University of Szeged

Albert Szent-Györgyi Medical School

Doctoral School of Experimental and Preventive Medicine

Computational analysis of protein-ligand complexes and peptides using molecular mechanics and machine learning methods

Ph.D. Thesis

Vivien Erzsébet Resch

Supervisor:

Dr. habil. Gábor Paragi

Szeged, 2026

Contents

LIST OF PUBLICATIONS

Publications related to the subject of the thesis:

1. Resch, Vivien, Marija Gjorgoska, Eva Hafner, Ildikó Bacsa, Benjamin Kovács, Tomaž Büdefeld, Attila Hunyadi, Ildikó Huliák, Mónika Kiricsi, Gábor Paragi, and et al. "Targeting Steroid-Metabolizing Enzymes with 15β-Substituted Estrone Analogues: Dual Discovery of AKR1C2/17β-HSD1 Inhibitors and a Fluorescent 17β-HSD1 Ligand" **Cancers**, 2026; 18(12): 1889. doi:10.3390/cancers18121889. **SJR indicator Q1, IF: 4.8** (2025)

2. Ali, Hazhmat, Péter Traj, Gábor J. Szebeni, Nikolett Gémes, Vivien Resch, Gábor Paragi, Erzsébet Mernyák, Renáta Minorics, and István Zupkó. "Investigation of the Antineoplastic Effects of 2-(4-Chlorophenyl)-13α-Estrone Sulfamate against the HPV16-Positive Human Invasive Cervical Carcinoma Cell Line SiHa" **International Journal of Molecular Sciences***,* 2023; 24(7): 6625. doi:10.3390/ijms24076625. **SJR indicator D1, IF: 4.9** (2025)

3. Vivien Erzsébet Resch, László Tóth, Anita Rácz, Dezső Virok, Gábor Paragi, "Machine Learning-based prediction of Cross Immunity" **Briefing in Bioinformatics**, 2026; (**accepted***),* **SJR indicator D1, IF: 7.3** (2025)

4. Vivien Erzsébet Resch, Anita Rácz, Gábor Paragi
   "Mutation-Induced Changes in Nucleation Scores of Aβ-peptide: A Machine Learning Study" **Journal of Cheminformatics** (in preparation)

5. Vivien Erzsébet Resch, Edit Wéber, Zsófia Hegedűs, Ádám Juhász, Edit Csapó, Bálint Lőrinczy, István Szatmári, Gábor Paragi
   "Binding properties of Kyuneric Acid Analogues to Serum Albumins" **Journal of Drug Delivery Science and Technology** (submitted)

Publications not related to the subject of the thesis:

6 György Szőllősi, Vivien Erzsébet Resch, Vanessza Judit Kolcsár, “Asymmetric transfer hydrogenation of 2,3-diphenylpropenoic acids over heterogeneous palladium catalysts modified by cinchona alkaloids”, **Journal of Catalysis**, 2024; 429: 115290, doi: 10.1016/j.jcat.2024.115290. **SJR indicator Q1, IF: 6.5** (2024)

7. Bózsity, Noémi, Viktória Nagy, Johanna Szabó, Balázs Pálházi, Zoltán Kele, Vivien Resch, Gábor Paragi, István Zupkó, Renáta Minorics, and Erzsébet Mernyák, "Synthesis of Estrone Heterodimers and Evaluation of Their In Vitro Antiproliferative Activity" **International Journal of Molecular Sciences**, 2024; 25(8): 4274. doi: 10.3390/ijms25084274. **SJR indicator Q1, IF: 5.6** (2025)

8. Njangiru, Isaac Kinyua, Noémi Bózsity-Faragó, Vivien Erzsébet Resch, Gábor Paragi, Éva Frank, György T. Balogh, István Zupkó, and Renáta Minorics, “A Novel 2-Methoxyestradiol Derivative: Disrupting Mitosis Inhibiting Cell Motility and Inducing Apoptosis in HeLa Cells In Vitro” **Pharmaceutics**, 2024; 16(5): 622. doi: 10.3390/pharmaceutics16050622**. SJR indicator Q1, IF: 6.9** (2025)

9. Kovács, Édua, Hazhmat Ali, Renáta Minorics, Péter Traj, Vivien Resch, Gábor Paragi, Bella Bruszel, István Zupkó, and Erzsébet Mernyák. 2023. "Synthesis and Antiproliferative Activity of Steroidal Diaryl Ethers" **Molecules**, 2023; 28(3): 1196. doi: 10.3390/molecules28031196. **SJR indicator Q1, IF: 5.1** (2025)

10. Mernyák, Erzsébet, Sándor Bartha, Lili Kóczán, Rebeka Jójárt, Vivien Resch, Gábor Paragi, Máté Vágvölgyi, et al. “Microwave-Assisted Phospha-Michael Addition Reactions in the 13α-Oestrone Series and in Vitro Antiproliferative Properties.” **Journal of Enzyme Inhibition and Medicinal Chemistry**, 2021; 36(1): 1931–37. doi:10.1080/14756366.2021.1963241. **SJR indicator Q2, IF: 5.3** (2025)

LIST OF ABBREVIATIONS

Alzheimer's disease	AD
aldo-keto reductase	AKR
amyloid-β	Aβ
amyloid precursor protein	APP
artificial intelligence	AI
bovine serum albumin	BSA
binding free energy	ΔG
central nervous system	CNS
feed-forward neural network	FFNN
force field	ff
human leukotic antigen	HLA-A
human serum albumin	HSA
isothermal titration calorimetry	ITC
kynurenic acid	KYNA
large language model	LLM
logistic regression	LR
machine learning	ML
major histocompatibility complex I	MHC-I
molecular dynamics	MD
molecular mechanics	MM
molecular mechanics/generalized born surface area	MM/GBSA
potential mean force	PMF
protein-ligand complex	PL complex
quantum mechanics	QM
random forest	RF
saturation transfer difference NMR	STD-NMR
support vector machine	SVM
support vector regressor	SVR
umbrella sampling	US

1. Introduction and biological background

Understanding complex biological systems requires both molecular and macroscopical perspectives. For instance, the overall function of an enzyme is influenced by its macroscopic environment whereas its catalytic activity arises from molecular and electronic interactions. Due to this inherent complexity, theoretical frameworks are essential for describing molecular structures and interactions in a computationally manageable manner. A possible approach used in molecular modelling is the application of quantum mechanics (QM) [1,2], where the molecular behaviour based on their electronic structure and the particles are treated as a collection of nuclei and electrons without predefined chemical bonds [1,2]. The energy of the system is obtained by solving the Schrödinger-equation, however, equation cannot be solved exactly especially for large molecular systems, like ligand-protein modelling. A probable solution of this problematic limitation is the classical molecular mechanics (MM) [2,3], which follows the Newtonian classical physics, where the molecules contain point-like charged particles with specific connection and the electronic system is not described explicitly [2].

A fundamentally new and successful paradigm in computational molecular modelling is based on the artificial intelligence (AI) techniques. This was recognized by the 2024 Nobel Prize awarded for the development of AlphaFold, which addressed the long-standing protein folding problem by enabling high accurate 3D protein structure prediction from amino acid sequences [4]. While, the most popular AI sort is the large language models (LLMs) like ChatGPT, Gemini etc., which is typically based on a neural network with transformer architecture [5], machine learning (ML) methods can offer several flexible ways to model complex and noisy biological data, particularly when the relationships are not easily captured by explicit rules [6]. Frequently used types are the deep learning methods (feed forward neural networks – FFNN) [7,8], shallow ML methods (support vector machine – SVM, support vector regressor - SVR) [9–11], supervised ML algorithms (logistic regression – LR) [12] and ensemble methods (random forest – RF) [13].

These computational approaches provide molecular-level interpretation of experimental observations and yield valuable insights; however, they are designed to complement rather than replace laboratory experiments at all. During my doctoral studies I applied wide range of the above mentioned simulation techniques from MM level molecular modelling to AI-based investigations across diverse physiochemical systems. Although, the affected biological systems were highly heterogenous, they were all connected to the central nervous system (CNS), either directly or indirectly to varying degrees. In the introduction part, I focus on these

CNS-related contacts; however, I am also aware that some of their individual investigations, such as those involving the immune system, have broader biological implications beyond the CNS.

1.1. *Role of the tubulin proteins in CNS*

Building on this biological context, part of the present research focused on the role of tubulin proteins in the CNS. Mostly the CNS is built up of nerve cells which consist of two main parts as axon and dendrite supported by stable microtubules and neurofilaments. The microtubules are a component of the cytoskeleton that contribute to the intracellular transport, cell division and the preservation of the cell shape [14]. The tubulin dimers contain α and β subunits and this structure can be polymerized as well as depolymerized, causing microtubules to assemble and disassemble in random cycles. That unique function is the dynamic instability [14], which can be modified by the interactions of microtubules with other proteins. Some cellular proteins, called as microtubule-associated proteins (MAPs) or tau proteins bind to microtubules and increase their stability while other proteins can disassemble microtubules. These proteins can lock the microtubules together and inhibiting their disassembly which blocks the mitosis [15]. The microtubule-targeting agents (e.g. taxene [16]) participate in that mechanism also causing cell cycle arrest and apoptosis like in the faster divided cancer cells [17]. Steroid molecules can act as regulators of microtubule dynamics by binding directly to MAP2, for example, the neurosteroid pregnenolone promotes microtubule assembly, whereas others steroids, such as progesterone can antagonize this effect [18].

1.2. *Role of the AKRs in CNS*

From a broader biological perspective of steroids, my work also related to the aldo-keto-reductases (AKRs), particularly the AKR1C subfamily, which are key managers of steroid metabolism (like estrogen, androgen, progesterone [19]). These steroid hormones are regulatory molecules in the human body that control cell growth, differentiation, stress response, and metabolic processes. Dysregulation of steroid metabolism can lead to pathophysiological conditions in peripheral organs because the elevated levels of biologically active steroids promote cellular proliferation and contribute to the development of prostate and breast or cervical carcinoma pathologies in hormone-sensitive tissues [20]. In addition to their endocrine effects, the members of AKR1C1-C3 subfamily also regulate neurosteroid homeostasis, modulate neurotransmission and support the antioxidant defence mechanism of the CNS. For

example, the well-known neurosteroid, dehydroepiandrosterone act as negative allosteric modulator of $GABA_A$ receptors, by reducing chloride reflux in neurons and decreasing neuronal activity [21]. These properties imply their involvement in neuroprotection, and neurodegenerative conditions, including Parkinson's and Alzheimer's disease (AD) [22].

1.3. *The Aβ peptide and its mutations*

Another major area of the present research focused on the hallmark protein abnormalities disrupt neural communications and synaptic conditions, especially in AD. One of them is extracellular amyloid-β (Aβ) plaques derived from amyloid precursor protein (APP) [23,24]. APP is a type 1 membrane glycoprotein localized primarily at synapses. By spanning the cellular membrane, it functions as a cell surface receptor that important for synaptic formation. Under physiological conditions, α-secretases (ADAM10) cleave APP within its transmembrane domain, which contains Aβ resultant $sAPP_\alpha$ fragments (see more in Figure 1). Therefore, the processing of APP by β- and γ-secretase proteases does not produce Aβ peptides. Consequently, α-secretases are neuroprotective, and their mechanism of APP processing is often referred to as the 'non-amyloidogenic pathway' [23]. Under pathological conditions, Aβ plaques arise primarily from abnormal processing of the APP, leading to the accumulation of toxic Aβ oligomers that impair synaptic transmission long before substantial neuronal loss occurs [25]. In this 'amyloidogenic pathway' [26], the β-secretase (BACE1) and γ-secretase proteases cleave APP at the N-terminal and C-terminal sites of Aβ region, generating small Aβ peptides that are soluble, secreted fragment known as $sAPP_\beta$ [25,27]. These Aβ peptides exist predominantly in two major isoforms, $Aβ_{40}$ (containing 40 amino acids) and the more aggregation-prone $Aβ_{42}$ (containing 42 amino acids). According to the prevailing hypothesis the overproduction and aggregation of $Aβ_{42}$ constitute the primary event in the pathogenesis of AD, possibly due to the abnormal processing of APP or clearance of Aβ [28].

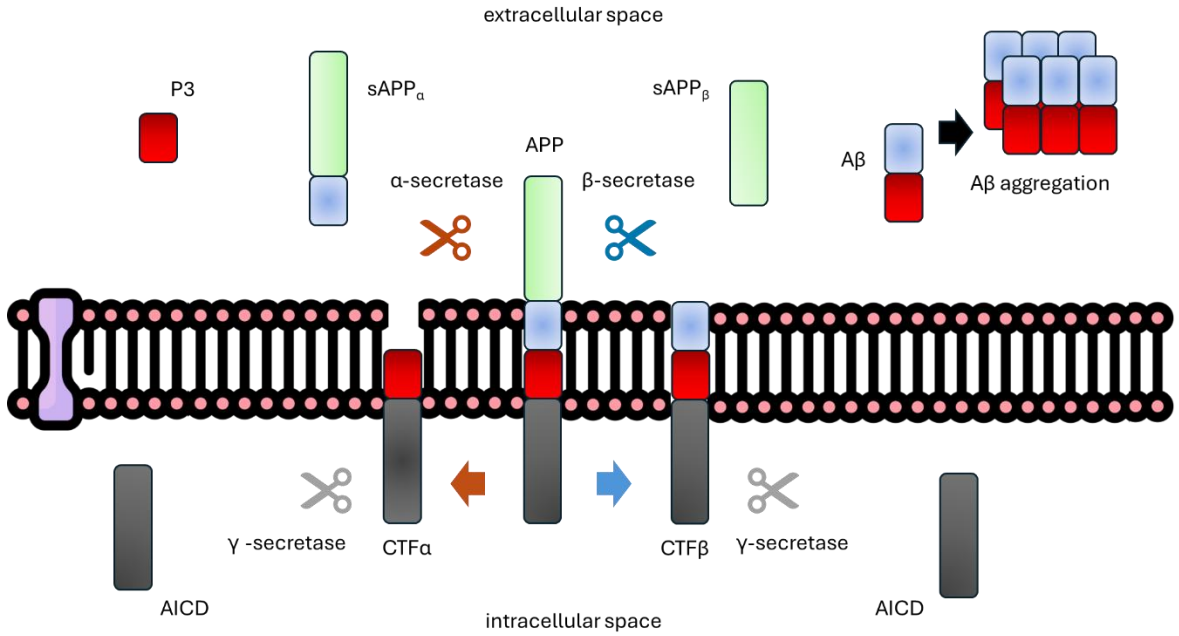

Figure 1. Physiological and pathogenic pathway in processing of APP. (AICD: amyloid precursor protein intracellular domain, CTFα: C-terminal fragment, CTFβ: C-terminal fragment, sAAP: soluble APP, P3: usually Aβ17-40/42 peptide) [23].

However, AD is a multifactorial condition; its development generally involves both genetic and environmental risk factors. Sequence alterations caused by genetic mutations can modify not only protein function but also its physiochemical properties. The well-known genetic factor in AD pathogenesis involves mutations within the APP gene particularly those altering the Aβ sequence region. Common mutations such as the *Swedish* (K670L and M671L double missense) mutation which increase the total amount of Aβ [29], while others such as the *Iberian* (I716F) mutation cause a massive increase in the Aβ42/Aβ40 ratio leading to very early-onset of AD [30]. The *Dutch* (E693Q) mutation leads to severe vascular amyloid deposition causes cerebral amyloid angiopathy [31] while *Arcitic* (E693G) mutation enhances protofibril formation and sources unique “ring-like” shaped plaques [32]. The *Italian* mutation (A673V) produces increasing aggregation propensity [33]. Not only aggregation-promoting mutations are known but also neuroprotective mutation located near to the *Italian* site but with opposite effect: which is the *Icelandic* mutation (A673T) causing reduced BACE1 cleavage efficiency with decreasing the neuroinflammation [34].

1.4. *Investigations of HLA-A epitope presentation*

Although, the third major topic appears less directly related to the CNS, it is closely connected to the immune system function and thus to processes that may indirectly influence

CNS homeostasis and pathology. The immune system detects foreign antigens by presenting peptide-Major Histocompatibility Class – I. (pMHC-I) complexes / Human Leukocyte Antigen (HLA-A) on the cell surface to T-cells. During this process, intracellular self and non-self proteins are cleaved into short fragments, predominantly 9-mer epitopes, which bind to MHC-I. molecules for T-cell recognition [35]. As a result of recognition mechanism, highly similar epitope sequences, differing only by a single amino acid, may not be recognized by the same T-cell receptor. By contrast, epitopes with multiple differences in sequences but with similar physiochemical properties can produce similar immune response [36]. The immune system maintains overall balance and its disfunction can significantly impact neuronal function. As a representative example, Aβ oligomers have neurotoxic effects, as they strongly activate microglial cells, which respond by releasing proinflammatory cytokines (e.g., IFN-γ, TNF-α, IL-1β) [37–39]. These cytokines upregulate MHC-I expression on neurons, which is normally tightly controlled [37]. It states functionally results in neurons becoming quasi-antigen-presenting cells which make them targets for immune-mediated destruction causing neuronal death [40].

So, it can be seen that neurodegenerative diseases cannot attributed to a single protein metabolism problem, genetic mutation or inflammation. Instead, the diseases arise as a result of combined effects of multiple interacting molecular, cellular and systemic abnormalities.

1.5. *Drug delivery and kynurenic acid*

An important factor in the design of neuroprotective compounds is their ability to cross the blood–brain barrier (BBB). In recent years, numerous therapeutic strategies have focused on improving the penetration of drug candidates across the BBB. Molecules may cross the BBB via passive diffusion, which depends on molecular size and physicochemical properties, or through facilitated or receptor-mediated transport pathways. Carrier proteins, like human serum albumin (HSA), interact with receptors expressed on brain endothelial cells, including gp60 and SPARC [41], which mediate albumin transcytosis via vesicular transport. Its capability to bind e.g. Aβ monomer and its suitability to drug formulations made the HSA a possible therapeutic strategy among drug delivery methods [42]. Structural and functional variations of HSA have been reported in several neurological disorders, further underscoring its relevance to AD pathology [43]. One of that well-known endogen molecule is synthetised by kynurenine pathway which is the principal route of tryptophan degradation cause neuroactive metabolites production [44]. Among these, kynurenic acid (KYNA) functions as an endogenous antagonist

of several neurotransmitter receptors, including ionotropic glutamate receptors such as the N-methyl-D-aspartate receptor [45,46]. Current therapeutic strategies aim to modulate the kynurenine pathway to selectively enhance KYNA levels or develop synthetic analogues that preserve neuroprotective properties while minimizing potential cognitive side effects [47–49]. Structural modification of KYNA is also being explored to improve its BBB permeability while maintaining its beneficial biological activity [47,49]. Therefore, the analysis of the transport of new KYNA derivatives across the BBB remains an important open question. This is not only due to the novelty of these compounds, but also because comparing theoretical and experimental results can support the appropriate selection of in silico methods in future studies where HSA is considered a potential transporter.

2. Aim of the study

The overall aim of this dissertation was to investigate the interactions of biologically relevant small molecules and peptides with protein targets using traditional molecular simulation and more recent ML approaches. The study focused on two complementary directions: first, the theoretical characterization of small-molecule binding to selected protein targets with different biological functions, and second, the development of ML models for predicting biologically relevant peptide properties.

Within the molecular modelling part of the work, I aimed to determine whether the experimentally observed binding preferences of modified sterane molecules toward AKR1C1, AKR1C2, and tubulin protein could be explained at the molecular level. To address this question, molecular docking, molecular dynamics (MD) simulations, and free-energy-based methods, including Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) and Umbrella Sampling (US), were applied. These approaches were used to characterize binding modes, key intermolecular interactions, and relative binding preference order.

A further objective was to investigate how structural modifications of KYNA derivatives influence their binding to HSA and bovine serum albumin (BSA). Since BSA represents a more affordable and widely used experimental model of serum albumin, the comparison between HSA and BSA was used to assess whether ligand-binding trends are conserved between the two proteins. The binding behaviour of the KYNA derivatives was evaluated using both theoretical and experimental approaches.

Within the ML part of the dissertation, I aimed to examine which level of peptide representations and their consensus are sufficient for predicting immune responses associated with the HLA-A*02:01 allele. For this purpose, different peptide encoding strategies were compared using FFNN, RF, SVM and LR models.

Finally, I intended to test whether nucleation score values of Aβ mutations can be predicted from different sequence representations. In this part of the work, four complementary sequence encoding strategies: position-weighted BLOSUM62 (pw.BLOSUM62), Z-scale descriptors, amino acid composition (AAC), and embeddings generated by the ESM-2 protein language model were used to encode Aβ variants and FFNN, RF and SVM models were applied to predict experimentally determined nucleation scores.

3. Material Methods

3.1. *Molecular mechanics level investigations*

Molecular docking is one of the widely used and computationally efficient MM-level methods for exploring molecular interactions. Docking algorithms aim to predict preferred binding orientation of a flexible, small molecule (e.g. ligand) within the binding site of a larger receptor, thereby forming stable protein-ligand (PL) complex. The resulting pose can be evaluated using scoring functions, which estimate binding free energy, interaction strength and complex stability [50]. Scoring functions are commonly classified as forcefield (ff)-based, empirical or knowledge-based [50]. The forcefield-based scoring functions, such as those used in AutoDock [51], GoldScore [52] are derived from MM terms, including van der Waals and electrostatic interactions. Empirical scoring functions, such as ChemScore [53] or Schrödinger Glide SP, XP [52,54] are parametrized against experimental binding affinities. Knowledge-based scoring functions, including PMF [55] and DrugScore [56], are derived from statistical analyses of observed pairwise atom-atom distribution. More recently, ML and deep-learning-based docking approaches have also emerged. These methods use AI-driven models to predict ligand binding poses and, in some cases, complex structures directly. For example, AlphaFold3 can model entire biomolecular complexes, including ligand-target assemblies [57], whereas DiffDock predicts ligand binding poses without prior knowledge of the binding site [58], while the DeepDock is another deep-learning approach designed for rapid pose generation [59]. Overall, the molecular docking establishes an initial hypothesis about potential binding properties of the ligand. Even through rapid conformational sampling the receptor has limited flexibility which may reduce the ability to capture subtle conformational changes.

Meanwhile, to overcome these limitations, MD simulations are employed to describe the temporal evolution of PL complexes within a classical mechanics framework. MD models the motions and interactions of atoms and molecules by numerically solving Newton's equations of motion. As in other MM simulations, atoms are treated as charged point-like particles whose trajectories evolve in time according to forces derived from a potential energy function (i.e., the ff). The applied ff is a mathematical model that defines the potential energy between atoms, the forces on atoms, molecules using parameters like bond lengths, partial charges and spring constants. As a result, it enables the investigation of conformational changes, binding events and thermodynamic properties of molecular systems.

Furthermore, we can quantify the time-dependent stability of PL complexes by calculating their binding free energy (ΔG) from MD snapshots using MM/GBSA [60]. This incorporates the energetic properties in implicit solvent model of the ligand, receptor and PL complex structures as well as energy differences relating to deformation and binding modes (equation (1) and (2)).

$$\Delta G = G_{complex} - G_{protein} - G_{ligand} \quad (1)$$

$$\Delta G = \Delta H - T\Delta S = \Delta E_{MM} + \Delta G_{polar} + \Delta G_{nonpolar} - T\Delta S \quad (2)$$

The G_{polar} calculates the polar solvation energy in implicit water model while $G_{nonpolar}$ based on the nonpolar solvation energy which is coming from the surface area. The E_{MM} is computed from the sum of the bonding energy, angle, dihedral, van der Waals and electrostatic energy and the entropic part estimates changes in conformational freedom upon binding (typically calculated via normal mode analysis on MD snapshots, but often overlooked). These free energies calculated on each species per snapshots.

Another popular technique to determine the free energy profile of molecular system is the US method [61,62] (see Figure 2.). The technique is dividing a trajectory the dissociation of ligand from the binding site along a predefined reaction coordinate (eg. distance of ligand from the binding site) into several segments (“windows”). In each window, harmonic constrain is applied to force the system to sample specific conformations, thereby making it possible to investigates states with higher energy barrier that would be rarely accessible. The histograms from each window are then combined using the Weighted Histogram Analysis method (WHAM) [63], which is the potential mean force (PMF) [55,62] curve and the ΔG can be determined. While MM/GBSA is quick method to comparison of binding stability, the US provides more detailed information on binding and dissociation mechanism.

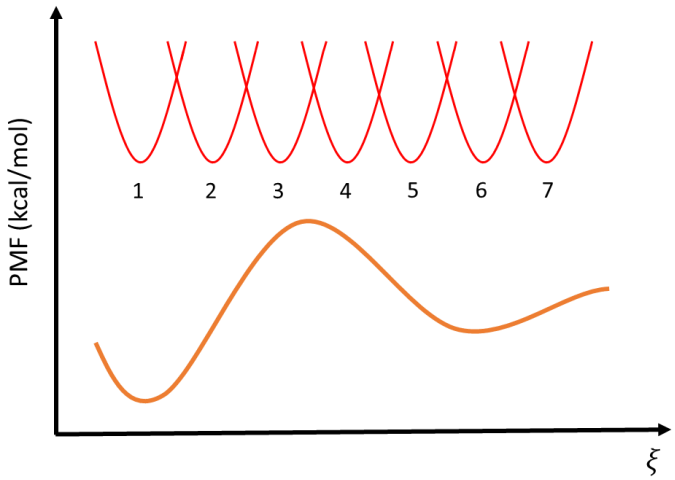

Figure 2. Schematic representation of US method. The ξ represent a reaction coordinate (can be two molecule distance or dihedral angle, etc.). The PMF is a free-energy profile that describes how the average force changes along a specific pathway or coordinate in a system *[55]*.

3.2. *Machine Learning techniques*

The applied classical ML algorithms: FFNN [7,8], SVM [9], SVR [11], RF [13] and LR [12] represent fundamentally different techniques (schematic figure of them can be found in Figure 3.).

FFNN is organized as fully connected layers arranged in a strict linear order, the input layer, one or more hidden layers and an output layer. Each neuron in a layer is connected to every neuron in the next layer and the information flows only in one direction without feedback or cycles. The network operates on fixed-size feature vectors, meaning that all inputs must be represented in the same dimensional space. The learning method occurs by adjusting numerical weights and biases associated with each connection using gradient-based optimization. The structure of an FFNN does not explicitly encode relationships between individual input features; instead, it relies on dense connections to implicitly learn feature interactions [7,8].

SVM is a type of geometrical ML method which works by finding the optimal hyperplane (decision margin) that maximizes the margin between the closest data points of different classes (support vectors) [9]. It is a deterministic model that focuses on strict geometric separation rather than calculating class probabilities. While SVR is a supervised ML technique that predicts continuous values by finding a line or curve within a set error margin [11].

RF has a hierarchical tree-based decision structure rather than layered neural units. It consists of an ensemble of independent decision trees, where each tree is built by recursively splitting the data based on feature thresholds. Structurally, each decision tree is a separating graph with internal decision nodes and terminal leaf nodes, but unlike neural networks, there

are no continuous weights or neuron layers. Each tree is trained on a random subset of samples and features, and the final prediction is obtained by collecting outputs from all trees (e.g., voting or averaging). Importantly, the trees do not interact with each other during training, making RF a collection of parallel models rather than a single interconnected structure [13].

In contrast, the LR is a probabilistic model that estimates the likelihood of a data point belonging a specific class using the sigmoid function. It performs the best on linearly separable data and provides highly interpretable coefficients, showing exactly how each feature impacts the prediction [12].

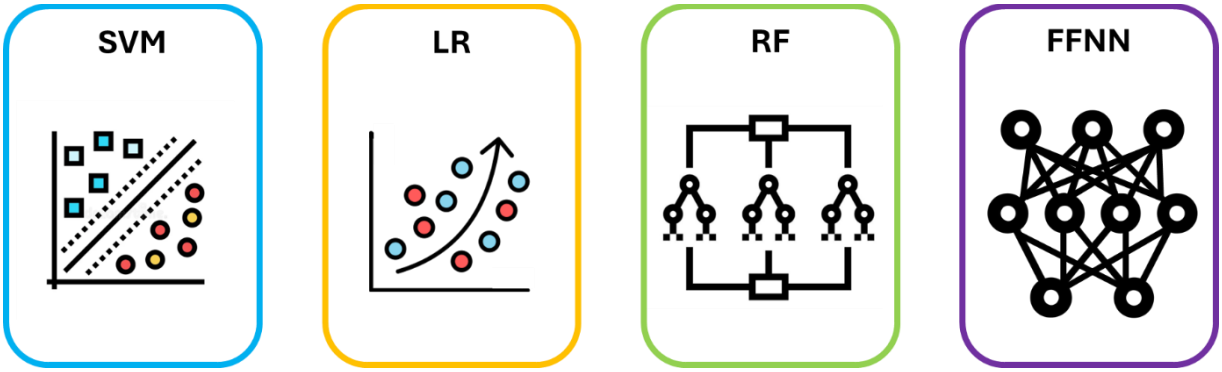

Figure 3. Schematic diagrams of commonly used ML algorithm (SVM: support vector machine, LR: logistic regression, RF: random forest, FFNN: feed-forward neural network).

Each ML technique balances complexity against interpretability. The LR is highly efficient and provides clear probabilities for linear data, but it fails to capture complex, non-linear relationships. The SVM excels in high-dimensional spaces and resist outliers by maximizing margins, yet they become computationally expensive on large datasets and lack native probability outputs. As opposed to the RF easily handles non-linear structures, require minimal data pre-processing and resist overfitting, though they act as complex "black boxes" that demand significant memory. Finally, FFNN offers unmatched flexibility for learning intricate, massive data patterns, but they require vast amounts of data, high computational power, and extensive hyperparameter tuning [64].

3.2.1. *Peptide representations*

We have several possibilities to represents a molecule at different levels. The numerical representation of the 9-mer epitopes with simple amino acid sequences is based on a one-to-one mapping as shown in Figure 4A, while the theoretical background of the other representations, such as similarity matrices and structural descriptors, are summarized in Figure

4B-E. Thus, five different input datasets (representations) were generated for the concerned peptide sequences, namely *aa.seq., p.w.BLOSUM62 and PAM250* [65–67], as well as traditional molecular descriptors [68] based on (*mol.desc*.) or Schrödinger's pharmacophore features (*pharmFP*) [69,70]. The applied similarity matrices are essential tools for quantifying sequences based on amino acid similarities. Molecular descriptors are direct or derived properties of a molecule that can be understood by supposing that molecules with similar descriptors have similar properties. Pharmacophore properties (H-bond donor/acceptor relation, charge, aromaticity, and hydrophobicity) include all the structural or chemical properties of a compound that is responsible for ligand binding according to the Canvas module in the Schrödinger suite [70]. For the molecular descriptors, we considered the complexes generated by the AlphaFold3 server [71] and used a set of 0D–3D descriptors of the epitopes produced by the Dragon software [24]. The generated molecular descriptors were unique to the peptides and were based on their 0D, 1D, 2D, and 3D features, resulting in approximately 3,000 descriptors. Pharmacophore fingerprints were generated using Canvas, a part of the Schrödinger Suite [25]. These pharmacophore fingerprints are also unique and are based on the three-point model used by Canvas on the Schrödinger platform, yielding approximately 10,000 data points.

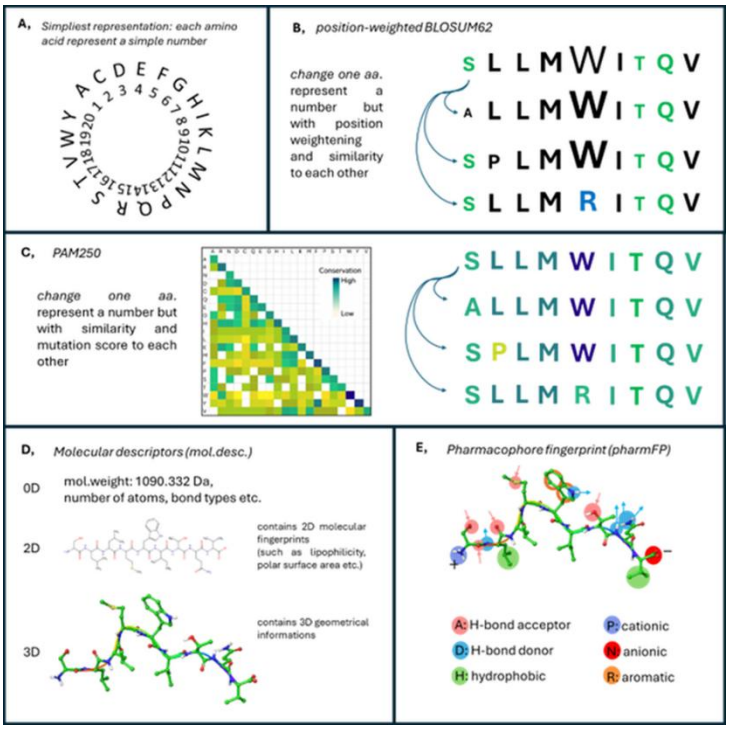

Figure 4. Numerical representation types of 9-mer epitopes. (A) Simple amino acid sequence (**aa.seq.**); (B) similarity-matrix-based encoding using position-weighted BLOSUM62 (**p.w.BLOSUM62**); (C) mutation matrix-based encoding using point accepted mutation matrix (**PAM250**); (D) multiple molecular descriptors based on physicochemical properties such as chemical formula, molecular weight, bond angles, surface characteristics, etc. (**mol.desc.**); (E) 3D pharmacophore fingerprint based on annotated secondary interactions of the ligand (**pharmFP**). The epitope–protein complex 3D structures were generated by AlphaFold3 [71].

The numerical representation of the Aβ peptides with simple amino acid sequences based on a amino acid composition (AAC) represents each sequence by the relative frequency of the 20 standard amino acids and provides a simple, low-dimensional baseline. AAC is widely implemented in peptide and protein feature-extraction framework. Its major disadvantage is the complete loss of residue order and position [72].

The Z-scale representation describes each amino acid through five continuous principal-property variables derived from a larger collection of physicochemical measurements. These components summarize properties related to hydrophobicity, steric characteristics, polarity and electronic behaviour [73,74].

Finally, the ESM-2 embeddings provide a data-driven alternative to predefined descriptors. ESM-2 is a transformer-based protein language model trained on large collections of natural protein sequences. Its latent representations capture contextual sequence dependencies and contain information relevant to protein structure and function. Residue-level embeddings preserve positional context, whereas mean or special-token pooling generates a fixed-length sequence representation. ESM-2 may therefore detect complex and non-linear

relationships between mutations and the surrounding sequence that are inaccessible to composition-based encodings [75–77].

3.2.2. *Evaluation metrics*

To qualify the outcomes, binary classification with the most frequently applied evaluation metrics was presented. The simplest but most common is *accuracy* which expresses the percentage of correctly classified instances among the members of the full set. The *sensitivity* value quantifies truly positive instances, whereas *specificity* expresses truly negative ones. In the literature, *sensitivity* is often referred to as *recall*. The applied formulas for these quantities are as follows [78]:

$$\textit{Accuracy} = \text{ACC.} = \frac{TP + TN}{TP + TN + FP + FN} \quad \in [0,1] \tag{3}$$

$$\textit{Sensitivity} = \text{Sen.} = \frac{TP}{TP + FN} \quad \in [0,1] \tag{4}$$

$$\textit{Specificity} = \text{Spe.} = \frac{TN}{TN + FP} \quad \in [0,1] \tag{5}$$

True positive (TP), true negative (TN), false positive (FP), and false negative (FN) refer to the number of predictions with these designations. It should be noted that *sensitivity* and *specificity* reveal more about the model than *accuracy*, especially when the number of real positive and negative instances is unbalanced. Furthermore, we calculated additional evaluation metrics, such as *balanced accuracy* that depend on all correctly classified values of the confusion matrix [78]. It gives equal weight to the performance in all classes, provides a more reliable assessment of performance, and provides a balanced picture.

$$\textit{Balanced accuracy} = \text{bACC.} = \frac{Sen. + Spe.}{2} \quad \in [0,1] \tag{6}$$

The *F1-score* is an important evaluation metric that measures the harmonic mean of *precision* (Pre.) and *sensitivity*, defined as [79]

$$F1 - score = \frac{2 * Pre. * Sen.}{Pre. + Sen.} \quad \in [0,1], \text{ where Pre.} = \frac{TP}{TP + FP} \quad \in [0,1] \tag{7}$$

while *Cohen's kappa* (κ), defined as [80]

$$\kappa = \frac{ACC. - p_e}{1. - p_e} \in [-\infty, 1] \qquad \text{where } p_e = \frac{(TP+FN)(TP+FP)+(TN+FP)(TN+FN)}{(TP+TN+FP+FN)^2} \tag{8}$$

The latter compares the performance of the binary classifier with that of the random accuracy p_e. κ is a quantity, with a probability-based correction that helps filter biases that arise when a model seems to perform well due to an unbalanced distribution or according to accuracy, but it is no better than a random guess. It can be applied to measure the agreement between predicted and real classes.

Another typical performance metric is the *receiver operating characteristic curve* (ROC) which is a visual representation of the performance [81]. This shows the model's ability to distinguish between positive and negative classes across all possible classification thresholds. The perfect model shows a square with unit sides and has an *area under the curve* (AUC) of 1.0. This means that it has 100% probability of a randomly selected positive case that will be correctly categorized by the model at a continuously varying threshold value.

The predictive performance of the regression models was evaluated using four complementary metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). Each metric emphasizes different aspects of predictive performance, and they were considered jointly throughout this study. MAE reflects the average prediction accuracy, MSE and RMSE emphasize the presence of large prediction errors, whereas R^2 evaluates how well the regression model captures the variability of the experimental nucleation scores. Reporting all four metrics provides a balanced and comprehensive evaluation of model performance and facilitates comparison with previous machine-learning studies on protein and peptide property prediction. Together, these metrics provide a comprehensive assessment of prediction accuracy by quantifying both the magnitude of prediction errors and the proportion of variance explained by the model.

3.3. *Experimental methods*

3.3.1. *Isothermal titration calorimetry*

Isothermal titration calorimetry [82] (ITC) was used to experimentally determine the thermodynamic parameters of the interactions between SA-s and the ligands. Thermometric titration experiments were performed at 303 K with a computer-controlled VP-ITC power-compensation microcalorimeter (MicroCal). Compounds were dissolved in pH 7.4 20mM phosphate buffer containing 150mM NaCl. The 1400-μL sample cell of the calorimeter was filled with the HAS or BSA solution ($C_{albumin}$=100 μM) and this protein solution was titrated under constant stirring with (KYNA or SZR compounds) solutions. Concentration of the applied ligand solutions were in the range of 5 – 10 mM. The heat released or absorbed during each injection was recorded. Titrations were performed with KYNA and with derivatives SZR72, SZR81, SZR759, SZR104 and SZR105 using 5-10 μL injections with 10-30 sec spacing. Raw data integration was performed using SVD analysis implemented in NITPIC [83], The resulting isotherms were fitted using SEDPHAT [84], for the derivatives using nonlinear regression to obtain the lgKa, equilibrium dissociation constant (K_d) and enthalpy (ΔH) including baseline and incompetent protein fraction as adjustable parameters. Figures were prepared by GUSSI [85].

3.3.2. *Saturation-transfer-difference – NMR*

The saturation-transfer-difference (STD) – NMR detects binding by transferring the selectively magnetic saturation (excitation) of the receptor (protein) to the ligand temporarily bound to it through spatial proximity. In the resulting difference spectrum, therefore only those parts of the ligands actually involved in the binding will produce a signal, which have come into direct physical contact with the protein. Using that additional experimental technique, we validated the computational and ITC results by STD-NMR [86]. Together, the PMF simulations, ITC measurements and STD-NMR data provide a coherent and complementary overview of the binding behaviour of the KYNA and its analogue series, enabling reliable comparison of affinities across human and bovine SAs and revealing structure-dependent trends in interaction strength. NMR spectra were acquired at 300 K with a Bruker Avance 600 MHz spectrometer equipped with a 5 mm z-gradient CP-TCI triple-resonance cryoprobe. Compounds were dissolved in phosphate buffer (20 mM, pH 7.4, 90% H2O, 10% D2O) containing 150 mM NaCl and 3 mM NaN3. Samples contained 500 μM ligand (KYNA or SZR compound) and 25 μM

HSA or BSA, thus the ligand excess was 20-fold. Selective saturation of the protein was achieved using 40 equally spaced 50 ms Gaussian-shaped pulses, resulting in a total saturation time of 2 s. The on- and off-resonance saturation frequencies were set to 0.84 ppm and 40 ppm, respectively. Spectra were collected with 512 scans. STD amplification factors were calculated as:

$$STD - AF = \frac{I_0 - I_{sat}}{I_0} * ligand\ excess$$

where I_0 is the integral of the proton signal in the unsaturated spectrum, and I_{sat} is the integral of the corresponding proton after saturation [87].

4. Results and Discussions

4.1. *Modified steroids binding to tubulin protein*

The development of new anticancer compounds with improved tolerability and safety profiles are essential because many therapeutic agents are associated with undesirable side effects and the progress of resistance can further limit treatment effectiveness [88]. In this context, the steroid scaffold represents a promising starting point for the development of new anticancer agents against cervical carcinoma, not only because of its favourable structural properties, but also because steroid derivatives often display complex biological activities beyond their classical hormonal effects [89]. A further attractive feature of the steroid backbone is that even minor structural modifications can lead to marked changes in the pharmacological profile [90]. For example, certain estrone sulfamates have shown anticancer activity against triple-negative breast cancer cells as well [91].

The first research line during my PhD studies was the application of US and MM/GBSA methods to explain at molecular level the binding preference order of three selected modified steroid compounds (3-O-sulfamoyl-13α-estrone derivatives: 13AES1-3) to tubulin protein, having promising biological activity (IC_{50} of 13AES3 > 13AES2 > 13AES1) in HeLa and SiHa cell lines [92].

According to the US calculations (ff14SB ff for protein, GAFF2 ff for the ligands were applied in 0.15M physiological salt concentration and TIP3P explicit water model while the simulation pulling phase was 1ns long and 200ps equilibration was performed in each window), ligand 13AES3 had the strongest interaction with the binding sites (−14 kcal/mol), whereas the other two compounds (13AES1 and 13AES2) were characterized by quite similar binding free energy values (−5.2 kcal/mol and −6.5 kcal/mol, respectively) in Figure 5a. The MM/GBSA calculations supported a very similar conclusion: 13AES3, 13AES2, and 13AES1 showed −69.38 ± 2.14 kcal/mol, −50.85 ± 2.85 kcal/mol, and −42.76 ± 2.35 kcal/mol ΔG values, respectively in Figure 5b. These findings confirmed the US results and allowed further analysis of ligand–protein interaction as each MM/GBSA calculations (OPLS4 ff and SPC water model) were based on 500 ns long explicit water simulations.

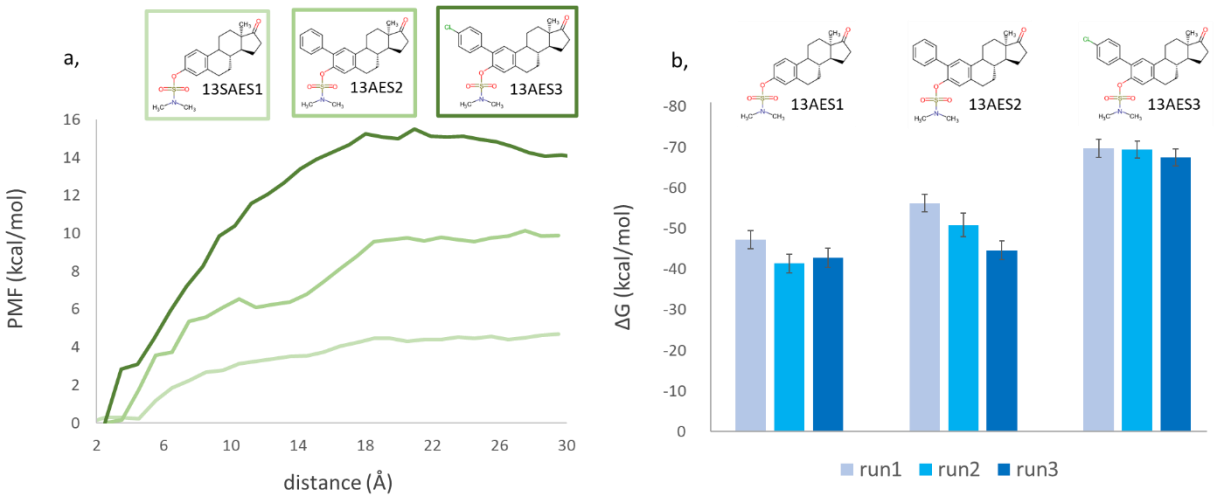

Figure 5. a, PMF curves of the modified steroid compounds bound to tubulin (PDB: 5SYF) protein using US method (AMBER [61]) **b,** ΔG values of PL complexes within the taxene binding site [16] across three independent MD runs (Schrödinger Desmond package [93]) in each molecule.

To provide a computational explanation for the experimental results, the present findings are consistent with the experimental observations, as both 13AES1 and 13AES2 exhibited less pronounced biological activity than 13AES3. To better understand the molecular background of the binding preference order, a ligand interaction diagram (LID) analysis [94] was performed. This type of analysis visualises the key residue-ligand contacts and helps identify which molecular interactions may contribute to binding stability and affinity. According to the LID results, ligand 13AES3 formed two stable interactions with residues THR276 and ARG369, which may play a main role in its higher ΔG value (Figure 6.). Interestingly, although the ligands with weaker binding showed a larger number of interacting residues, the interaction stability, quantified by the fraction of persistent contacts, was the highest for 13AES3.

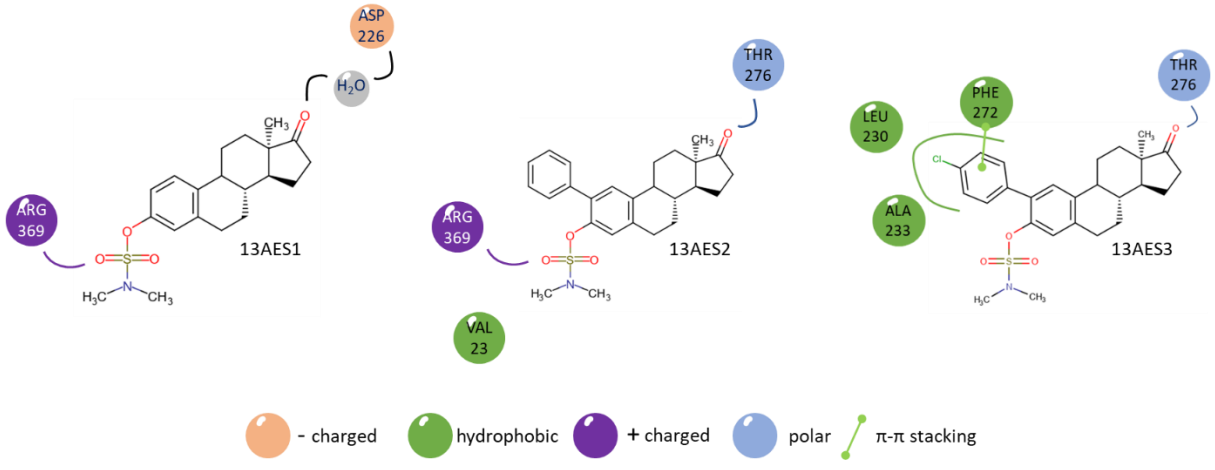

Figure 6. Ligand interaction diagram of modified steroids binding to tubulin protein to taxene binding site.

Computational simulation results suggest that the chlorophenyl ring of 13AES3 at C-2 helps to stabilize the binding pose of the sterane skeleton, explaining the stronger interaction with β-tubulin. Namely, increased hydrophobic interaction between the chlorophenyl ring of ligand 13AES3 and the phenyl ring of PHE272 residue was revealed. It is important to note that a hydrophobic interaction between ligand 13AES2 and the phenyl ring of PHE272 was also established. However, this hydrophobic interaction was less significant and was not exclusively formed between the phenyl rings. The phenyl ring of PHE272 partially interacted with the aromatic ring A of the estrane skeleton. Consequently, the sterane scaffold was less oriented for 13AES2, and thus it could not establish a stable link to the protein through the 17-keto and 3-*O*-sulfamoyl groups. Furthermore, visual inspection of the trajectories suggested that in the presence of a halogen atom, the chlorophenyl ring could penetrate deeper into the binding pocket, which considerably restricted the ligand's mobility. Thus, it can be concluded that the presence of an unsubstituted phenyl ring (13AES2) promotes the hydrophobic connection, but it is insufficient; increased hydrophobicity must be achieved, which could be realized by the introduction of a halogen atom.

4.2. *Modified steroids binding to AKRs*

Well known, that the aldo-keto reductase isoforms (namely the AKR1C1-3) are key enzymes in steroid metabolism and validated targets in hormone-dependent cancers. The literature reveals several structurally different inhibitors of AKR1C1–3 enzymes, albeit selective inhibition of the three isoforms is rather challenging [95]. Especially that of AKR1C1 and AKR1C2 is difficult, as these enzymes differ by only a single amino acid residue in the active site.

Firstly, molecular docking calculations than interaction analysis were performed to gain a molecular-level insight into the possible binding mode of two modified steroids in the active sites of AKR1C1 and AKR1C2 enzymes [96]. To evaluate the reliability of the predicted binding modes, docking scores (Glide scores) were compared to the ΔG values derived from the experimental inhibition data.

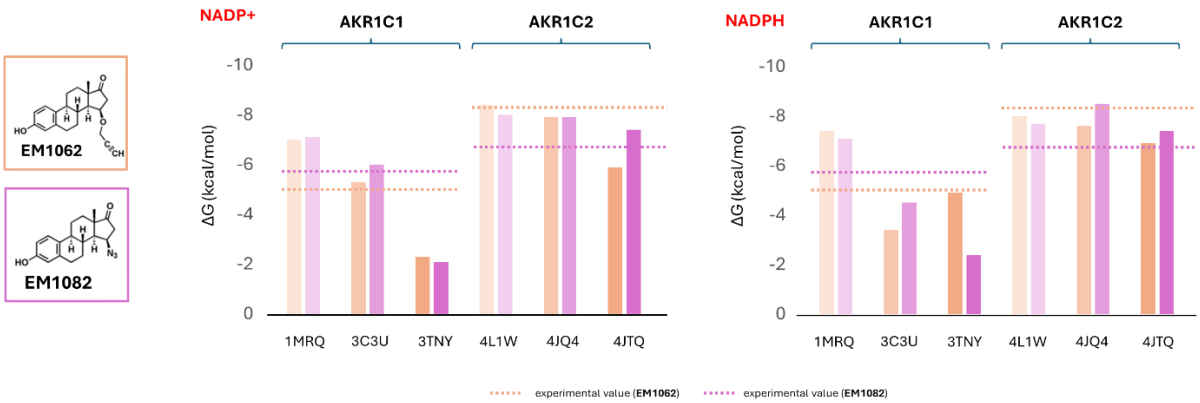

Figure 7. Docking scores of modified steroids into AKR1C1 and AKR1C2 binding sites using different (AKR1C1 PDB codes: 1RMQ, 3C3U, 3TNY and AKR1C2 PDB codes: 4L1W, 4JQ4, 4JTQ) crystall structures. The dotted line represent the experimental binding free energies (calculated from inhibition %) [96].

As Figure 7. shows, similar tendencies can be observed between theoretical and experimental data for both receptors. Moreover, the re-docking pose of the crystal steroid ligand closely matched the experimental geometry, further supporting the reliability of the applied docking protocol. More precisely, we have in case of AKR1C1: 0.248 Å and in case of AKR1C2: 0.476 Å RMSD values. To ensure that the docking protocol is independent of the selected PDB structure and the protonation state of the cofactor (NADP+ or NADPH), docking calculations were performed using multiple crystal structures (AKR1C1: 1RMQ, 3C3U, 3TNY; AKR1C2: 4L1W, 4JQ4, 4JTQ), with both NADP+ and NADPH considered in all cases. A comparative study involving AKR1C1 and AKR1C2 was performed to gain insight into the markedly different inhibitory activities of these (EM1062 and EM1082) compounds toward the two structurally highly similar enzymes.

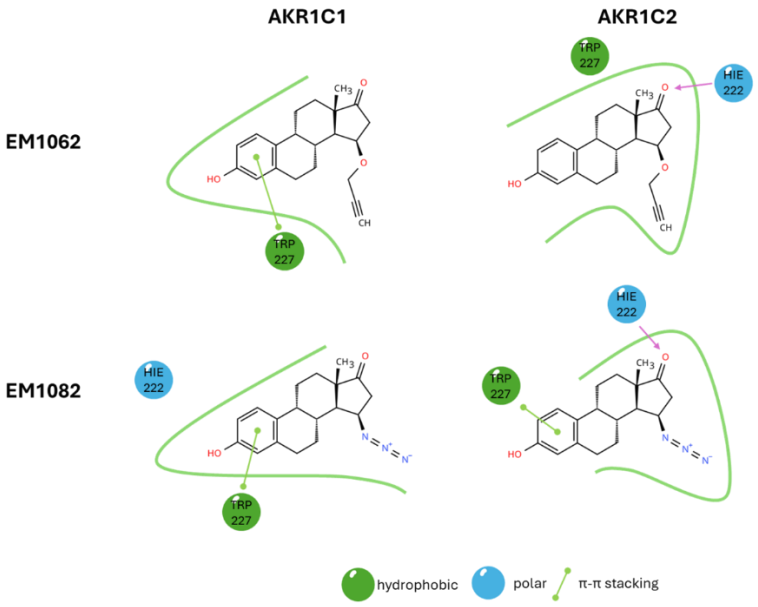

Figure 8. LID of modified steroids in the AKR1C1 and AKR1C2 binding pocket.

Docking results indicate that both ligands exhibit stronger binding affinity for the AKR1C2 isoenzyme compared with AKR1C1 (according to Figure 7.). This trend is consistent with the experimental inhibition data, where both compounds showed higher inhibition against AKR1C2 (inhibition approximately 97–99%) than against AKR1C1 (inhibition approximately 32–63%). Analysis of the binding poses revealed that the principal difference between the ligand–protein complexes arise from distinct ligand orientations. In AKR1C1, both ligands preferentially adopt a binding mode in which the A-ring is oriented toward the NADPH cofactor. In contrast, in AKR1C2, the ligands favor an orientation in which the sterane skeleton positions the D-ring in proximity to NADPH. These findings suggest that subtle structural differences between the two enzymes significantly influence ligand orientation and stability within the active site. In particular, the residue at position 54 (Leu54 in AKR1C1 vs Val54 in AKR1C2) appears to modulate the hydrophobic and steric environment of the binding pocket, thereby contributing to more favorable ligand binding in AKR1C2. Ligand Interaction Diagram (LID) analysis further indicates that this difference may be associated with a hydrogen bond between the ligand keto group and HIE222 in AKR1C2, which compensates for or outweighs the π–π stacking interaction observed between the ligands and TRP227 in AKR1C1 (Figure 8.). Overall, these results explain the observed selectivity of compounds **EM1062** and **EM1082** toward AKR1C2 and highlight the importance of subtle active-site variations in determining isoenzyme selectivity.

In summary, computational simulations of the binding properties of the modified steroids provided a molecular-level interpretation of the experimental results and offered a theoretical augmentation for understanding their biological activity.

4.3. *Computational and experimental investigation of KYNA derivatives binding to HSA and BSA*

Drug efficiency is strongly influenced by several factors, including formulation conditions, solubility, lipophilicity, and pH. According to previous investigations, KYNA favours stronger electrostatic interactions and greater incorporation into the serum albumin nanoparticle carrier under acidic conditions [97]. Building on this line of research, the direct binding of KYNA and selected derivatives (SZR72 [49], SZR104 [48]) to SAs was investigated to improve the pharmacological profile of these compounds while preserving biological activity. Therefore, another aim of my doctoral research activity related to these investigations was to comprehensively characterize the binding properties of KYNA and selected analogues to *Sudlow binding site II* in both HSA and BSA [98] (see Figure 9).

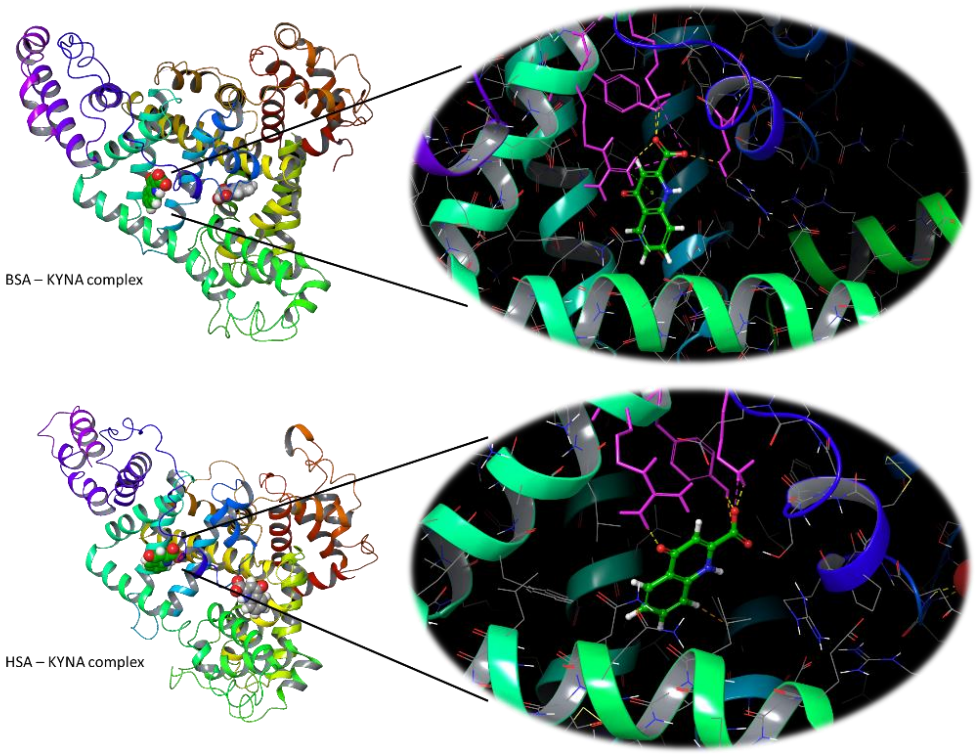

Figure 9. Global and enlarged views of *Sudlow binding site II*. binding pocket in BSA and HSA, illustrating the structural similarity and local differences between two SAs.

As a starting point, I first present the theoretical results from computational simulations in Figure 10, which show the US PMF curves for the selected ligands. The computational simulations were evaluated by umbrella sampling (US) molecular dynamics simulations using the AMBER 2024 version program package [61]. The ff19SB force field (proteins) and GAFF2 (ligands) were used in combination with the TIP3P explicit water model and 0.15 M NaCl to mimic physiological conditions. The systems were equilibrated for 200 ns at 315 K temperature, after which 1000 ns short simulations windows with distance restrain along the reaction coordinate (defined as the distance from *Sudlow binding site II*.) were performed for 100 ns per window, with a harmonic biasing potential applied to ensure adequate sampling. Finally, the WHAM was applied to reconstruct the free energy profile of ligand binding, allowing quantification of binding strength and identification of key interaction regions.

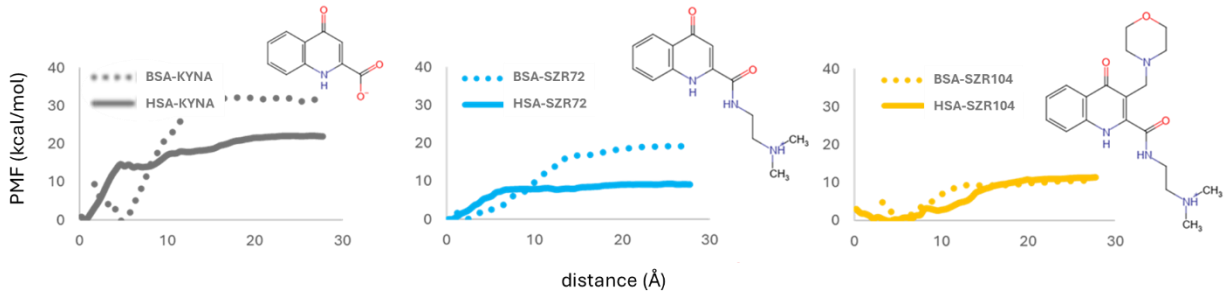

Figure 10. PMF curves of KYNA and its derivatives bound to HSA and BSA proteins.

According to Figure 10, KYNA binds more strongly to BSA than to HSA (larger min/max difference along the curve), which supports the use of BSA in studies involving the direct binding of KYNA molecule to the BSA molecule, which differ from the previously investigated SA nanoparticles. The C-2 modified SZR72 analogue showed a lower binding propensity for HSA than KYNA and only a weak preference for BSA, while their overall binding ability remained relatively low. Considering the morpholine ring containing derivative SZR104, it presented weak binding to both SAs with no or only a weak preference order.

Regarding the character of the US-PMF curves of the parent compound, they show a clear albumin dependent trend, where BSA producing a steadily rising PMF curve that reaches ~32 kcal/mol at 15 Å. It indicates significantly stronger stabilization by BSA, whereas the HSA profile remains shallow with plateaus around ~22 kcal/mol. A similar pattern is observed for the SZR72, but with decreased interaction strength. Considering all the curves for the BSA complexes, KYNA exhibited the strongest interaction, and only the SZR72 compound showed considerable, albeit weak, binding, whereas the SZR104 compound exhibited binding strengths close to 10 kcal/mol, with shallow US-PMF curve profiles. As a conclusion, the PMF results highlight a structure–dependent affinity pattern across the ligand types: the original KYNA with the free carboxyl group showed the strongest interaction, while the SZR72 and SZR104 reveal weaker interactions with both albumins.

It is important to note that the BSA is slightly more negatively charged and hydrophilic protein whereas HSA is generally a more hydrophobic protein. This suggests that positively charged derivatives may be favourable if the steric properties are not considered. The morpholine type derivative can be sterically hindered due to its more space-filling conformations which causes the flattered PMF curves. The *Sudlow site II.* highly conserved between HSA and BSA, with the available comparative data indicating that the amino acid composition of this binding pocket is essentially the same in both proteins.

ITC measurements reveal a moderate binding affinity across the ligand series in serum albumins. In our experiments, the parent compound (KYNA) displays sub-millimolar dissociation constants (BSA **K_d = 158.9 ± 81.4 µM** and to HSA is **K_d = 5.7 ± 6.0 µM**,), indicative of relatively moderate binding interactions (see more in Figure 11.).

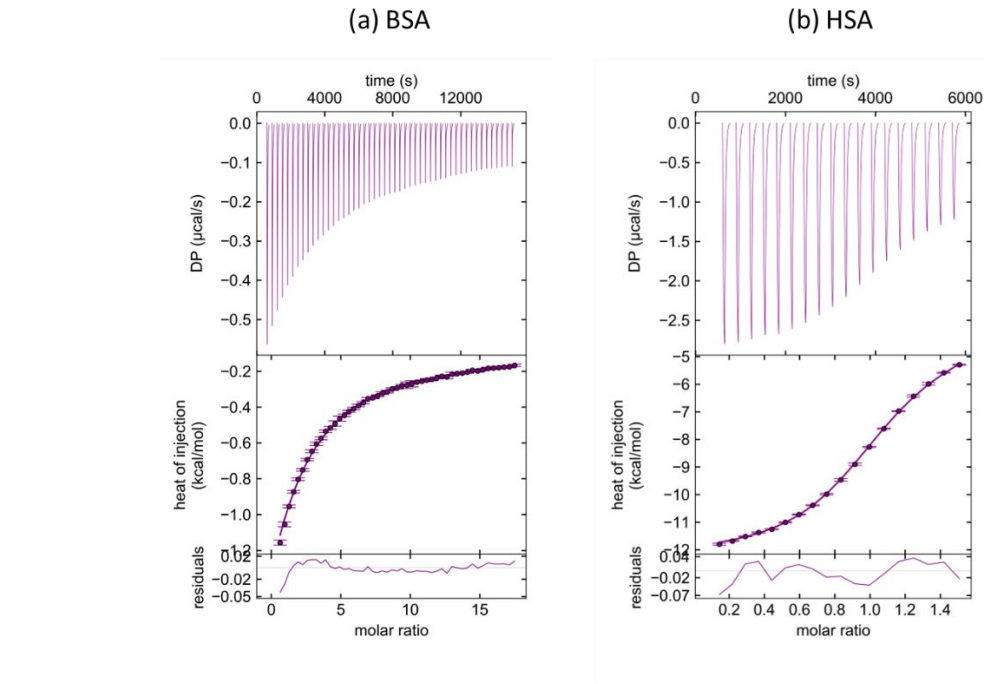

Figure 11. ITC results of KYNA to BSA (a) and HSA (b). The best fit of the models were observed into 1:1 binding model.

The results of the progressive structural modifications show systematic reduction in affinity with K_d values spanning above the detectable range (**1 mM < K_d**). The observed affinity trend likely reflects of altered hydrogen bonding capacity, electrostatic interactions and conformational flexibility introduced by chemical substitutions. However, ITC is most reliable within a specific affinity window (typically low micromolar to mid millimolar) and very weak binders often fall below the detectable heat threshold.

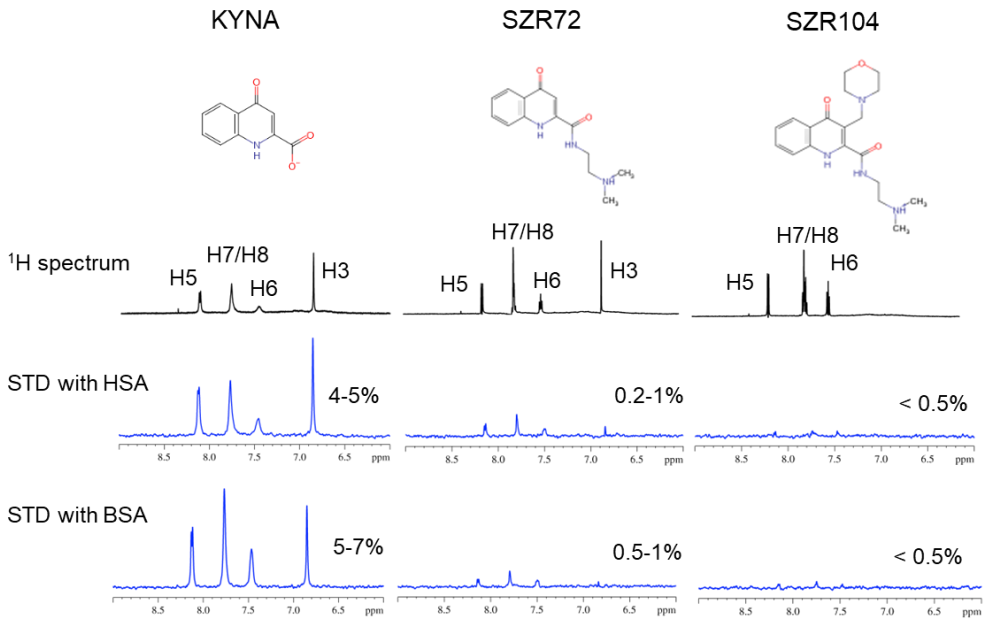

Figure 12. STD-NMR results of the selected compounds.

Based on the STD-NMR experiments, KYNA displays clear and well-defined STD responses in the presence of both albumins (Figure 12.), demonstrating direct binding under the applied experimental conditions (see more in Method section). The binding of KYNA derivative SZR compounds to HSA/BSA was clearly detectable, however, the overall STD amplification factors of the derivatives were markedly lower than those of KYNA (in the range of 4-7% and 0.2-1% for KYNA and for the derivatives, respectively), consistent with weaker binding interactions relative to the parent compound. Importantly, signals of the morpholine ring and the aliphatic protons were not observed in the STD spectra, indicating that these atoms are not in direct contact with the protein. These results suggest that KYNA forms a more stable and instantly detectable complex with SAs, whereas the modified derivatives show altered binding behaviour like characterized by moderate affinity toward BSA and weaker interactions with HSA. It is important to note that STD intensity reflects proximity and exchange dynamics within the ligand–protein complex and therefore provides qualitative rather than strictly quantitative information about binding affinity. Nonetheless, the STD-NMR findings complement the ITC and PMF results, collectively reinforcing the observed structure–affinity trends across the ligand series. In our STD-NMR experiments, KYNA showed the strongest albumin binding, whereas SZR72 displayed weaker but still detectable interaction, and the more structurally modified derivative (SZR104) exhibited only minimal or no detectable binding. These findings suggest that subtle structural modifications can markedly alter carrier affinity and thus potentially influence CNS delivery.

4.4. *Prediction of epitopes immune activation using ML algorithms and different epitope representations*

The fourth research direction of my PhD studies involved the application of ML techniques to various peptide-related questions [99]. The first focused on variations in T-cell activation that may arise from structural differences among interacting peptides, which in turn modulate the molecular surface of epitope–MHC-I (pMHC) complex. Since the sequence and the primary geometry of the MHC-I protein remain constant, it is a trivial assumption that the chemical properties of the epitope are the key points that predominantly determine the formation of the complex and subsequent T-cell recognition. Based on the response of T-cells activated by the formed pMHC complexes, we classified the peptides as active or inactive ones (based on relative activation). A peptide was considered active if it was capable of retaining its specific biological functions, such as TCR binding or T-cell activation. Although these two effects are not independent of each other, the applied experimental biological data typically address only one of them at a time. Thus, diverse representations of epitopes could be explored, ranging from simple amino acid sequences to similarity matrix-derived scores or three-dimensional structural characters. By integrating these distinct layers of molecular information, including peptide sequence-based information (*aa.seq., p.w.BLOSUM62, PAM250*) or molecular descriptors (*mol.desc*.) and pharmacophore features (*pharmFP*), we constructed a comprehensive predictive framework.

In our models training, the 135-epitopes dataset was randomly divided into a 70% training subset and a 30% hold-out (internal) validation subset. Model optimization was performed on the training subset using five-fold cross-validation, and final performance was evaluated on the independent 30% hold-out set (Figure 13). Here, I present the consensus function results based on *similarity matrices* (*p.w.BLOSUM62* + *PAM250*) + *structural descriptors* (*mol.desc.* + *pharmFP*), where consensus function based on the mean of predicted probabilities from these representations). The quality of the methods was represented by five selected performance parameters: ACC, bACC, Cohen's κ, ROC AUC, and F1-score, whose definitions can be found in the Methods section.

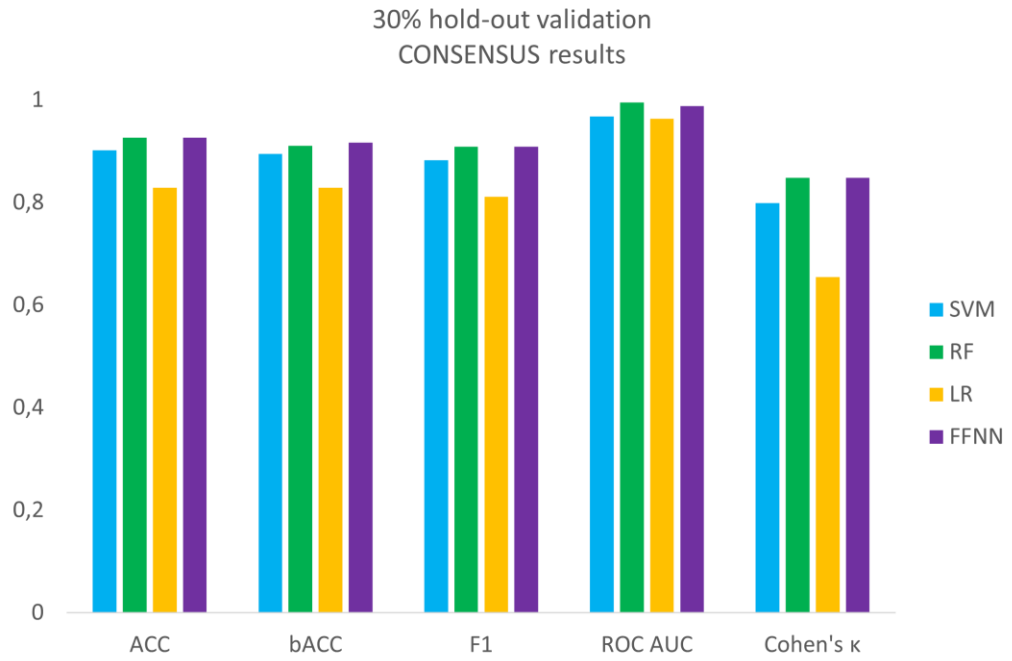

Figure 13. Qualitative attributes of **30% hold-out (internal) validation test**, used different ML models (Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF) and the Feed Forward Neural Network (FFNN)) across consensus function of different epitope representations. The qualitative descriptors are: ACC = accuracy, bACC = balanced accuracy, ROC AUC = receiver operating characteristic area under the curve.

The consensus-based hold-out validation results (see Figure 13.) showed strong overall performance for all four classifiers across the 30% validation set. RF and FFNN achieved the most consistent performance, with ACC, bACC, and F1-score values around 0.90 or higher, indicating robust classification of both active and inactive epitopes. SVM also performed well, particularly in terms of ROC AUC, although its Cohen's κ was slightly lower than that of RF and FFNN. LR showed the weakest overall performance among the four models, especially for F1-score and Cohen's κ, suggesting a reduced ability to capture the more complex relationships present in the combined peptide representations. Overall, the consensus approach produced high ROC AUC values for all classifiers, indicating that the combined input representations retained strong discriminative information. These results support the use of multiple complementary classifiers rather than relying exclusively on a single modelling strategy.

An important aspect of the first results is that although the FFNN and RF method also achieved a good performance comparable during internal cross-validation the main challenge of AI methods remains their applicability to external datasets. Therefore, external validation and comparative analysis were performed using the full set of previous samples (135 epitopes) as training data. During this phase, the same AI methodologies as in the initial training and

validation stages were applied, and classification was conducted on an independent external dataset comprising 85 previously unseen epitopes. Notably, measurement data for the new set of epitopes were provided by the same research group using the same experimental techniques. This consensus results can be found in Figure 14.

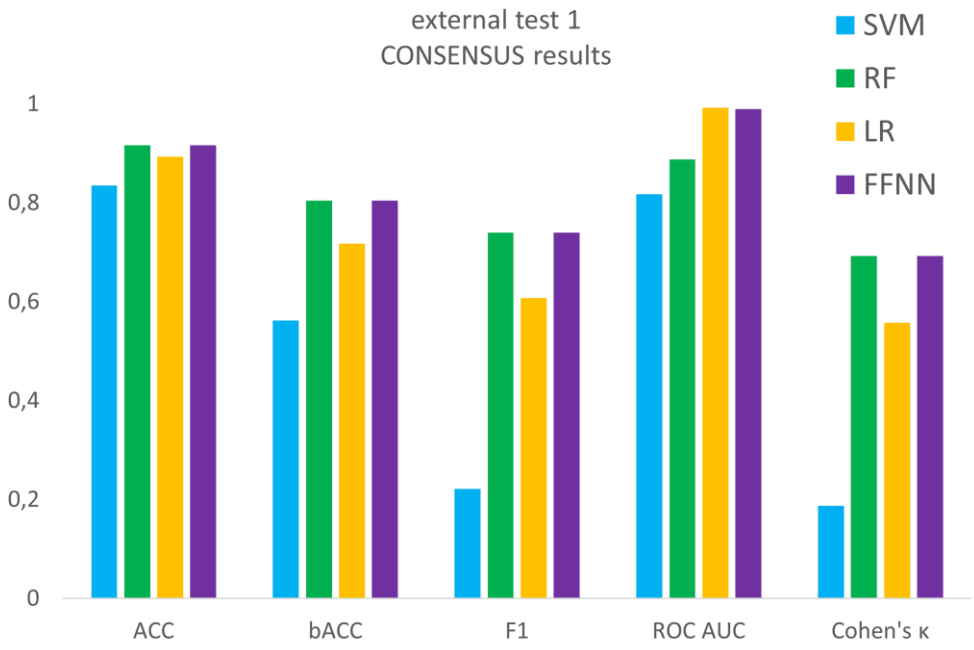

Figure 14. Performance metrics for the **1st external test** set evaluated using different AI models (Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF) and the Feed Forward Neural Network (FFNN)) across consensus function. The results are based on the classification of a dataset comprising 85 additional, previously unseen epitopes originating from the same research group that generated the initial 135-epitopes dataset. The qualitative descriptors are: ACC = accuracy, bACC = balanced accuracy, ROC AUC = receiver operating characteristic area under the curve.

The 1st external test set supported the transferability of the consensus-based approach, although clear classifier-dependent differences were observed. RF and FFNN showed the most consistent performance across threshold-dependent metrics, including bACC, F1-score, and Cohen's κ, suggesting a more balanced classification of activating and non-activating epitopes. LR achieved the highest ROC AUC, indicating strong discriminative ranking of peptide activity probabilities; however, its lower F1-score and Cohen's κ suggest that this ranking performance did not fully translate into optimal binary classification. SVM also showed good overall ACC, but its weaker F1-score and Cohen's κ indicated reduced performance regarding the identification of the active class. Together, these results suggest that the consensus input strategy remained informative in an independent external validation setting, with RF and FFNN providing the most balanced classification profiles.

To further assess the robustness and cross-dataset generalizability of the models, I evaluated their performance on a second independent external dataset (2nd external test set). The 2nd external test set contains 172 additional epitopes reported by different research group [100] and measured in three altered TCR-defined T-cell populations, referred to as TCR1, TCR2, and TCR3. These populations represent distinct T-cell lineages, with TCR1 corresponding to the avian equivalent of mammalian γδ T cells, while TCR2 and TCR3 represent two αβ T-cell lineages. Therefore, this dataset provided a more stringent validation scenario, allowing us to test whether the predictive performance (observed for 1st external test set) could be maintained across biologically distinct TCR response contexts.

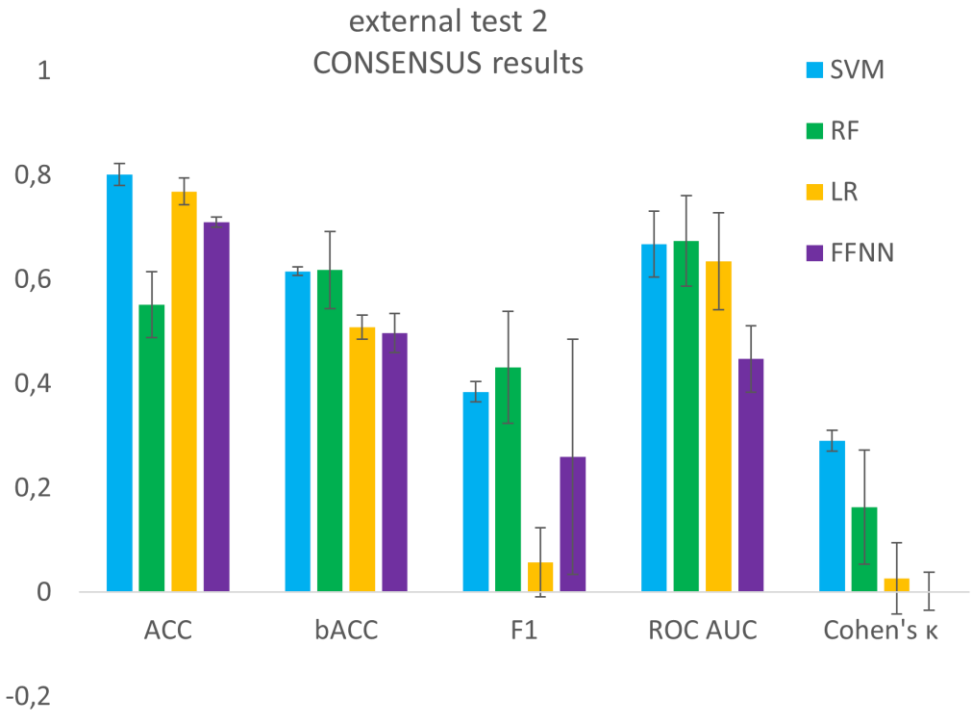

Figure 15. Qualitative descriptors' results of the similarity-matrix-based and structural-descriptor-based consensus models, and of the four-component consensus model in case of the **2nd external test set** using different AI architectures (Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF) and the Feed Forward Neural Network (FFNN)). The **external test 2** contains TCR1, TCR2 and TCR3 cell lines, and their results mean and SD can be found. The abbreviation for descriptors means: ACC = accuracy, bACC = balanced accuracy, ROC AUC = receiver operating characteristic area under the curve.).

Thus, the 2nd external test set represented a more challenging validation scenario, as reflected by the lower and more variable performance across classifiers (Figure 15.). SVM showed the most favourable overall profile, achieving the highest ACC and Cohen's κ, together with moderate bACC and ROC AUC. RF provided comparable bACC and the highest F1-score, suggesting a relatively better ability to identify activating epitopes, although its overall ACC

and Cohen's κ were lower than those of SVM. LR achieved high ACC but showed markedly reduced F1-score and Cohen's κ, indicating that its performance was strongly influenced by the dominant class and it was less effective to detect the active class. FFNN showed the weakest overall performance on this external dataset, with lower ACC and ROC AUC, suggesting limited generalization under this more heterogeneous external validation setting. Overall, these results indicate that the 2nd external test set was substantially more difficult to classify than the previous ones (internal hold-out and 1st external test set), highlighting the impact of diverse laboratory conditions, activity definitions, and experimental variability on cross-dataset model transferability. The discrepancy between ACC and F1-score, particularly for LR, indicates that ACC alone is insufficient for evaluating performance on this 2nd external dataset, and balanced metrics are required to assess active-class recognition.

4.5. *Prediction of nucleation score of Aβ mutations using ML algorithms*

In parallel with the investigations of the 9-mer epitopes and the MHC-I complexes, further ML studies focused on the nucleation score (NS) function of the 42-amino-acid Aβ peptide (Aβ1–42) [101], a key player peptide in AD progression. Nucleation process describes the initial formation of small Aβ assemblies that can subsequently grow into fibrils and plaques. This event can be quantified by NS function, which may be derived from experimental or computational methods. Recently, deep mutational scanning assays assign NS values to large number of single and double Aβ variants [102]. NS is a highly complex, multiparametric biological descriptor that reflects residue-specific effects, long-range interactions, sequence context, and cooperative structural rearrangements. As a key determinant of Aβ aggregation, NS provides a quantitative measure of mutational effects, and its value may change substantially even after a single amino acid (sAA) or double amino acid (dAA) substitution. However, despite considerable experimental effort, the currently available mutational datasets remain incomplete. Given that larger training datasets can yield more robust ML outcomes in this thesis I focus on the dAA dataset, compromising approximately 15,000 experimental measurements. The principal challenge arises from the substantial disparity between the number of these known dAA measurements and the estimated number of unknown dAA variants (approximately 280,000).

ML analysis (SVM, RF, FFNN) of Aβ variants requires the conversion of peptide sequences into numerical representations (*AAC, Z-scale, p.w.BLOSUM62 and ESM-2 embeddings*) that capture complementary aspects of amino-acid identity, position, physicochemical properties and sequence context. This is particularly important for Aβ1–42

because all variants have the same length and often differ from the wild-type peptide by only one or two residues. Consequently, representations that preserve residue position may distinguish mutational effects more effectively than global composition-based descriptors.

To investigate the relationship between sequence-derived features and Aβ nucleation propensity, three complementary regression algorithms (FFNN, RF, and SVR) representing distinct ML paradigms were evaluated. These algorithms were selected because they differ substantially in their ability to capture nonlinear relationships, feature interactions, and complex decision boundaries. Model performance was quantified using R^2, MAE, MSE, and RMSE values, thereby providing complementary measures of prediction accuracy, robustness, and goodness-of-fit.

In Figure 16. FFNN (16a), RF (16b), and SVR (16c) predicted NS values are plotted against experimentally determined data, where the red diagonal line denotes perfect agreement between calculations and experiments. Individual variants are coloured according to the position of the mutated residue within the Aβ1–42 sequence. In each case the first row presents predictions for the randomly partitioned hold-out validation dataset (30% of the dAA mutation dataset, “internal” training), the second row shows predictions for the external sAA mutation dataset, which was not used during model training.

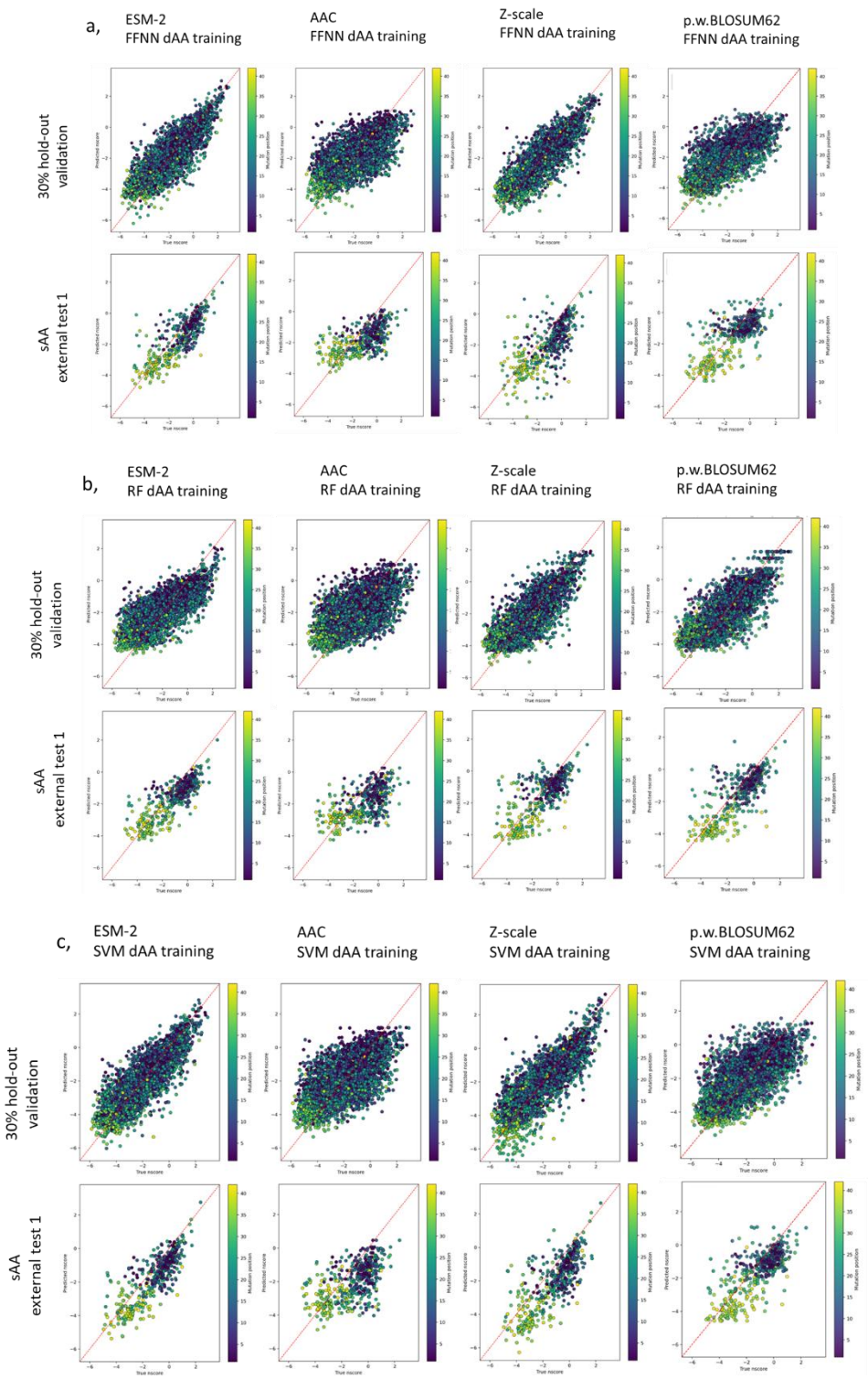

Figure 16. Comparison of regression models (feed-forward neural network: FFNN (a), random forest: RF (b) and support vector regressor: SVR (c) for predicting Aβ nucleation scores (NSs) using sequence-based representations.

.

Overall, all three regression algorithms successfully captured the relationship (see also Table 1.) (30% of the dAA mutation dataset, “internal” training), the second row shows predictions for the external sAA mutation dataset, which was not used during model training.

Table 1. Performance comparison of the four (ESM-2, Z-scale, AAC, and p.w.BLOSUM62) regression models (feed-forward neural network: FFNN, random forest: RF and support vector regressor: SVR).

	Test	Metric	SVR	RF	FFNN
ESM-2	5-fold cross validation (70%)	MAE ± SD	**0.715**±0.018	0.911±0.022	**0,688**±0.012
		MSE ± SD	**0,877**±0.044	1.287±0.042	**0,813**±0.023
		RMSE ± SD	**0,936**±0.023	1.134±0.019	**0.901**±0.013
		R^2 ± SD	**0.718**±0.012	0.586±0.009	**0.738**±0.009
	30% hold-out validation	MAE ± SD	**0.691**±0.007	0.897±0.009	**0.688**±0.017
		MSE ± SD	**0.824**±0.023	1.260±0.030	**0.813**±0.038
		RMSE ± SD	**0.908**±0.013	1.122±0.013	**0.901**±0.021
		R^2 ± SD	**0.734**±0.005	0.593±0.007	**0.737**±0.010
	sAA test (external test)	MAE	**0.879**	0.828	**0.775**
		MSE	**1.109**	0.992	**0.946**
		RMSE	**1.053**	0.996	**0.972**
		R^2	**0.508**	0.560	**0.580**
Z-scale	5-fold cross validation (70%)	MAE ± SD	0.734±0.012	**0.716**±0.011	0.663±0.014
		MSE ± SD	0.923±0.038	**0.881**±0.029	0.765±0.040
		RMSE ± SD	0.960±0.020	**0.939**±0.015	0.875±0.023
		R^2 ± SD	0.703±0.016	**0.716**±0.013	0.754±0.016
	30% hold-out validation	MAE ± SD	0.723±0.007	**0.708**±0.011	0.652±0.016
		MSE ± SD	0.900±0.024	**0.867**±0.023	0.750±0.027
		RMSE ± SD	0.949±0.013	**0.931**±0.012	0.866±0.015
		R^2 ± SD	0.709±0.006	**0.720**±0.005	0.758±0.006
	sAA test (external test)	MAE	1.149	**0.865**	1.157
		MSE	1.863	**1.175**	2.070
		RMSE	1.365	**1.084**	1.439
		R^2	0.173	**0.478**	0.081

	Test	Metric	SVR	RF	FFNN
AAC	5-fold cross validation (70%)	MAE ± SD	1.072±0.018	1.127±0.019	1.063±0.019
		MSE ± SD	1.823±0.041	1.957±0.057	1.733±0.041
		RMSE ± SD	1.350±0.015	1.399±0.020	1.316±0.016
		R^2 ± SD	0.414±0.012	0.370±0.011	0.442±0.013
	30% hold-out validation	MAE ± SD	1.066±0.010	1.111±0.007	1.063±0.011
		MSE ± SD	1.819±0.034	1.927±0.021	1.749±0.031
		RMSE ± SD	1.349±0.013	1.388±0.007	1.322±0.012
		R^2 ± SD	0.412±0.012	0.377±0.008	0.435±0.009
	sAA test (external test)	MAE	1.249	1.181	1.135
		MSE	2.277	2.083	1.835
		RMSE	1.509	1.443	1.355
		R^2	-0.01	0.075	0.185
pw.BLOSUM62	5-fold cross validation (70%)	MAE ± SD	0.984±.0.012	0.747±0.011	0.957±0.014
		MSE ± SD	1.571±0.036	0.957±0.037	1.485±0.038
		RMSE ± SD	1.253±0.014	0.978±0.019	1.219±0.015
		R^2 ± SD	0.494±0.015	0.692±0.015	0.522±0.014
	30% hold-out validation	MAE ± SD	0.969±0.011	0.750±0.022	0.971±0.016
		MSE ± SD	1.532±0.025	0.955±0.030	1.505±0.033
		RMSE ± SD	1.238±0.061	0.977±0.014	1.227±0.018
		R^2 ± SD	0.516±0.011	0.698±0.026	0.524±0.015
	sAA test (external test)	MAE	0.896	0.855	0.864
		MSE	1.272	1.193	1.201
		RMSE	1.128	1.092	1.096
		R^2	0.435	0.470	0.467

Overall, all three regression algorithms successfully captured the relationship (see also Table 1.) between sequence-derived descriptors and experimental NS on the independent validation dataset. As expected, prediction accuracy decreased for the external sAA dataset, reflecting the increased difficulty of extrapolating beyond the distribution represented in the training data.

To quantitatively compare the predictive capabilities of the evaluated ML algorithms, regression performance was assessed using four complementary metrics (Table 1). Overall, the

FFNN consistently achieved the highest predictive performance across almost all the investigated sequence representations, yielding the highest coefficients of determination together with the lowest MAE, MSE, and RMSE values in both the internal validation and external test (30% hold-out test: R^2: 0.737 and sAA external test: R^2: 0.580 in ESM-2). The RF model also demonstrated competitive performance, although its prediction errors were generally higher than those of the FFNN (FFNN RMSE: 0.901 vs RF RMSE: 0.908). In contrast, SVR exhibited lower predictive accuracy, particularly for previously unseen variants, indicating a reduced ability to generalize to sequence patterns outside the training distribution. The agreement between the numerical performance metrics and the scatter plots shown in Figure 16. demonstrates that lower prediction errors are accompanied by a tighter clustering of data points around the identity line. Collectively, these results suggest that deep neural networks are better suited to modelling the highly nonlinear relationship between amino acid sequence and Aβ nucleation propensity, whereas ensemble tree methods provide a robust alternative and kernel-based regression offers comparatively limited predictive capability for this task.

After evaluating the predictive performance of individual sequence representation–model combinations, an additional consensus analysis was performed to investigate whether integrating predictions could improve the robustness and generalization of the regression models. Rather than averaging all available models, only those combinations that demonstrated superior predictive performance were included in the consensus.

The consensus model combined the predictions of **ESM-2–FFNN, ESM-2–SVR,** and **Z-scale–RF**, whereas these have the highest predictive performances. This strategy allowed us to evaluate whether extending the ensemble with an additional competitive predictor improved overall performance and generalization.

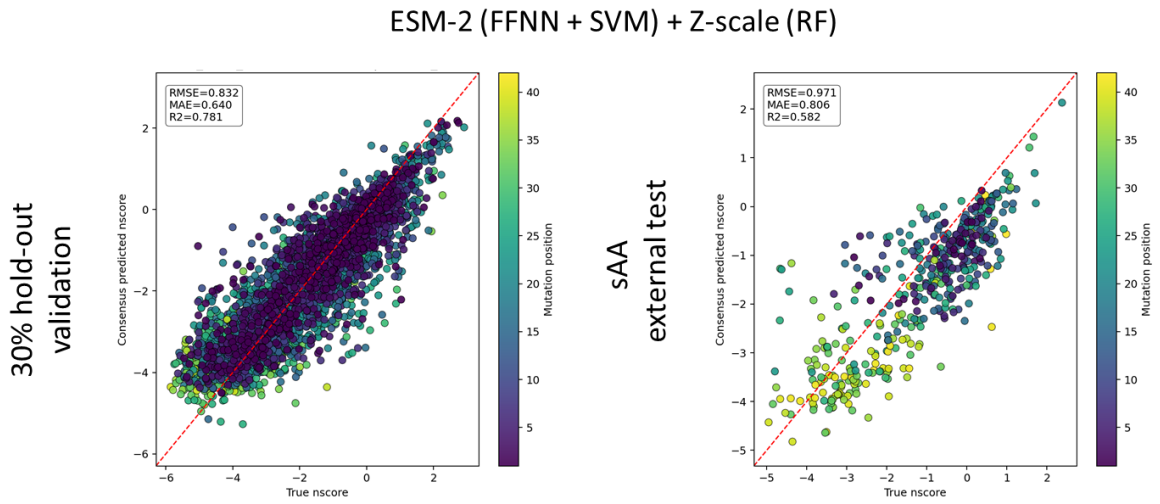

Figure 17. Comparison of model-representation wise consensus (consensus of input and model types: ESM-2: FFNN + SVR and Z-scale: RF).

The performance of the consensus model is summarized in Figure 17. The consensus strategy yielded highly accurate predictions on the independent hold-out validation dataset, with prediction points closely distributed around the identity line. The consensus model achieved the highest validation performance, producing an R^2 value of **0.781** and an RMSE of 0.832 while originally the highest performance in the validation test (ESM-2: FFNN) was R^2 value of **0.737** and an RMSE of 0.901. These results indicate that incorporating the three picked models improved prediction accuracy on data originating from the same distribution as the training set.

A similar trend was observed for the independent external mutation dataset; however, the overall applicability to the external test set remains rather limited. Specifically, the consensus model demonstrated modest generalization, achieving an R^2 of **0.582** and an RMSE of 0.971, compared to the originally highest-performing external test model (ESM-2: FFNN), which yielded an R^2 of **0.580** and an RMSE of 0.972. These values indicate only a marginal improvement, as well as overall weak predictive performance. Nevertheless, this observation highlights an interesting characteristic of mutational effects in peptides: the physicochemical differences between, for example, single and double mutations can be substantially larger than what is typically inferred from experience with protein similarity. Specifically, based on the substitutions observed in double mutants, we initially hypothesized that single mutations could be predicted with reasonable accuracy by applying reasoning analogous to protein homology. However, the low R^2 values obtained on the external test set contradict this expectation,

suggesting that mutational effects in peptides may not transfer as straightforwardly as in full-length proteins.

Returning to the composition of the consensus model, the outcomes are readily interpretable from both biological and computational perspectives. The ESM-2 representation contributes contextual sequence information learned from large-scale protein databases, while the Z-scale descriptors capture complementary physicochemical characteristics of amino acids. Furthermore, the participating regression algorithms represent different learning paradigms: FFNN models complex nonlinear relationships, SVR provides kernel-based regression with strong regularization, and RF captures nonlinear feature interactions through ensemble decision trees. Consequently, the first consensus integrates both complementary sequence representations and complementary regression strategies, likely explaining its improved robustness on the independent external dataset. Overall, these findings suggest that careful selection of complementary predictors is more important than simply increasing the size of the ensemble. The best generalization performance was achieved by combining a limited number of high-performing and diverse model–representation pairs rather than by averaging all available predictors.

5. Conclusions

During my doctoral studies I combined several computational and experimental approaches to investigate structure–affinity relationships across diverse biological systems.

Molecular dynamics simulations suggested that the C-2 positions chlorophenyl substituent of 13AES3 plays a key role in stabilizing the steroid scaffold within the taxene binding site in **tubulin protein**. The deeper insertion of the chlorophenyl ring into the hydrophobic pocket, together with a stronger interaction with PHE272, resulted in reduced ligand mobility and a more stable binding pose than that observed for 13AES2. Although 13AES2 also formed hydrophobic contacts with PHE272, these interactions were weaker and less well defined, which may explain why the ligand could not maintain a comparably stable orientation.

Docking and interaction analyses on the **AKR1C isoenzymes** further demonstrated that both ligands bind more strongly to AKR1C2 than to AKR1C1, in agreement with the experimental inhibition data. The selectivity appears to arise from subtle differences in binding orientation and active-site geometry, particularly the Leu54/Val54 substitution. In AKR1C1, the

ligands orient their A-ring toward NADPH, whereas in AKR1C2 the D-ring is positioned closer to the cofactor. LID analysis suggested that a hydrogen bond with HIE222 in AKR1C2 may contribute to stronger binding, whereas in AKR1C1 a π–π interaction with TRP227 dominates. These findings highlight how small active-site differences can strongly affect isoenzyme selectivity.

Following the steroid-related investigation, I turned my attention to another line in which I provided theoretical, ITC and STD-NMR evidences for the interactions of **KYNA** and its derivatives with **SAs**. KYNA showed clear STD responses with both HSA and BSA, indicating direct and robust binding. In contrast, the modified derivatives SZR72 and SZR104 exhibited weaker STD amplification, consistent with reduced affinity relative to KYNA. Although the derivatives showed broadly similar STD behaviour with the two albumins, the overall signal intensities were lower, suggesting altered binding dynamics and weaker complex formation. Together with ITC and PMF data, these results support a clear structure–affinity trend across the ligand series.

Finally, in the **ML-based MHC-I epitope activation study**, my results did not support a single universal classifier across all validation settings. Instead, FFNN, RF, LR, and SVM each contributed complementary predictive strengths depending on the external test scenario. RF offered stable nonlinear performance across the two external datasets, whereas FFNN performed well on the hold-out (internal test set) and on the external test case (1st external test set), but was less robust on the more distinct external test (2nd external test set). These results support the application of consensus-based representations in combination with multiple complementary classifiers and underscore the importance of independent external validation and balanced performance metrics when evaluating TCR-related epitope activation models.

Using similar ML-based investigations, the consensus model itself proved informative computational perspective in **Aβ1-42 peptide related mutational examinations**. The ESM-2 added context-aware sequence information derived from large protein databases, whereas Z-scale descriptors contributed complementary physicochemical features of Aβ peptides. The use of diverse regression paradigms (FFNN, SVR, and RF) likely explains the improved generalization of the internal consensus model to predict NS values of Aβ mutations. Overall, the findings indicate that carefully selected, consensus of the complementary model–representation pairs can outperform larger but less diverse ensembles, highlighting the value of combining biological interpretability with computational diversity.

Thesis points

1. The computational investigation (MM/GBSA and US methods) of modified steroid molecules binding to taxene binding site of tubulin protein showed that the chlorophenyl ring in 13AES3 makes stronger hydrophobic interaction and increase the binding stability than the 13AES2 and 13AES1. The experimental binding preference order was supported by the computational results [92].
2. The EM1062 and EM1082 modified steroids bind stronger to the AKR1C2 isoenzyme than AKR1C1, indicating that subtle structural differences in the active site significantly influence selectivity in agreement with the experimental results [96].
3. The binding and binding preference order of KYNA to HSA and BSA proteins is confirmed by experimentally and computationally. However, the same experimental and theoretical techniques for modified derivatives exhibit weaker affinity than the parent compound which indicates that a different drug delivery method is necessary for the modified compounds [98].
4. Considering various ML classifiers, there is no single best model for predicting epitope activation in MHC-I complexes; FFNN, RF, LR, and SVM each offer distinct advantages under different validation scenarios. A consensus model combining different peptide representations provided the most robust performance on both the internal test set and an independent external test set [99].
5. ML algorithms (FFNN, SVM, and RF) were used to predict the NS values of Aβ1-42 peptide double mutational variants, I found that the most accurate predictions were obtained using a consensus model combining different algorithms and amino acid representations. I also observed that a model trained on dAA mutations is not suitable for predicting the NSs of sAA mutations [101].

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to Prof. Dr, Tamás Martinek, the Head of the Department of Medical Chemistry in University of Szeged, who gave me a chance to work here. I would like to thank the help and support of my colleges and excolleges in the department: Ibolya Zita Papp, Vencel Petrovicz, Attila Tököli, András Márton, Edit Wéber and Zsófia Hegedűs. I am appreciative to Györgyi Váradi, for the opportunity to be involved in the education of first-year medical students. My sincere thanks also go to Ferenc Bogár, Lajos Kovács and Zoltán Kupihár for their valuable life lessons and for ensuring that my caffeine levels never dropped to low.

I am glad to my friends for supporting me, although usually they do not understand what I am talking about. The words cannot describe how thankful I am to my family: my dad and my mum. I cannot go on without mentioning my boyfriend for giving inspiring thoughts during difficult times and believed me, especially toward to the end.

Last but not least, my acknowledgement dedicated to my supervisor: Dr. habil. Gábor Paragi. He stood by me and believed in me when no one else did. His interdisciplinary expertise and experience set an example for me, moreover his resilience, patience and grit showed me how to be a good and enthusiastic researcher and teacher.

REFERENCES

1. Dr. Roman Schmied. *An Introduction to Computational Quantum Mechanics*. 2nd edn, 2014.

2. Hehre WJ. *A Guide to Molecular Mechanics and Quantum Chemical Calculations*. n.p.: Wavefunction, Inc., 2003.

3. Kleinschmidt AT, Chen AX, Pascal TA *et al.* Computational Modeling of Molecular Mechanics for the Experimentally Inclined. *Chem Mater* 2022;**34**(17):7620–34. https://doi.org/10.1021/acs.chemmater.2c00292.

4. Abramson J, Adler J, Dunger J *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 2024;**630**(8016):493–500. https://doi.org/10.1038/s41586-024-07487-w.

5. Jay Alammar, Maarten Grootendorst. *Hands-On Large Language Models*. n.p.: O'Reilly, 2024. https://www.oreilly.com/library/view/hands-on-large-language/9781098150952/.

6. Choudhary K, DeCost B, Chen C *et al.* Recent advances and applications of deep learning methods in materials science. *Npj Comput Mater* 2022;**8**(1):59. https://doi.org/10.1038/s41524-022-00734-6.

7. Sazlı Murat. A Brief Review of Feed-Forward Neural Networks. *Commun Fac Sci* 2006;**50**(1):11–7. https://doi.org/10.1501/commua1-2_0000000026.

8. BISHOP CM. *Neural Networks for Pattern Recognition*. OXFORD: CLARENDON PRESS, 1995.

9. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;**20**(3):273–97. https://doi.org/10.1007/BF00994018.

10. Asa Ben-Hur, Cheng Soon Ong, Sören Sonnenburg *et al.* Support Vector Machines and Kernels for Computational Biology. *PLOS Comput Biol* 2008;**4**(10). https://doi.org/10.1371/journal.pcbi.1000173.

11. Awad M, Khanna R. Support Vector Regression. In: Awad M, Khanna R, *Efficient Learning Machines*. Berkeley, CA: Apress, 2015, 67–80. https://doi.org/10.1007/978-1-4302-5990-9_4.

12. Das A. Logistic Regression. In: Michalos AC (ed.), *Encyclopedia of Quality of Life and Well-Being Research*. Dordrecht: Springer Netherlands, 2014, 3680–2. https://doi.org/10.1007/978-94-007-0753-5_1689.

13. Breiman L. Random Forests. *Mach Learn* 2001;**45**(1):5–32. https://doi.org/10.1023/A:1010933404324.

14. Geoffrey M. Cooper. *The Cell: A Molecular Approach*. 2nd edn, n.p.: Sinauer Associates, 2000.

15. Linda A Amos, Daniel Schlieper. Microtubules and maps. *Adv Protein Chem* 2005;**71**:257–98. https://doi.org/10.1016/S0065-3233(04)71007-4.

16. M Abal, J M Andreu, I Barasoain. Taxanes: microtubule and centrosome targets, and cell cycle dependent mechanisms of action. *Curr Cancer Drug Targets* 2003;**3**:193–203. https://doi.org/10.2174/1568009033481967.

17. Xingyu Wang, Benoît Gigant, Xi Zheng *et al.* Microtubule-targeting agents for cancer treatment: Seven binding sites and three strategies. *MedComm – Oncol* 2003;**2**(3):e46. https://doi.org/10.1002/mog2.46.

18. Fontaine-Lenoir, Virginie, Chambraud, Béatrice, Fellous, Arlette *et al.* Microtubule-associated protein 2 (MAP2) is a neurosteroid receptor. *Proc Natl Acad Sci* 2006;**103**(12):4711–6. https://doi.org/10.1073/pnas.0600113103.

19. Trevor M Penning, Phumvadee Wangtrakuldee, Richard J Auchus. Structural and Functional Biology of Aldo-Keto Reductase Steroid-Transforming Enzymes. *Endocr Rev* 2018;**40**(2):447–75. https://doi.org/10.1210/er.2018-00089.

20. Rižner TL, Penning TM. Role of aldo–keto reductase family 1 (AKR1) enzymes in human steroid metabolism. *Steroids* 2014;**79**:49–63. https://doi.org/10.1016/j.steroids.2013.10.012.

21. Ruili Guan, Changhao Yang. Dehydroepiandrosterone alleviates hypoxia-induced learning and memory dysfunction by maintaining synaptic homeostasis. *CNS Neurosci Ther* n.d.;**28**(9):1339–50. https://doi.org/10.1111/cns.13869.

22. S Luchetti, I Huitinga, D F Swaab. Neurosteroid and GABA-A receptor alterations in Alzheimer's disease, Parkinson's disease and multiple sclerosis. *Neuroscience* 2011;**191**:6–21. https://doi.org/10.1016/j.neuroscience.2011.04.010.

23. Azargoonjahromi A. The duality of amyloid-β: its role in normal and Alzheimer's disease states. *Mol Brain* 2024;**17**(1):44. https://doi.org/10.1186/s13041-024-01118-1.

24. Zhang J, Zhang Y, Wang J *et al.* Recent advances in Alzheimer's disease: mechanisms, clinical trials and new drug development strategies. *Signal Transduct Target Ther* 2024;**9**(1):211. https://doi.org/10.1038/s41392-024-01911-3.

25. Valérie Wilquet, Bart De Strooper. Amyloid-beta precursor protein processing in neurodegeneration. *Curr Opin Neurobiol* 2004;**5**:582–8. https://doi.org/10.1016/j.conb.2004.08.001.

26. Hampel H, Hardy J, Blennow K *et al.* The Amyloid-β Pathway in Alzheimer's Disease. *Mol Psychiatry* 2021;**26**(10):5481–503. https://doi.org/10.1038/s41380-021-01249-0.

27. Zhang Y, Chen H, Li R *et al.* Amyloid β-based therapy for Alzheimer's disease: challenges, successes and future. *Signal Transduct Target Ther* 2023;**8**(1):248. https://doi.org/10.1038/s41392-023-01484-7.

28. Lin Wai Hung, Giuseppe D. Ciccotosto, Eleni Giannakis, Deborah *et al.* Amyloid-β Peptide (Aβ) Neurotoxicity Is Modulated by the Rate of Peptide Aggregation: Aβ Dimers and Trimers Correlate with Neurotoxicity. *J Neurosci* 2008;**28**(46):11950–8. https://doi.org/10.1523/JNEUROSCI.3916-08.2008.

29. Haass C, Lemere CA, Capell A *et al.* The Swedish mutation causes early-onset Alzheimer's disease by β-secretase cleavage within the secretory pathway. *Nat Med* 1995;**1**(12):1291–6. https://doi.org/10.1038/nm1295-1291.

30. Guerreiro RJ, Baquero M, Blesa R *et al.* Genetic screening of Alzheimer's disease genes in Iberian and African samples yields novel mutations in presenilins and APP. *Neurobiol Aging* 2010;**31**(5):725–31. https://doi.org/https://doi.org/10.1016/j.neurobiolaging.2008.06.012.

31. E Levy, M D Carman, I J Fernandez-Madrid *et al.* Mutation of the Alzheimer's disease amyloid gene in hereditary cerebral hemorrhage, Dutch type. *Science* 1990;**248**(4959):1124–6. https://doi.org/10.1126/science.2111584.

32. Nilsberth C, Westlind-Danielsson A, Eckman CB *et al.* The 'Arctic' APP mutation (E693G) causes Alzheimer's disease by enhanced Aβ protofibril formation. *Nat Neurosci* 2001;**4**(9):887–93. https://doi.org/10.1038/nn0901-887.

33. Sorbi S, Nacmias B, Mortilla M *et al.* Molecular genetics of Alzheimer's disease in Italian families. *Neurochem Int* 1994;**25**(1):81–4. https://doi.org/10.1016/0197-0186(94)90057-4.

34. Jonsson T, Atwal JK, Steinberg S *et al.* A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature* 2012;**488**(7409):96–9. https://doi.org/10.1038/nature11283.

35. Thomas Trolle, Curtis P McMurtrey, John Sidney *et al.* The length distribution of class I restricted T cell epitopes is determined by both peptide supply and MHC allele specific binding preference. *J Immunol* 2016;**196**(4):1480–7. https://doi.org/10.4049/jimmunol.1501721.

36. Galina Petrova, Andrea Ferrante, Jack Gorski. Cross-Reactivity of T Cells and Its Role in the Immune System. *Crit Rev Immunol* 2012;**32**(4):349–72. https://doi.org/10.1615/critrevimmunol.v32.i4.50.

37. Cebrián C, Loike JD, Sulzer D. Neuronal MHC-I expression and its implications in synaptic function, axonal regeneration and Parkinson's and other brain diseases. *Front Neuroanat* 2014;**Volume 8-2014**. https://doi.org/10.3389/fnana.2014.00114.

38. Gunjan Dhawan, Angie M Floden, Colin K Combs. Amyloid-β oligomers stimulate microglia through a tyrosine kinase dependent mechanism. *Neurobiol Aging* 2011;**33**(10):2247–61. https://doi.org/10.1016/j.neurobiolaging.2011.10.027.

39. Sondag CM, Dhawan G, Combs CK. Beta amyloid oligomers and fibrils stimulate differential activation of primary microglia. *J Neuroinflammation* 2009;**6**(1):1. https://doi.org/10.1186/1742-2094-6-1.

40. Benjamin D Hobson, David Sulzer. Neuronal Presentation of Antigen and Its Possible Role in Parkinson's Disease. *J Park Dis* 2022;**12**(1):137–47. https://doi.org/10.3233/JPD-223153.

41. Shastri D, Raj V, Lee S. Revolutionizing Alzheimer's treatment: Harnessing human serum albumin for targeted drug delivery and therapy advancements. *Ageing Res Rev* 2024;**99**:102379. https://doi.org/https://doi.org/10.1016/j.arr.2024.102379.

42. Litus EA, Kazakov AS, Sokolov AS *et al.* The binding of monomeric amyloid β peptide to serum albumin is affected by major plasma unsaturated fatty acids. *Biochem Biophys Res Commun* 2019;**510**(2):248–53. https://doi.org/https://doi.org/10.1016/j.bbrc.2019.01.081.

43. Helen F Stanyon, John H Viles. Human Serum Albumin Can Regulate Amyloid-β Peptide Fiber Growth in the Brain Interstitium. *J Biol Chem* 2012;**287**(33):28163–8. https://doi.org/10.1074/jbc.C112.360800.

44. Schwarcz R, Bruno JP, Muchowski PJ *et al.* Kynurenines in the mammalian brain: when physiology meets pathology. *Nat Rev Neurosci* 2012;**13**(7):465–77. https://doi.org/10.1038/nrn3257.

45. Stone TW. Neuropharmacology of quinolinic and kynurenic acids. *Pharmacol Rev* 1993;**45**(3):309–79. https://doi.org/10.1016/S0031-6997(25)00441-7.

46. Inseyah Bagasrawala, Nada Zecevic, Nevena V Radonjić. N-Methyl D-Aspartate Receptor Antagonist Kynurenic Acid Affects Human Cortical Development. *Front Neurosci* 2016;**10**(435). https://doi.org/10.3389/fnins.2016.00435.

47. Shinsuke Fukui, Robert Schwarcz, Stanley I. Rapoport *et al.* Blood–Brain Barrier Transport of Kynurenines: Implications for Brain Synthesis and Metabolism. *J Neurochem* 1991;**56**(6):2007–17. https://doi.org/10.1111/j.1471-4159.1991.tb03460.x.

48. Szabo M, Lajkó N, Dulka K *et al.* The kynurenic acid analog SZR104 induces cytomorphological changes associated with the anti-inflammatory phenotype in cultured microglia. *Sci Rep* 2023;**13**(1):11328. https://doi.org/10.1038/s41598-023-38107-8.

49. Balla Z, Kormányos ES, Kui B *et al.* Kynurenic Acid and Its Analogue SZR-72 Ameliorate the Severity of Experimental Acute Necrotizing Pancreatitis. *Front Immunol* 2021;**Volume 12-2021**. https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2021.702764.

50. Li J, Fu A, Zhang L. An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking. *Interdiscip Sci Comput Life Sci* 2019;**11**(2):320–8. https://doi.org/10.1007/s12539-019-00327-w.

51. Trott, O., Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 2010;**32**(2):455–61. https://doi.org/10.1002/jcc.21334.

52. Friesner RA, Banks JL, Murphy RB *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J Med Chem* 2004;**47**(7):1739–49. https://doi.org/10.1021/jm0306430.

53. Verdonk, Marcel L., Cole, Jason C., Hartshorn, Michael J. *et al.* Improved protein–ligand docking using GOLD. *Proteins Struct Funct Bioinforma* 2003;**52**(4). https://doi.org/10.1002/prot.10465.

54. Halgren TA, Murphy RB, Friesner RA *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J Med Chem* 2004;**47**(7):1750–9. https://doi.org/10.1021/jm030644s.

55. Muegge I, Martin YC, Hajduk PJ *et al.* Evaluation of PMF Scoring in Docking Weak Ligands to the FK506 Binding Protein. *J Med Chem* 1999;**42**(14):2498–503. https://doi.org/10.1021/jm990073x.

56. Dittrich J, Schmidt D, Pfleger C *et al.* Converging a Knowledge-Based Scoring Function: DrugScore2018. *J Chem Inf Model* 2018;**59**(1):509–21. https://doi.org/10.1021/acs.jcim.8b00582.

57. Abramson J, Adler J, Dunger J *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 2024;**630**(8016):493–500. https://doi.org/10.1038/s41586-024-07487-w.

58. Gabriele Corso, Hannes Stärk, Bowen Jing *et al. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking*. published online 2023. https://doi.org/10.48550/arXiv.2210.01776.

59. Gentile F, Agrawal V, Hsing M *et al.* Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Cent Sci* 2020;**6**(6):939–49. https://doi.org/10.1021/acscentsci.0c00229.

60. Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin Drug Discov* 2015;**10**(5):449–61. https://doi.org/10.1517/17460441.2015.1032936.

61. Case, D.A. *AMBER*. n.d.

62. Aho N, Groenhof G, Buslaev P. Do All Paths Lead to Rome? How Reliable is Umbrella Sampling Along a Single Path? *J Chem Theory Comput* 2024;**20**(15):6674–86. https://doi.org/10.1021/acs.jctc.4c00134.

63. Li Y, Patel JM, Terrell A. WHAM: A High-Throughput Sequence Alignment Method. *ACM Trans Database Syst* 2012;**37**(4):1–39. https://doi.org/10.1145/2389241.2389247.

64. Ian Goodfellow, Yoshua Bengio, Aaron Courville. *Deep Learning*. n.p.: MIT Press, 2016.

65. Steven Heinkoff, Jorja G. Heinkoff. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences. *Proc Natl Acad Sci* 1992;**89**(22):10915–9. https://doi.org//10.1073/pnas.89.22.10915.

66. M.O. Dayhof, R.M. Schwartz, B. C. Orcutt. *A Model of Evolutionary Change in Proteins*, vol. 5. n.p.: National Biomedical Research Foundation Silver Spring, 1978.

67. Martin Christen Frølund Thomsen, Morten Nielsen. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res* 2012;**40**(1):281–7. https://doi.org//10.1093/nar/gks469.

68. Andrea Mauri, Viviana Consonni, Manuela Pavan *et al.* Dragon Software: an easy approach to molecular descriptor calculations. *Match Commun Math Comput Chem* 2006;**56**(2):237–48.

69. Deborah Giordano, Carmen Biancaniello, Maria Antonia Argenio *et al.* Drug Design by Pharmacophore and Virtual Screening Approach. *Pharmaceuticals* 2022;**15**(5):646. https://doi.org//10.3390/ph15050646.

70. *Schrödinger Release Canvas*. versions 2025-1. 2025.

71. Josh Abramson, Jonas Adler, Jack Dunger et.al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 2024;**630**:493–500. https://doi.org//10.1038/s41586-024-07487-w.

72. Chen Z, Liu X, Zhao P *et al. iFeatureOmega:* an integrative platform for engineering, visualization and analysis of features from molecular sequences, structural and ligand data sets. *Nucleic Acids Res* 2022;**50**(W1):W434–47. https://doi.org/10.1093/nar/gkac351.

73. Sandberg M, Eriksson L, Jonsson J *et al.* New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *J Med Chem* 1998;**41**(14):2481–91. https://doi.org/10.1021/jm9700575.

74. Kubyshkin V. Experimental lipophilicity scale for coded and noncoded amino acid residues. *Org Biomol Chem* 2021;**19**(32):7031–40. https://doi.org/10.1039/D1OB01213D.

75. Nguyen TT, Jiang Z, Nguyen VN *et al.* Integrating ESM-2 and Graph Neural Networks with AlphaFold-2 Structures for Enhanced Protein Function Prediction. *ACS Omega* 2025;**10**(33):38103–11. https://doi.org/10.1021/acsomega.5c05484.

76. Vieira LC, Handojo ML, Wilke CO. Medium-sized protein language models perform well at transfer learning on realistic datasets. *Sci Rep* 2025;**15**(1):21400. https://doi.org/10.1038/s41598-025-05674-x.

77. Lin Z, Akin H, Rao R *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**(6637):1123–30. https://doi.org/10.1126/science.ade2574.

78. Yasen Jiao, Pufeng Du. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quant Biol* 2016;**4**(4):320–30. https://doi.org//10.1007/s40484-016-0081-2.

79. C.J. Van Rijsbergen. *Information Retrieval*, vol. 6. 2nd edn, n.p.: Butterworths, 1979.

80. Mary L McHugh. Interrater reliability: the kappa statistic. *Biochem Med* 2012;**22**(3):276–82. https://doi.org//10.11613/BM.2012.031.

81. Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit* 1997;**30**(7):1145–59. https://doi.org//10.1016/S0031-3203(96)00142-2.

82. Isothermal titration calorimetry. *Nat Rev Methods Primer* 2023;**3**(1):18. https://doi.org/10.1038/s43586-023-00216-z.

83. Scheuermann TH, Brautigam CA. High-precision, automated integration of multiple isothermal titration calorimetric thermograms: New features of NITPIC. *Methods* 2015;**76**:87–98. https://doi.org/10.1016/j.ymeth.2014.11.024.

84. Zhao H, Piszczek G, Schuck P. SEDPHAT – A platform for global ITC analysis and global multi-method analysis of molecular interactions. *Methods* 2015;**76**:137–48. https://doi.org/10.1016/j.ymeth.2014.11.012.

85. Chad A. Brautigam. *GUSSI*. 2015.

86. Viegas A, Manso J, Nobrega FL *et al.* Saturation-Transfer Difference (STD) NMR: A Simple and Fast Method for Ligand Screening and Characterization of Protein Binding. *J Chem Educ* 2011;**88**(7):990–4. https://doi.org/10.1021/ed101169t.

87. Meyer B, Peters T. NMR Spectroscopy Techniques for Screening and Identifying Ligand Binding to Protein Receptors. *Angew Chem Int Ed* 2003;**42**(8):864–90. https://doi.org/10.1002/anie.200390233.

88. Kumar BS, Raghuvanshi DS, Hasanain M *et al.* Recent Advances in chemistry and pharmacology of 2-methoxyestradiol: An anticancer investigational drug. *Steroids* 2016;**110**:9–34. https://doi.org/10.1016/j.steroids.2016.03.017.

89. El-Kady DS, Ali NA, Sayed AH *et al.* Assessment of the Antitumor Potentiality of Newly Designed Steroid Derivatives: Pre-Clinical Study. *Asian Pac J Cancer Prev* 2019;**20**(10):3057–70. https://doi.org/10.31557/APJCP.2019.20.10.3057.

90. Minorics R, Zupko I. Steroidal Anticancer Agents: An Overview of Estradiol-related Compounds. *Anticancer Agents Med Chem* 2018;**18**(5):652–66. https://doi.org/10.2174/1871520617666171114111721.

91. Howarth NM, Purohit A, Reed MJ *et al.* Estrone sulfamates: potent inhibitors of estrone sulfatase with therapeutic potential. *J Med Chem* 1994;**37**(2):219–21. https://doi.org/10.1021/jm00028a002.

92. Ali H, Traj P, Szebeni GJ *et al.* Investigation of the Antineoplastic Effects of 2-(4-Chlorophenyl)-13α-Estrone Sulfamate against the HPV16-Positive Human Invasive Cervical Carcinoma Cell Line SiHa. *Int J Mol Sci* 2023;**24**(7):6625. https://doi.org/10.3390/ijms24076625.

93. Schrödinger. *Schrödinger Release Desmond*. n.d.

94. Schrödinger. *Schrödinger Release Maestro*, versions 2022-3. n.d.

95. Rižner TL, Penning TM. Aldo-keto reductase 1C3—Assessment as a new target for the treatment of endometriosis. *Pharmacol Res* 2020;**152**:104446. https://doi.org/10.1016/j.phrs.2019.104446.

96. Resch V, Gjorgoska M, Hafner E *et al.* Targeting Steroid-Metabolizing Enzymes with 15β-Substituted Estrone Analogues: Dual Discovery of AKR1C2/17β-HSD1 Inhibitors and a Fluorescent 17β-HSD1 Ligand. *Cancers* 2026;**18**(12):1889. https://doi.org/10.3390/cancers18121889.

97. Hornok V, Juhász Á, Paragi G *et al.* Thermodynamic and kinetic insights into the interaction of kynurenic acid with human serum albumin: Spectroscopic and calorimetric approaches. *J Mol Liq* 2020;**313**:112869. https://doi.org/10.1016/j.molliq.2020.112869.

98. Vivien Erzsébet Resch, Edit Wéber, Zsófia Hegedűs *et al.* Binding properties of Kyuneric Acid Analogues to Serum Albumins. ??? 2026.

99. Vivien Erzsébet Resch, László Tóth, Anita Rácz *et al.* Machine Learning-based prediction of Cross Immunity. *Brief Bioinforma* published online 2026. https://doi.org/accepted.

100. Lun Cai, Leidy D Caraballo Galva, Yibing Peng *et al.* Preclinical Studies of the Off-Target Reactivity of AFP158-Specific TCR Engineered T Cells. *Front Immunol* 2020;**11**(607). https://doi.org//10.3389/fimmu.2020.00607.

101. Vivien Erzsébet Resch, Anita Rácz, Gábor Paragi. Mutation-Induced Changes in Nucleation Scores of Aβ-peptide: A Machine Learning Study. *J Cheminformatics* 2026.

102. Seuma M, Faure AJ, Badia M *et al.* The genetic landscape for amyloid beta fibril nucleation accurately discriminates familial Alzheimer's disease mutations. *eLife* 2021;**10**:e63364. https://doi.org/10.7554/eLife.63364.