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

1. Introduction

Understanding complex biological systems requires multiscale perspective. Since quantum mechanics (QM) is not feasible for large biomolecules, classical molecular mechanics (MM) and machine learning (ML) models (e.g., feed-forward neural networks (FFNN), support vector machine and regressor (SVM, SVR), random forest (RF), logistic regression (LR)) provide efficient alternatives. I applied these computational methods to study different biological systems which are connected to the central nervous system (CNS) into different level.

One of the main focuses of my research was the investigation of tubulin protein, which is a component of the cytoskeleton that contributes intracellular transport, cell division, etc. The function of microtubules relies on dynamic instability, which mechanism can be modulated by various microtubule-associated proteins, microtubule-targeting agents (e.g., taxene), and steroid molecules, thereby regulating the cell cycle or inducing apoptosis (like in the faster divided cancer cells).

Closely related to the biological significance of steroids, I also investigated the role of aldo-keto reductases (AKRs),

specifically the AKR1C subfamily, which are key regulators of steroid metabolism. These enzymes play a role not only in the pathological processes of peripheral organs (such as the development of hormone-sensitive tumors) but also influence the neurosteroid homeostasis of the CNS.

It is known, regarding the immune system relevance, that the presentation of self or non-self epitopes to T-cells (via MHC-I proteins) can trigger an immune response. However, recent studies have also linked amyloid toxicity to immune system function, as Aβ oligomers can activate microglial cells, which release proinflammatory cytokines. Aβ oligomers are produced when the processing of the neuroprotective amyloid precursor protein shifts toward the pathological "amyloidogenic" pathway, leading to the production of toxic Aβ40/42 peptides. The dissertation highlights how genetic mutations alter the value of Aβ aggregation related function and whether this can be analyzed using computational methods.

Finally, I addressed the challenge of delivering neuroprotective agents, such as kynurenic acid (KYNA) and its analogues across the blood-brain barrier. The research investigated to use the human and bovine serum albumins

(HSA, BSA, together SAs) as potential carrier proteins that could be applied in the near future for drug development.

2. Aims of the PhD work

1. Computational investigation (MM/GBSA and umbrella sampling (US) methods) of modified steroid molecules binding to taxene binding sites of tubulin protein.
2. Computational investigation (docking studies) of the EM1062 and EM1082 modified steroids binding to AKR1C2 and AKR1C1 isoenzymes.
3. Characterization the binding of KYNA and its derivatives to SAs by experimentally and computationally, to highlighting the structure-affinity relationships.
4. Application of AI models (FFNN, RF, LR and SVM) to predict T-cell activation of 9-mer epitopes. We compare their performance across two different test scenarios.
5. Exploring how well different AI models (FFNN, RF, SVR) can predict mutated Aβ nucleation score values using different (ESM-2 embeddings, Z-scale, AAC and p.w.BLOSUM62) representations.

3. Results and Discussions

Since many chemotherapeutic agents are associated with undesirable side effects and resistance can further limit the efficacy of treatment, steroid skeleton become an attractive feature, as even minor structural modifications can lead to marked changes in the pharmacological profile. The **first** research **line** during my PhD studies was the application of MM/GBSA and US methods to explain at molecular level the binding preference order of three selected modified steroid compounds to tubulin protein having promising biological activity in HeLa and SiHa cell lines (IC_{50} of 13AES3 > 13AES2 > 13AES1).

According to the US calculations, ligand 13AES3 had the strongest interaction with the binding site: PMF_{13AES3}: -14 kcal/mol, PMF_{13AES2}: -6.5 kcal/mol and PMF_{13AES1}: -5.2 kcal/mol values. Furthermore, the MM/GBSA calculations supported a very similar conclusion regarding the calculated binding free energy results: ΔG_{13AES3}: −69.38 ± 2.14 kcal/mol, ΔG_{13AES2}: −50.85 ± 2.85 kcal/mol, ΔG_{13AES1}: −42.76 ± 2.35 kcal/mol, respectively. The computational results suggest that the chlorophenyl ring of 13AES3 at C-2 helps to stabilize the binding pose of the

sterane skeleton, namely hydrophobic interaction with the PHE272 residue in the taxene binding site.

It is also known, that the aldo-keto reductase isoforms are key enzymes in steroid metabolism and validated targets in hormone-dependent cancers. The development of selective inhibitor molecules to the AKR1C isoforms is rather challenging. Especially that of AKR1C1 and AKR1C2 is difficult, as these enzymes differ by only a single amino acid residue in the active site. So, the **second line** of my studies is the investigating the EM1062 and EM1082 possible binding modes to AKR1C1 and AKR1C2 isoenzymes. Molecular docking calculations were performed to gain a molecular-level insight into the possible binding mode of two modified steroids. 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. Docking results indicate that both ligands exhibit stronger binding affinity for the AKR1C2 isoenzyme compared with AKR1C1 (according to Figure 1.).

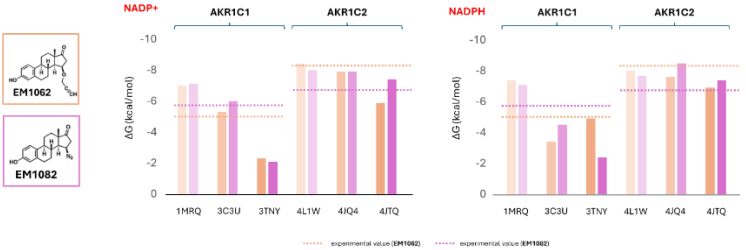

Figure 1. Docking scores of modified steroids into AKR1C1 and AKR1C2 binding sites using different (AKR1C1 PDB codes: 1RMQ, 3C3U, 3TNY and AKR1C2 PDB codes: 4L1W, 4JQ4, 4JTQ) crystall structures. The dotted line represent the experimental binding free energies (calculated from inhibition %).

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%).

In summary, computational investigation 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.

Following this, in the frame of the **third line**, the direct binding of KYNA and selected derivatives (SZR72, SZR104) to SAs was investigated to improve the pharmacological profile of these compounds while preserving biological activity.

According to US results, KYNA binds more strongly to BSA protein than to HSA, which supports the use of BSA in experiments. 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 (C-3) containing derivative SZR104, it presented weak binding to both SAs with no or only a weak preference order. Experimental investigations, like ITC measurements; BSA_{KYNA} **K_d = 158.9 ± 81.4 µM** and to HSA_{KYNA} is **K_d = 5.7 ± 6.0 µM** dissociation constants. However, the results of progressive structural modifications (in case of SZR72 and SZR104) they showed systematic reduction in affinity **1 mM < K_d** (above the detectable range). Therefore, to clarify the binding of these molecules we performed STD-NMR experiments. In our STD-NMR experiments, KYNA showed the strongest albumin binding, whereas SZR72 displayed weaker, but still detectable interaction, and the derivative with combined structural modification (SZR104) exhibited only minimal or no detectable binding.

The **fourth** research **direction** of my PhD studies involved the application of ML techniques to various

peptide-related questions. The first focused on variations in T-cell activation that may arise from structural differences among interacting peptides, which are presentated by epitope–MHC-I complex on the cell surface. In thesis book, I would like to focus on one evaluation metric for classification, namely the *balanced accuracy* (bACC), which is particularly valuable in real-world scenarios where unbalanced datasets are common (if the bACC value is 1, then the classification model is perfect). In the case of machine learning regression, I focus on the coefficient of determination (R^2), which indicates the quality of the AI model's fit (if R^2 equals 1, the regression model is perfect).

The consensus-based (30% hold-out) validation results showed strong overall performance where $bACC_{FFNN}$, $bACC_{RF,}$, $bACC_{SVM}$, $bACC_{LR}$ are around 0.9 or higher, indicating robust classification of both active and inactive epitopes. But the main challenge of AI methods remains their applicability to external datasets. The classification in this case conducted on an independent external dataset compromising 85 unseen epitopes. The 1st external test set supported the transferability of the consensus-based approach, although clear classifier-dependent differences

were observed. The results are the following: $bACC_{FFNN}$: 0.805, $bACC_{RF,}$: 0.804, $bACC_{SVM}$: 0.562, $bACC_{LR}$: 0.718.

To further assess the robustness and cross-dataset generalizability of the models, I evaluated their performance on a 2nd external test set. The 2nd external test set contains 172 additional epitopes reported by different research group and measured in three altered T-cell receptors (TCR1, TCR2, and TCR3). Thus, the 2nd external test set represented a more challenging validation scenario, as reflected by the lower and more variable performance across classifiers. Namely: $bACC_{FFNN}$: 0.497±0.037, $bACC_{RF,}$: 0.618±0.074 , $bACC_{SVM}$: 0.615±0.008, $bACC_{LR}$: 0.508±0.023. 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, and experimental variability on cross-dataset model transferability.

The **fifth line** of my research during my PhD is connected to the prediction of nucleation score (NS) of Aβ mutations (Aβ1–42), using ML algorithms. 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. 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 sets can provide more robust ML results, I focused on the dAA dataset. Model performance was quantified using R^2. The model trained on the dAA dataset (~15,000 data points), using 70/30% split in 5-fold cross validation and 30% hold-out (internal) validation and tested on sAA (~500 data points) external test.

The FFNN consistently achieved the highest predictive performance across almost all the investigated sequence representations (30% hold-out test: R^2: 0.737 and sAA external test: R^2: 0.580 in **ESM-2**). 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.

The consensus model combined the predictions of **ESM-2–FFNN, ESM-2–SVR,** and **Z-scale–RF**, whereas these have the highest predictive performances. The consensus model achieved the highest validation performance, producing an R^2 value of **0.781** while originally the highest performance in the validation test (ESM-2: FFNN) was R^2 value of 0.737. These results indicate that incorporating the three picked models improved prediction accuracy on data originating from the same distribution as the training set. While in the external test (sAA) case it performed R^2 of **0.582**.

4. Conclusions

During my PhD studies, I combine several computational and experimental approaches to clarify 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.

Docking and interaction analyses on the AKR1C isoenzymes further demonstrated that both ligands (EM1062 and EM1082) 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 geomertry, particularly the Leu54/Val54 substitution.

The computational and experimental results of KYNA and its derivatives binding properties to HSA and BSA suggest a clear structure dependent trend; namely the C-2 and C-3 modification of the parent compound (KYNA) resulting weaker complex to the SAs.

In the ML-based investigations of MHC-I. epitope activation study, the results did not support a single universally classifier across all validation settings. Instead, FFNN, RF, LR, and SVM each contributed complementary predictive strengths depending on the external test scenario.

Finally, the Aβ1-42 peptide mutational investigations suggest that the applied consensus model itself proved informative computational perspective. Use of diverse regression paradigms (FFNN, SVR, RF) and wide range of peptide representations, improved generalization of the internal consensus model to predict NS values of Aβ

mutations while application on the sAA external test it generalized weakly.

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.
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.
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.

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.

5. ML algorithms (FFNN, SVM, and RF) were used to predict the NS values of Aβ1-42 peptide double mutated 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.

Publications to serve as the basis 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**

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) 2026;

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) 2026;