cross-validation is a robust statistical technique employed in machine learning to assess model performance and generalizability. It involves partitioning data into subsets, iteratively training and validating models on different combinations. K-fold cross-validation is commonly utilized, where data is divided into k subsets. This method mitigates overfitting, provides reliable performance estimates, and optimizes hyperparameter selection, enhancing model robustness and predictive accuracy [93].

d. **Hyperparameter tuning**:

Hyperparameter tuning is a critical process in machine learning that optimizes model performance by selecting optimal configuration variables. It addresses the bias-variance tradeoff and enhances generalization. Methods include random search, grid search, and Bayesian optimization. This iterative process impacts model structure, training efficiency, and accuracy. While computationally intensive, it is essential for maximizing model effectiveness and resource efficiency in both supervised and unsupervised learning paradigms [94].

e. **Overfitting and underfitting**:

Overfitting occurs when a model learns training data too well, including noise, leading to poor generalization on new data. Underfitting happens when a model is too simple to capture underlying patterns, resulting in poor performance on both training and new data. Balancing model complexity is crucial to avoid these issues and achieve optimal generalization. Techniques like cross-validation, regularization, and early stopping help mitigate overfitting and underfitting [95].

1.3.4. Performance metrics

Selecting appropriate metrics is essential for evaluating model performance effectively. The following metrics are commonly employed:

1.3.4.1. Classification model validation using confusion matrix

Validating the performance of classification models is essential to ensure accurate outcome predictions. A widely used method for this validation is the confusion matrix, which summarizes the performance of a classification algorithm by comparing predicted classifications with actual classifications.

The confusion matrix comprises four key components:

a. **True Positive (TP)**: The number of instances correctly predicted as positive.
b. **True Negative (TN)**: The number of instances correctly predicted as negative.
c. **False Positive (FP)**: The number of instances incorrectly predicted as positive.
d. **False Negative (FN)**: The number of instances incorrectly predicted as negative [96].

From these components, several performance metrics can be derived to evaluate the model's effectiveness:

i. **Accuracy**: Accuracy in machine learning quantifies the proportion of correct predictions relative to the total number of cases evaluated. While widely used, it can be misleading for imbalanced datasets. It is calculated as (True Positives + True Negatives) / Total Observations. Accuracy should be considered alongside other metrics like precision, recall, and F1-score for comprehensive model evaluation, especially in classification tasks.

ii. **Balanced accuracy:** Balanced accuracy is a robust performance metric for classification models, particularly in imbalanced datasets. It is calculated as the arithmetic mean of sensitivity and specificity, ranging from 0 to 1. This metric mitigates bias towards majority classes, providing a more equitable evaluation of model performance across all classes. Balanced accuracy is crucial for assessing models in domains with class imbalance, such as medical diagnostics.

iii. **Precision:** Precision indicates the proportion of true positive predictions among all positive predictions made by the model. High precision suggests that when the model predicts a positive class, it is often correct, which is particularly important in scenarios where false positives have significant costs.

iv. **Recall (Sensitivity)**: Recall measures the proportion of actual positives that were correctly identified by the model. A high recall indicates that most actual positive cases are captured, which is crucial in applications such as disease detection, where failing to identify a positive case can have serious implications [96].

v. **F1 Score**: The F1 score is a crucial performance metric in machine learning, particularly for binary classification tasks. It represents the harmonic mean of precision and recall, providing a balanced measure of a model's accuracy. The F1 score ranges from 0 to 1, with 1 indicating optimal performance. It is especially valuable in scenarios with imbalanced datasets, offering a more comprehensive evaluation than accuracy alone [97].

vi. **Matthews correlation coefficient (MCC)**: MCC is a statistical measure that quantifies the quality of binary classifications in machine learning. It is derived from the confusion matrix and provides a balanced assessment of model performance, particularly when dealing with imbalanced datasets. This metric is valued for its ability to provide a single, comprehensive measure of classification performance, even when class sizes vary significantly. MCC ranges from -1 to +1, where +1 indicates perfect prediction, 0 suggests no better than random prediction, and -1 implies total disagreement between prediction and observation [96].

1.4. Multi-target-directed ligands (MTDLs)

MTDLs represent a paradigm shift in drug discovery, addressing the limitations of single-target therapies for complex multifactorial diseases. These compounds are designed to simultaneously modulate multiple pathological targets, offering synergistic therapeutic effects and improved clinical outcomes [98, 99]. The rationale for MTDLs stems from the interconnected nature of disease pathways, where single-target inhibition often fails to halt progression due to compensatory mechanisms or network redundancy [100, 101]. MTDLs integrate structural features of distinct pharmacophores, enabling interactions with diverse targets such as enzymes, receptors, and protein aggregates [99, 100]. This approach has gained prominence in neurodegenerative disorders, cancer, and metabolic diseases, where polypharmacology enhances efficacy while reducing adverse effects [98, 101]. Key advantages include reduced drug-drug interactions, simplified pharmacokinetic profiles, and lower development costs compared to combination therapies [100, 102]. Modern MTDL design employs computational strategies like molecular hybridization and pharmacophore fusion, coupled with ML to optimize target affinity [101, 103].

1.4.1. Molecular hybridization in MTDLs

Molecular hybridization has emerged as a cornerstone strategy for designing MTDLs, addressing the inherent complexity of multifactorial diseases. This approach combines structural motifs from distinct pharmacophores into a single chemical entity, enabling simultaneous modulation of multiple disease-relevant targets [104]. By merging fragments of known bioactive molecules, hybrid MTDLs retain or enhance parent compounds' activities while mitigating off-target effects and pharmacokinetic limitations [105]. The rational design of hybrid MTDLs leverages computational tools like pharmacophore mapping and molecular docking to identify optimal fusion points, ensuring minimal steric clashes and preserved target interactions [106]. As polypharmacology gains traction, molecular hybridisation remains pivotal for developing next-generation therapeutics with enhanced therapeutic windows and reduced side-effect profiles.

1.4.2. Lead optimisation in MTDLs

Lead optimisation is a critical phase in the development of MTDLs, involving systematic modifications of a lead compound to enhance its polypharmacological profile while maintaining desirable pharmacokinetic and physicochemical properties. This process requires a careful balance between achieving potent activity at multiple targets and minimising off-target effects, toxicity, and poor bioavailability [107, 108]. Through iterative cycles of SAR studies, computational modelling, and *in vitro/in vivo* validation, lead optimisation enables the refinement of molecular scaffolds to improve efficacy and target selectivity in a coordinated manner. Advances in cheminformatics, machine learning, and molecular docking have further accelerated this process by predicting multi-target interactions and prioritizing compounds with optimal polypharmacological profiles [109, 110]. Moreover, absorption, distribution, metabolism, excretion, and toxicity (ADMET) profiling during lead optimization ensures that the multifunctionality of MTDLs is coupled with drug-like properties, increasing the likelihood of clinical success. Thus, lead optimization serves as an indispensable tool in the rational design of MTDLs for the treatment of complex disorders.