Evaluation of Cell Permeability in Macrocyclic Peptides

An Exploratory Computational Chemistry Approach

Honia Rasul

Master's thesis
2025:E55

Lund University

Faculty of Science
Centre for Mathematical Sciences
Applied Computational Science

Centrum Scientiarum Mathematicarum

Evaluation of Cell Permeability in Macrocyclic Peptides

An Exploratory Computational Chemistry Approach

Honia Rasul

LUND UNIVERSITY
Faculty of Science

Thesis supervisor: Ingemar André
Thesis examiner: Stefan Olin

Date: June 16, 2025

Abstract

Macrocyclic peptides (MCPs) are a promising drug modality for targeting protein–protein interactions, offering high selectivity, metabolic stability, and potential cell permeability. However, predicting their membrane permeability remains challenging due to their conformational flexibility, environment-dependent polarity, and ability to adopt chamaleonic behavior, e.g. by changing the number of intramolecular hydrogen bonds. These features, combined with their frequent violation of Lipinski's Rule of Five, make traditional small-molecule models poorly suited for MCPs. This exploratory thesis investigates whether computational tools can accurately predict permeability-relevant properties of MCPs and how well these predictions align with experimental data. Machine learning models were developed using the CycPeptMPDB database and trained on both 2D and 3D molecular descriptors. A curated subset of 60 matched MCP pairs, differing slightly in structure but showing significant differences in permeability, was analyzed to evaluate descriptor–permeability relationships. To bridge the gap between computational predictions and experimental validation, two matched pairs (four peptides) were synthesized. Their logP values were determined using the shake-flask method, and their conformations were analyzed in polar and apolar solvents using NMR spectroscopy. These experimental data, provided by colleagues at the company, were compared with structures generated from molecular dynamics (MD) simulations. The results showed that key descriptors such as logP, PSA, and the presence of intramolecular hydrogen bonds strongly correlate with permeability, but also that static descriptor calculations often fail to capture stereochemical effects or dynamic folding behavior. MD simulations and structural shape metrics provided more nuanced insights into conformational differences affecting permeability. Overall, this exploratory study demonstrates that integrating machine learning, dynamic modeling, and experimental validation offers a powerful approach for understanding and predicting the permeability of MCPs, advancing their rational design in the beyond-rule-of-five chemical space.

Keywords: Descriptors, Machine learning, Macrocyclic peptides, Membrane permeability, Molecular dynamics

Populärvetenskaplig sammanfattning

Makrocykliska peptider är en ny typ av läkemedelskandidater som kan påverka mål i kroppen som vanliga läkemedel ofta inte kommer åt – till exempel när två proteiner ska interagera. Trots att dessa peptider är relativt stora och komplexa, har vissa visat sig kunna tas upp i kroppen vid tablettintag, något som annars är ovanligt för större molekyler. Men det är fortfarande oklart vilka egenskaper som gör att vissa av dessa peptider kan ta sig igenom cellens skyddande membran och bli aktiva i kroppen. Syftet med det här explorativa examensarbetet var att undersöka om datorbaserade metoder kan hjälpa till att förutsäga vilka makrocykliska peptider som har god förmåga att passera cellmembran. För att göra detta analyserades nästan 7000 olika peptider med hjälp av maskininlärning och molekyldynamiska simuleringar. Datorn tränades att hitta samband mellan olika egenskaper, som fettlöslighet, förmåga att binda till vatten (polär yta) och inre vätebindningar, och hur lätt peptiderna tar sig in i celler. För en mer detaljerad analys valdes 60 peptidpar ut där varje par bestod av två nästan identiska molekyler med olika förmåga att ta sig in i celler. Det visade sig att små strukturella förändringar kunde påverka hur peptiderna veckade sig och hur de interagerade med sin omgivning, faktorer som i sin tur påverkar deras upptag i kroppen. Fyra av peptiderna tillverkades i labb av kollegor på företaget och testades med experimentella metoder, där resultaten stämde väl överens med datorns förutsägelser. Slutsatsen är att en kombination av datorberäkningar och experimentella tester kan ge en bättre förståelse för hur framtidens läkemedel kan utformas för att fungera effektivt i kroppen, även när de bryter mot de klassiska reglerna för läkemedelsdesign.

Acknowledgments

I would like to express my sincere gratitude to my supervisor at RG Discovery Ricardo Ferreira for his valuable guidance, continuous support, and encouragement throughout this project. His scientific insight and constructive feedback have been crucial to the development of this thesis. I would also like to thank my colleagues at RG Discovery for providing a welcoming and inspiring research environment. A special thanks to the different teams that were part of the experimental part of this thesis for their expert contributions, and to everyone who took the time to answer questions, discuss ideas, or offer technical help. I am grateful to Ingemar André, my supervisor at the department of chemistry, for his time and helpful comments during all stages of this thesis. I would also like to express gratitude to my examiner Mikael Lund.
Finally, I would like to thank my friends and family for their support, patience, and motivation throughout my five years of studies.

Lund, June 2025
Honia Rasul

Contents

Chapter 1

Introduction

Drug discovery continues to evolve in response to emerging diseases and increasing resistance mechanisms. Small molecules, that account for approximately 90% of today's medicines [36], are essential in drug development but they often fail to engage so-called "undruggable" targets, particularly protein–protein interactions. These interactions play central roles in key biological processes such as enzymatic activity, hormone signaling, immune responses, and cellular communication. As a result, there is a growing interest in alternative drug modalities capable of addressing these complex interfaces. One such modality is macrocyclic peptides (MCPs), which have emerged as promising candidates due to their high binding specificity, chemical stability, cell permeability, and suitability for synthesis. However, their complex properties complicate prediction of behavior across biological membranes. Modeling the permeability of MCPs using *in silico* methods remains a complex task. Their intrinsic properties, such as conformational flexibility, polarity changes depending on the environment, and intramolecular hydrogen bonds (IMHB), which can reduce exposed polarity and enable membrane crossing, introduce challenges that are not typically encountered with small molecules. Although current computational methods are well-established for small molecule drug discovery, they tend to fail when branched out to MCPs or other molecules lying beyond the known small molecule chemical space. This underlines the need for more tailored approaches that can predict both the conformations and permeability-relevant properties of MCPs with higher accuracy.

To address these challenges and better understand what drives membrane permeability in MCPs, one of the most critical parameters for the development of orally available peptide drugs, this thesis investigates both computational and experimental perspectives. The study begins by evaluating which molecular descriptors are most relevant for predicting permeability, using machine learning models trained on a large dataset of 6986 MCPs from CycPeptMPDB, one of the largest publicly available databases of MCPs [32]. From this dataset, a subset of 30 matched MCP pairs were manually selected for further analysis. Within these pairs, the two MCPs have minor structural differences, such as stereochemical inversion of one amino acid or *N*-methylation on one amide, but were their experimental permeability values were significantly different. An example of a matched pair is visualized in Figure 1.1

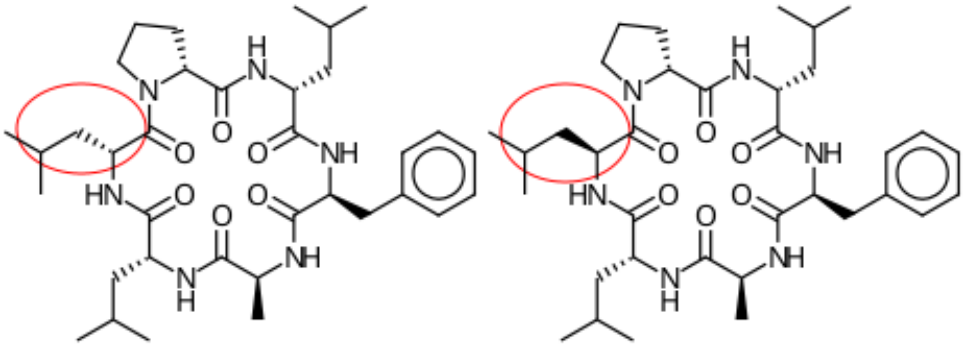


Figure 1.1: Example of a MCP pair differing by a single stereocenter (red circle). Despite the minor structural change, these peptides showed significantly different membrane permeability.

This difference in permeability within the pairs was one aspect that was of interest for this project to try to understand and evaluate. Descriptor calculations, based on literature-reported correlations with permeability in orally bioavailable drugs, were performed using multiple computational tools. Finally, two matched pairs (four MCPs in total) were synthesized and experimentally characterized, selected based on raw material availability and practical time constraints, by colleagues at the company. Their conformational behavior in both polar and apolar environments was assessed using NMR spectroscopy and *in silico* modeling (performed by colleagues at the company), offering a reference for evaluating the accuracy of computational predictions. Machine learning and 3D descriptors are essential because MCPs explore a wide range of conformations that static rules and 2D metrics fail to capture.
Based on the above described objectives the following research questions were formulated:

(i) How accurately can current computational tools predict the conformational landscape and permeability-related physicochemical properties of MCPs?

(ii) Do permeability descriptors developed for small molecules apply to more complex MCPs?

(iii) How can existing computational tools be adapted or improved to better predict 3D structures and passive membrane permeability in structurally complex MCPs?

1.1 Macrocycles and macrocyclic peptides

Macrocycles are a broad class of compounds defined by a ring structure containing 12 or more atoms [52]. They have attracted increasing interest in pharmaceutical research due to their ability to bind complex and otherwise undruggable targets with high affinity and selectivity. This is often attributed to their conformational rigidity and capacity to form IMHBs, which help stabilize bioactive conformations. In contrast, linear analogs are more flexible, adopting a wide range of conformations that can increase entropy and reduce binding specificity [54].

MCPs are a subclass of macrocycles characterized by a closed-loop backbone formed through linkages between amino acid residues [35]. In addition to the general advantages of macrocycles, their peptide-based nature often results in low or no toxicity and enables efficient synthesis using automated solid-phase peptide synthesis methods [60]. Their structural flexibility allows

for chemical modification and facilitates both characterization and handling. A particularly important property of some MCPs is their *chameleonic behavior*, the ability to adjust conformation and polarity in response to the surrounding environment through dynamic formation of IMHBs [50]. While this behavior can enhance membrane permeability, it also makes conformational prediction more difficult. In this study, matched pairs of MCPs were analyzed that exhibit significant differences in permeability despite minor structural variations. These modifications, such as *N*-methylation or inversion from L- to D-amino acids, are known to impact properties like polarity, flexibility, and hydrogen bonding, and have been shown to influence permeability [6].

As interest in MCPs continues to grow, so does the need for curated datasets that support their computational analysis and design. To support structure–property modeling and ensure broad chemical diversity, this thesis leverages the CycPeptMPDB database, which contains 7,991 structurally diverse cyclic peptides with experimental permeability data and a wide range of computationally calculated molecular descriptors [32]. This dataset provides the foundation for the machine learning and modeling approaches used throughout the study.

1.2 Permeability

Permeability refers to a compound's ability to cross biological membranes and thus become bioavailable in the body [14]. In 1997, Lipinski introduced the widely used Rule of Five (Ro5), which proposes that orally active drugs should meet at least three of the following criteria: molecular weight $\leq$ 500 Da, hydrogen bond donors $\leq$ 5, hydrogen bond acceptors $\leq$ 10 and calculated LogP $\leq$ 5 [33]. Many MCPs fall beyond these limits and are thus classified as beyond Rule of Five (bRo5) compounds. Typically, bRo5 compounds suffer from poor solubility, limited permeability, and unfavorable pharmacokinetic properties. However, MCPs are a notable exception: despite their size and polarity, some exhibit surprisingly high membrane permeability, often attributed to their chameleonic behavior and ability to form IMHBs that reduce effective polarity. As illustrated in Figure 1.2, this phenomenon has sparked growing interest in MCPs as orally bioavailable drug candidates [29]. Nonetheless, the factors governing MCP permeability remain poorly understood, posing a significant challenge for rational design and prediction.

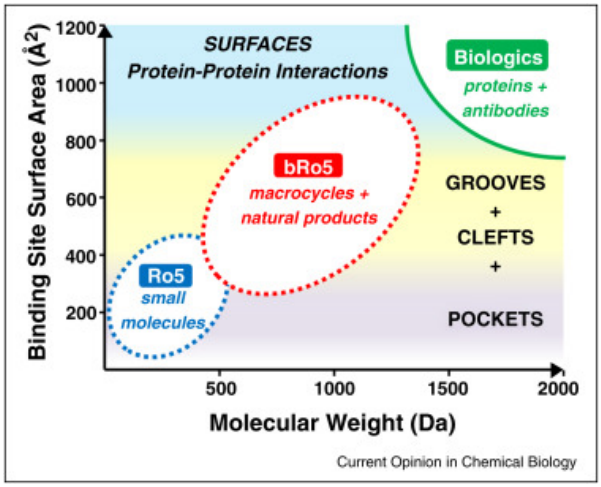


Figure 1.2: Illustration of pharmaceutical opportunity for bRo5 drugs to target larger biological binding sites not accessible to conventional small molecules. Retrieved from [42]

Oral bioavailability is a desirable property for drug candidates, as it improves treatment compliance and patient accessibility [10]. This has intensified interest in the development of orally bioavailable MCPs and in computational tools that can reliably predict their permeability. To evaluate permeability, one must consider a compound's ADME (absorption, distribution, metabolism, and excretion) profile. A widely used in vitro method for assessing passive permeability is the parallel artificial membrane permeability assay (PAMPA), which mimics membrane transport across a phospholipid layer separating two aqueous compartments. The compound is placed in one compartment, and after incubation, concentrations in both compartments are measured to estimate a permeation rate, typically reported in cm/s [2]. Another version of this method is the Caco-2 Cells permeability assay. The Caco2 is a human colon epithelial cancer cell line which is used to model the human intestinal permeation. [5]

1.3 Descriptors used to assess permeability

A variety of molecular descriptors have been proposed to estimate membrane permeability, but some consistently show stronger correlations with experimental data based on literature precedent and relevance to MCPs:

Polar Surface Area (PSA) refers to the total surface area occupied by polar atoms, typically nitrogen and oxygen, along with their attached hydrogens [45]. It is widely used as a predictor of passive permeability, as high PSA reflects greater polarity and hydrogen bonding with water, which hinders membrane diffusion. Another reason for it being used is due to the many computational methods developed to accurately predict PSA and its variants; topological polar surface area (TPSA) and effective polar surface area (ePSA). PSA is also applicable in many permeability models, such as PAMPA.

Solvent-Accessible Surface Area (SASA) is a geometric measure of the surface area of a molecule that is accessible to solvent molecules [12]. It reflects how a compound interacts with its environment and is frequently used to assess membrane permeability. For MCPs, SASA is

particularly informative because these compounds can adopt conformations that shield polar atoms from solvent exposure via IMHBs. This reduction in solvent-exposed polar area lowers the desolvation penalty during membrane translocation, a key step in passive diffusion. As a result, MCPs with lower polar SASA values often show enhanced permeability, even when their total polarity appears high in static descriptors. Incorporating SASA into permeability models is therefore especially valuable for capturing the dynamic, chameleonic behavior of MCPs [11].

Partition Coefficient (logP) quantifies the distribution of a compound between a hydrophobic (typically octanol) and a hydrophilic (water) phase, serving as an indicator of lipophilicity [31]. Molecules with higher logP values tend to favor lipid-rich environments, such as the hydrophobic core of cell membranes, which may facilitate passive diffusion and often correlates with higher membrane permeability [17]. For MCPs, logP remains a useful descriptor, although its interpretation can be complicated by conformational flexibility and the potential for polarity shielding. MCPs can exhibit lower apparent polarity through IMHBs, which may allow them to cross membranes despite having logP values that would otherwise suggest poor permeability. In cases where compounds possess ionizable groups, the *distribution coefficient* (logD) is used instead; it reflects the ratio of all species, ionized and un-ionized, between octanol and water at a specific pH. For non-ionizable compounds like those in this study, logD and logP are equivalent, as the molecules remain neutral across physiological pH ranges. [24].

Several computational methods exist for estimating logP, including calculated logP (cLogP), which uses fragment-based values; atom-based logP (ALogP), based on atomic contributions; and XlogP, which introduces correction factors for features such as hydrogen bonding and aromaticity [18, 55]. Together with PSA, logP remains one of the most widely used descriptors in permeability prediction models.

As mentioned previously, *intramolecular hydrogen bonds* (IMHBs) are considered an important factor in influencing membrane permeability, especially for polar or conformationally flexible compounds such as MCPs [1]. By forming hydrogen bonds between polar functional groups within the same molecule, IMHBs can reduce solvent exposure of these groups, thereby lowering the molecule's apparent polarity. This may increase its effective lipophilicity and reduce the energetic cost of desolvation, both of which can facilitate passive diffusion across lipid membranes.

In MCPs, the formation of IMHBs is often dependent on the surrounding environment and contributes to their so-called chameleonic properties, adopting less polar conformations in non-polar environments while maintaining solubility in aqueous conditions [44]. As such, IMHBs may help explain how some MCPs achieve membrane permeability despite having physicochemical properties that would not typically predict it based on standard descriptors alone.

1.4 Physical composition and crossing of the cell membrane

MCPs *3D conformation* plays a critical role in their membrane permeability. Unlike rigid small molecules, MCPs are conformationally flexible and can adopt a range of low- and high-polarity structures depending on their environment. These conformational changes influence how polar groups are exposed or internally shielded, thereby affecting both desolvation energy

and lipophilicity. Studies have shown that MCPs are capable of forming compact, low-polarity conformations that often are stabilized by IMHBs, and exhibit significantly higher passive permeability across membranes. [34, 15].

Experimental and computational investigations confirm that even minor modifications, such as N-methylation or stereochemical inversion, can shift the conformational ensemble and result in measurable differences in permeability. Conformational datasets like CREMP have further emphasized that relying on static descriptors alone, such as PSA or logP, is often insufficient for capturing the dynamic nature of MCPs [19]. As such, 3D conformation should be viewed as a dynamic descriptor that underpins how MCPs interact with the lipid bilayer and determines their potential for passive diffusion.

To better capture the structural features relevant to permeability, this study evaluated both 2D and 3D molecular descriptors. While 2D descriptors like PSA and MolLogP provide useful estimates of polarity and lipophilicity, they do not account for conformational flexibility. Therefore, several 3D descriptors were also included. Among these, *autocorrelation descriptors* received particular attention. These quantify how atomic properties, such as mass or electronegativity, are distributed across the structure by evaluating their relationships over different atomic distances. This is especially useful for flexible molecules like MCPs, where local interactions and folding patterns influence exposure to the solvent.

Other 3D descriptors included in this study were used to capture mass distribution and spatial arrangement. For instance, GRAV-1 encodes how heavy atoms are distributed throughout the molecule [25], while TDB1e and TDB1m measure how electronegativity and mass correlate across short topological distances, respectively [28]. Additionally, Van der Waals surface area (VSA) descriptors estimate atomic surface exposure based on atomic radii and bond lengths [30], and Electrotopological State (EState) indices reflect how each atom's electronic environment is shaped by the molecule as a whole [20]. Combinations of these, such as VSA_EState9, provide insight into how spatial and electronic features together influence permeability. Collectively, these descriptors help bridge the gap between static molecular structure and the dynamic, environment-dependent behavior that governs passive membrane diffusion in MCPs.

1.5 Computational Chemistry and Molecular Dynamics

Numerous computational chemistry libraries and software tools are available for predicting the permeability and conformational behavior of MCPs. Many of these are implemented in Python, which offers a flexible, readable, and well-supported environment for scientific computing. Its extensive ecosystem of molecular modeling libraries, along with reproducibility-friendly workflow design, makes it particularly well-suited for building custom pipelines for permeability prediction and descriptor generation.

While molecular descriptors derived from static structures can provide useful approximations of membrane permeability, they have important limitations, especially for flexible compounds like MCPs. Static descriptors often fail to capture conformational variability, solvent responsiveness, and time-dependent behavior. Molecular Dynamics (MD) simulations address this

gap by modeling the motion of atoms over time using classical force fields. By generating trajectories that reflect how a molecule explores its conformational space in different environments, such as water or membrane-like solvents, MD simulations enable the evaluation of dynamic features such as IMHBs, conformational shielding, and SASA.

In the context of MCPs, MD simulations offer critical insight into structural adaptability, revealing how environment-dependent conformations may enhance membrane permeability. As such, MD simulations serves as a valuable complement to static descriptor-based models, providing a more realistic representation of molecular behavior during passive diffusion across lipid bilayers.

1.5.1 NMR Spectroscopy for Structural Validation and Evaluation

Due to the conformational flexibility and environment-dependent behavior of MCPs, structural validation through experimental techniques is critical. In this project, Nuclear Magnetic Resonance (NMR) spectroscopy was used by colleagues at the company to characterize the conformations of selected MCPs in both polar and apolar solvents. NMR enables comparison between experimental observables, such as chemical shifts, Nuclear Overhauser Effect (NOE), and/or scalar couplings, and computationally predicted structures. NOEs are especially informative for conformational analysis, as they arise from magnetization transfer between spatially close nuclei. The intensity of NOE cross-peaks, observed at short mixing times in NOESY experiments, is proportional to the inverse sixth power of the interproton distance:

$$I_{\mathrm{NOE}} \propto \frac{1}{r^6} \tag{1.5.1}$$

In practice, NOEs are typically observed for proton pairs within 5 Å. This makes them particularly valuable for studying the compact, environment-sensitive conformations adopted by MCPs. [8] Detailed experimental conditions and spectral analysis procedures are described in the Methods section. While the project includes experimental synthesis, purification, and NMR-based structural characterization of selected peptides, these procedures were carried out by the Peptide Chemistry, Analytical, and NMR teams at the host company. My primary role in the project focused on computational modeling, including the use of NOE-derived distance restraints to generate and validate conformational ensembles through MD simulations.

Chapter 2

Material and method

2.1 Computational Tools and Techniques

This study used a multi-step computational strategy to investigate the permeability of macrocyclic peptides (MCPs) using both descriptor-based modeling and structure-based simulations. The workflow integrated cheminformatics, molecular modeling, and machine learning to evaluate whether physicochemical descriptors and solvent-sensitive conformations could predict passive membrane permeability.

The following sections detail the tools and procedures used for: (i) handling peptide structures and generating descriptors; (ii) training machine learning models on a large MCP dataset; (iii) constructing and analyzing a manually curated subset of 60 matched peptides; and (iv) performing Molecular Dynamics (MD) simulations and free energy calculations to evaluate solvation behavior. Together, these components form the basis for assessing permeability trends in both large-scale data and experimentally validated peptide systems. The experimental synthesis and structural analysis presented in this study were performed by company collaborators and are briefly summarized for context. My contributions were limited to the computational components of the project, including structure generation, descriptor analysis, and the use of NOE-derived distances in MD simulations for conformational evaluation.
Together, these computational and experimental workflows formed the foundation for evaluating permeability-related features across the full dataset, matched peptide pairs, and synthesized MCPs.

2.1.1 Peptide Structure Handling and Descriptor Generation

Peptide Sequence and Structure Handling

pyPept [43, 16] is a Python library used for peptide sequence processing, molecular structure conversion, and conformer generation, supporting both 2D and 3D representations. A key feature is its ability to incorporate customized monomer libraries, including both natural and modified residues, an essential capability for this study. Additionally, *pyPept* also supports HELM notation [59], which offered a clearer alternative to SMILES for MCPs containing modified amino acids, as these could be built directly within the HELM framework.

Molecular Property Calculations

To develop predictive models concerning permeability, molecular descriptors were calculated for the full CycPeptMPDB dataset. A total of 200 different 2D descriptors were generated using *RDKit* [46] through the DescriptaStorus python package [26], while 458 different 3D descriptors were obtained using *PaDEL-Descriptor* [58] and *Mordred* [41]. Regression models were trained and evaluated in *PyCaret*, using continuous experimental permeability values (logPeff) as the target. The model with the highest coefficient of determination (R^2) was selected for further analysis. *OpenBabel* and *DataWarrior* were used for molecular format conversion and conformer generation, respectively. Similar to *pyPept*, conformers generated with DataWarrior were constructed in vacuum using its internal low-energy bias algorithm, without inclusion of solvent effects. Additional tools included *ChemProp* [21], which employs message-passing neural networks to train models able to predict molecular properties, and *ABSOLV* [4], a Python API for solvation free energy calculations based on thermodynamic cycles using the OpenMM framework [13]. Finally, *Melodia_py* [40] was used to compute Differential Geometry and Knot Theory descriptors such as curvature and torsion for a selected peptide subset. Concerning logD, a ML model using AstraZeneca public logD data at pH 7.4 ($N = 4.351$) [57] was used for predicting MCPs logD (external validation set, R^2, 0.710; RMSE = 0.638). Two other permeability prediction models, trained from publicly available PAMPA ($\log P_{eff}$, $N = 14.801$, validation set: $R^2 = 0.98$, RMSE = 0.440) and Caco-2 ($\log P_{app}$, $N = 2.799$, validation set: $R^2 = 0.935$, RMSE = 0.768) data, were used to infer the MCPs permeabilities. All three models were trained using *ChemProp*.

Molecular Dynamics and Visualization

MD simulations were generated to extract dynamic properties relevant to permeability, such as SASA, cPSA (calculated PSA), and the number of IMHBs over time. This analysis was performed using *Cpptraj* [48], its Python interface *Pytraj* [49], and *MDtraj* [38]. Structure drawing and visualization were carried out using *ChemDoodle* [7], *VMD* [22] or *PyMol* [51].

The computational workflow was structured in three sequential parts to systematically evaluate permeability prediction in MCPs:

(i) Development of general regression models using the full CycPeptMPDB dataset and either 2D or 3D molecular descriptors.

(ii) Computational analysis of 30 matched peptide pairs with similar structures but varying permeability values, to assess descriptor-permeability relationships.

(iii) Selection and experimental validation for some of the most representative matched pairs, including synthesis, permeability measurements, and structural comparison with *in silico* predictions.

2.1.2 Machine Learning Workflow on Full Dataset

The following steps summarize the workflow used to build predictive models from the full CycPeptMPDB dataset, with all calculations performed in Google Colab:

(i) The dataset was filtered to exclude molecules with invalid characters (e.g., <, >) or permeability values outside the assay detection range. Only permeability values derived from PAMPA assays were retained, resulting in a final set of 6,986 peptides from the original 7,991.

(ii) Feature selection and outlier removal were performed in *PyCaret*, using permeability as the target. The resulting data shapes were (6741, 41) for 2D descriptors and (6741, 92) for 3D descriptors using the command

```
s = setup(df, target='permeability', session_id=randrange(1000),
feature_selection=True, remove_outliers=True)
```

(iii) Then, all available models were then compared using

```
best = compare_models(exclude=['lar','par','dummy','lr'])
```

The CatBoost model, which is a regression model based on gradient boosting decision trees, achieved the highest $R^2 = 0.6268$ for both descriptor sets and was therefore selected for the analysis of the dataset. Model performance of the Catboost model was evaluated using prediction error, feature importance, and residual plots.

2.1.3 Matched Pair Selection and Analysis

The second part of the workflow focused on 30 matched peptide pairs selected for their minimal structural differences and varying permeability. Selection and preparation of the 60 peptides used in this analysis required extensive manual effort. The matched pairs were not pre-defined but identified through a systematic, manual review of the full CycPeptMPDB dataset (∼7,000 peptides), focusing on subtle structural modifications linked to permeability shifts. This process included visual inspection, monomer-level comparison, and literature-informed curation to ensure biological relevance. The entire process, including structure building, monomer generation, and integration into the modeling pipeline, spanned approximately two months and represented one of the most labor-intensive parts of the thesis.

(i) From the CycPeptMPDB database, a subset of 30 pairs (60 peptides) was selected, with permeability values in the range of –4 to –8, reflecting relevance to passive permeability studies. The matched pairs were curated manually and all pairs with low % recovery (%R < 10) [37] or low % transport (%T < 2.0) [23] were removed.

(ii) The structures were examined to identify monomers not present in the *pyPept* library. These missing monomers were manually built in *ChemDoodle*, with attachment points defined at the N- and C-termini. The resulting SMILES strings were converted to 2D structure files using *OpenBabel* in an adequate format (output-ID, sdf) format:

```
obabel [-:"SMILES"] [-o <output-ID>] [-O outfile] [OPTIONS]
```

(iii) Each new monomer was assigned a unique HELM identifier and entered into the HELM Web Editor to generate corresponding HELM strings. These were processed with the *map_monomers.py* script:

```
map_monomers.py --input 'new_monomer.sdf' --output 'new_monomer_processed.sdf'
```

and manually added to the *pyPept* monomer library.

(iv) 3D conformations of all peptides were then generated using *pyPept*:

```
run_pyPept --helm 'HELM_string'
```

The 3D conformers generated by *pyPept* were constructed in vacuum, as the tool does not apply any implicit or explicit solvation model during conformer generation. Thus, no solvent effects were considered at this stage.

(v) Molecular properties such as logP and TPSA were calculated using RDKit via: *rdkit.Chem .Crippen.MolLogP* and *rdkit.Chem.rdMolDescriptors.CalcTPSA*.

(vi) Solvent-accessible Polar Surface Area (fPSA) was additionally calculated with *FreeSASA*:

```
freesasa peptide_#.pdb
```

2.1.4 Molecular Dynamics Simulations

Solvation and free energies differences were estimated using ABSOLV, an open-source Python implementation of non-equilibrium alchemical transitions. For a final set of five matched pairs (ten MCPs), cubic simulation boxes were constructed, each containing one MCP and either 600 water molecules or 125 1-octanol molecules. Octanol and water was selected here to mimic a apolar and polar solvent, respectively. Systems were parameterized using the OpenFF 2.2.0 force field [13].

All MD simulations were performed using Langevin Dynamics, with a timestep of 2 fs, at a temperature of 298.15 K and pressure of 1 bar. Simulations were run for 1 million steps in vacuum and 2 million steps in solvent. Only the second half of each trajectory was used, assuming equilibration in the first half.

For solvent environments, alchemical decoupling was performed in two stages: electrostatic interactions were annihilated over the first 12 ps, followed by van der Waals (vdW) decoupling over the next 38 ps. In vacuum simulations, only the fully decoupled state was considered. Equilibrium configurations were generated for both the solvated and decoupled endpoints of the thermodynamic cycle.
The work W along each non-equilibrium transition path was computed as:

$$W = \sum_{i=0}^{n-1} \left[u_{\lambda_{i+1}}(x_i) - u_{\lambda_i}(x_i) \right] \tag{2.1.1}$$

$$W = \sum_{i=0}^{n-1} \left[u_{\lambda_{i+1}}(x_i) - u_{\lambda_i}(x_i) \right] \tag{2.1.2}$$

Free energies (ΔG) were estimated by solving the Crooks fluctuation theorem equation:

$$\sum_{i=1}^{N} \frac{1}{1 + \exp(\beta(W_i^f - \Delta F))} = \sum_{i=1}^{N} \frac{1}{1 + \exp(\beta(W_i^r + \Delta F))} \tag{2.1.3}$$

where W_i^f and W_i^r are the forward and reverse work values, respectively. and N refers to the total number of equilibrium snapshots. The transfer free energy from water to 1-octanol was computed using a thermodynamic cycle

$$\Delta G_{A->B} = \Delta G_1 - \Delta G_3 \tag{2.1.4}$$

while the partition coefficient (logP) was estimated as:

$$LogP = \frac{G_s^{wat} - G_s^{oct}}{2.303RT} \tag{2.1.5}$$

where $R = 1.985 \times 10^{-3}\,\text{kcal·mol}^{-1}\text{·K}^{-1}$ is the gas constant and $T = 298.15\,\text{K}$ is the temperature [27].

To enable additional analysis, the *ABSOLV* code was modified to save all MD trajectory files in `.dcd` atomic trajectory format and subsequently used in *Cpptraj*, *Pytraj*, and *MDtraj* to calculate key molecular properties, including PSA, SASA and IMHBs in both water and 1-octanol. Water and 1-octanol were selected to approximate polar and apolar environments, respectively, in line with standard logP-based permeability estimation. While DMSO and $CDCl_3$ were used in experimental NMR studies, they were not used in simulations due to the lack of a corresponding thermodynamic cycle and the standardization of the octanol–water system in permeability prediction.

2.1.5 Experimental Synthesis and Validation

The synthesis, purification, and structural characterization of the selected MCPs were performed by the Peptide Chemistry, NMR, and Analytical teams at the company. Two matched pairs (four peptides) were synthesized via solid-phase peptide synthesis, followed by cleavage, cyclization, and purification using LC-MS. Structural validation was conducted using NMR spectroscopy in polar DMSO-d_6 and apolar $CDCl_3$, with 1D and 2D experiments including TOCSY, NOESY, and HSQC. logD values were determined experimentally using a shake-flask partitioning method followed by LC-UV analysis. My contribution to the experimental workflow was limited to the use of NOE-derived distance restraints from the NMR data to generate and validate conformations through MD simulations.

2.1.6 Computational Analysis of Experimentally Validated Peptides

The following workflow summarizes the computational analysis performed on the experimentally validated peptides:

(i) **NOE-based distance calculations and structure generation**
Proton–proton distances were calculated from the intensities of NOESY cross-peaks using the following equation:

$$r_i = r_{\text{ref}} \left(\frac{S_{\text{ref}}}{S_i} \right)^{\frac{1}{6}} \tag{2.1.6}$$

where r_{ref} is a known reference distance, and S_{ref} and S_i are the integrated cross-peak intensities of the reference and the proton pair of interest, respectively.

NOE-distance-restrained conformations and MD trajectories in both water and chloroform were generated using the default *noeETKDG* [56] protocol, along with the *mlddec* [3], *mdfptools* [47] Python packages. Root-mean-square deviations (RMSD) between conformers were calculated using the *spyrmsd* package [39].

(ii) **Hydrogen bonding and surface area calculations**
IMHBs were calculated and visualized using the *Hbond* class in *Pytraj*, based on MD trajectory files for peptides simulated in water (ABSOLV or noeETKDG), chloroform (noeETKDG), or 1-octanol (ABSOLV). The different solvents are used to mimic cell membrane transportation, as the cell membrane has a polar outer layer and an apolar interior.

SASA and PSA were also computed from the trajectories using the *shrake_rupley* algorithm available in the *MDTraj* package.

(iii) **Conformational comparison via differential geometry descriptors**
To assess structural differences across solvent environments, NOE-restrained conformations in water and chloroform were analyzed using *Melodia_py* [40]. Again, the water and chloroform is used to simulate the inner and outer layer of a cell membrane.

Atomic coordinates were fit to a parametric vector function:

$$\vec{r}(t) = x(t)\hat{x} + y(t)\hat{y} + z(t)\hat{z} \tag{2.1.7}$$

encoding the 3D spatial trajectory of the macrocyclic ring. From this vector, two key differential geometry descriptors—curvature (κ) and torsion (τ)—were computed at each Cα position t:

$$\kappa = \frac{\left| \dot{\vec{r}} \times \ddot{\vec{r}} \right|}{\left| \dot{\vec{r}} \right|^3} \tag{2.1.8}$$

$$\tau = \frac{\left(\dot{\vec{r}} \times \ddot{\vec{r}} \right) \cdot \dddot{\vec{r}}}{\left| \dot{\vec{r}} \times \ddot{\vec{r}} \right|^2} \tag{2.1.9}$$

In addition to curvature and torsion, two other shape descriptors were calculated: arc length and writhing. Arc length was evaluated over a 3-residue-long segment, while writhing was computed for a 5-residue window centered around each residue. The writhing number [9] serves as a geometric measure of spatial twisting and, unlike curvature, can assume both positive and negative values.

Chapter 3

Results

The goal of this study was to evaluate whether computational tools can predict physicochemical properties and conformational features that contribute to the passive permeability of macrocyclic peptides (MCPs), a promising class of drug candidates for targeting protein–protein interactions with high binding specificity. To address this, the study was divided into three parts.
In **Part 1**, multiple machine learning models were trained on a dataset of 6,986 MCPs from CycPeptMPDB, using either 2D and 3D molecular descriptors to predict membrane penetration. The models were evaluated based on performance metrics and visualized using diagnostic plots to identify the most accurate algorithm.
In **Part 2**, a curated subset of 60 MCPs with experimentally measured permeability values was analyzed to investigate descriptor–permeability correlations in more detail.
In **Part 3**, four selected peptides were synthesized and experimentally evaluated to validate key predictions and assess conformational accuracy.
All molecular descriptors were calculated from computationally generated structures. For 3D descriptors, a single conformer was used per peptide unless otherwise stated. Model performance was assessed using R^2 values, residual plots, and feature importance analysis via the CatBoost algorithm.
The synthesis, purification, and experimental evaluation of the selected MCPs were performed by the Peptide, ADME and NMR teams at the company. My role focused on the computational analysis of the resulting data, including structure–descriptor comparison, conformer validation, and integration of NMR-derived distances into modeling workflows.

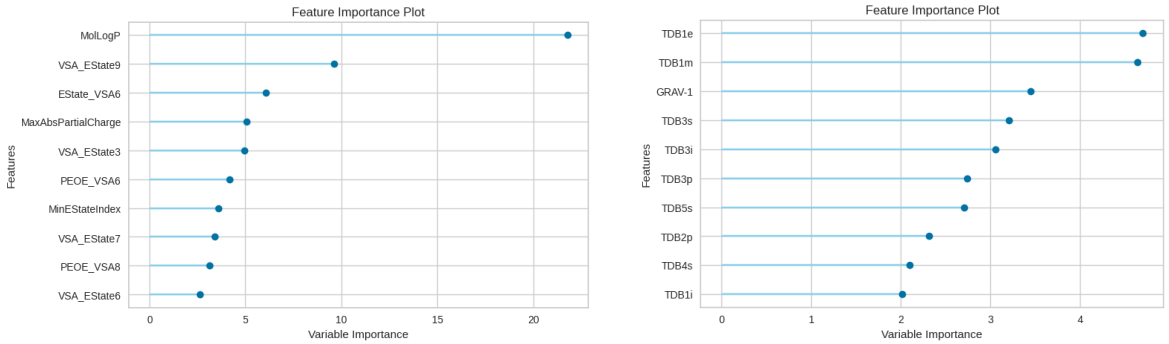
3.1 Computational Approaches

3.1.1 Part 1 – Model Building

This first part, as mentioned previously, aimed to investigate which molecular features are most relevant for passive permeability in MCPs. In order to do this, *Pycaret* was used, which is a machine learning library in Python. The model was trained on the full dataset of 6,986 MCPs with experimentally determined membrane penetration labels (permeable of non-permeable) from CycPeptMPDB. The 200 different 2D descriptors were loaded from RDKit and 458 different 3D descriptors were loaded from Morded and PaDeL, which all capture both physicochemical

and spatial properties. *Pycaret*'s *Compare models* function was used which trains and evaluates the performance of all the estimators using cross-validation. The machine learning models described here were trained using the passive permeability values reported in CycPeptMPDB. Each peptide entry includes a binary label indicating high or low membrane penetration, based on experimental assays, which served as the dependent variable for training. As is common in high-throughput modeling, all 3D descriptors were calculated from a single conformer per peptide, which does not capture dynamic structural variability.

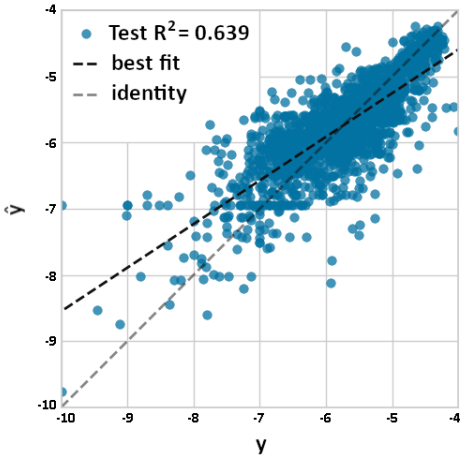
From this analysis, the CatBoost model was ranked as the model with highest performance with the highest R^2 value for both models and therefore selected. The Catboost model is a high-performance gradient boosting algorithm that efficiently handles categorical data, prevents overfitting, works well with small and imbalanced datasets, and offers fast training and inference. From this, a feature importance analysis, of the CatBoost model was conducted for the 2D- and 3D-descriptors separately. The feature importance is an indication of how much each feature contributed to the model prediction, and uses tree-based algorithms, and the plots are visualized in Figure 3.1 For the 2D model, the top-ranked descriptors were MolLogP, VSA_EState9, and EState_VSA6, which reflect lipophilicity and electronic distribution within the molecule. For the 3D model, GRAV-1, TDB1e, and TDB1m were identified as the most important features, reflecting mass distribution and short-range autocorrelations weighted by atomic properties. In order to assess and visualize the model, prediction error plots and residual plots were made and are seen in Figure 3.2 for both models. For the residual plots, the 2D model achieved a training $R^2 \approx 0.790$ and a test $R^2 \approx 0.639$. The 3D model showed a training $R^2 \approx 0.907$ and a test $R^2 \approx 0.543$. The residual plot further indicate that errors are symmetrical distributed around zero, with no signs ofskew or systematic deviation. Regarding the prediction error plots, they align relativery close with the experimental values, and follow the diagonal.



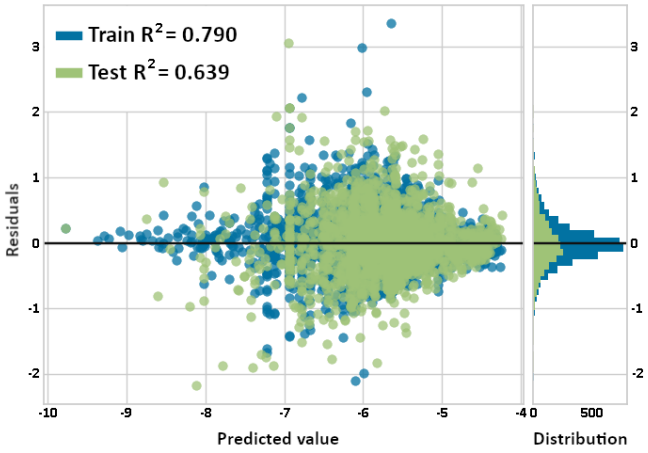
(a) Feature importance for CatBoost model on 2D descriptors

(b) Feature importance for CatBoost model on 3D descriptors

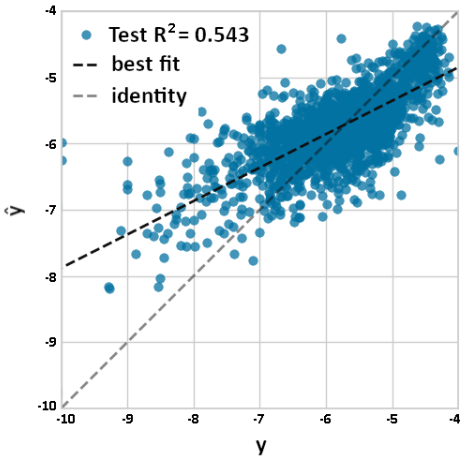
Figure 3.1: Feature importance for CatBoost model trained on 2D (a) and 3D (b) molecular descriptors. Feature importance is derived from the average gain across all decision trees in the model, representing how much each descriptor contributes to reducing prediction error during training. Higher bars indicate features with greater influence on permeability prediction.



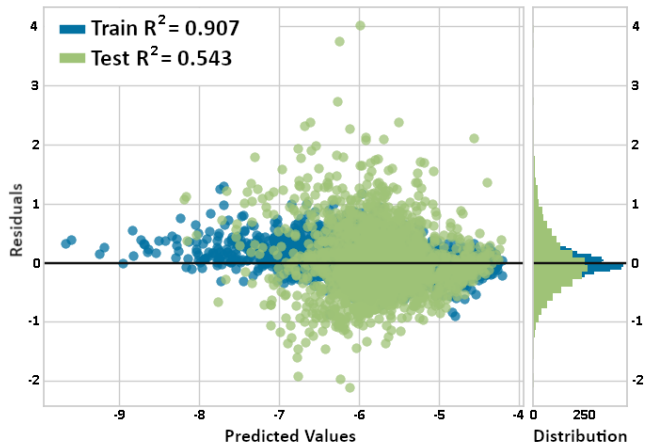
(a) Prediction error plot for 2D descriptors



(b) Residuals for 2D descriptors



(c) Prediction error plot for 3D descriptors



(d) Residuals for 3D descriptors

Figure 3.2: Prediction error and residual plots for CatBoost models trained on 2D (a, b) and 3D (c, d) descriptors. Residuals are symmetrically distributed. In prediction error plots (a, c) the x-axis represents the experimental permeability values (y) and the y-axis represents the model-predicted permeability values ($\hat{y}$). This gives an idea of the variance of the model, and how much the predicted values differ from the experimental permeability. Residuals (b, d) are computed as $\hat{y} - y$, and gives insight on how the predicted permeability values differ from the experimental values. Here, the residuals for both the 2D and 3D descriptors are symmetrically distributed around zero.

3.1.2 Part 2 – 60 MCPs

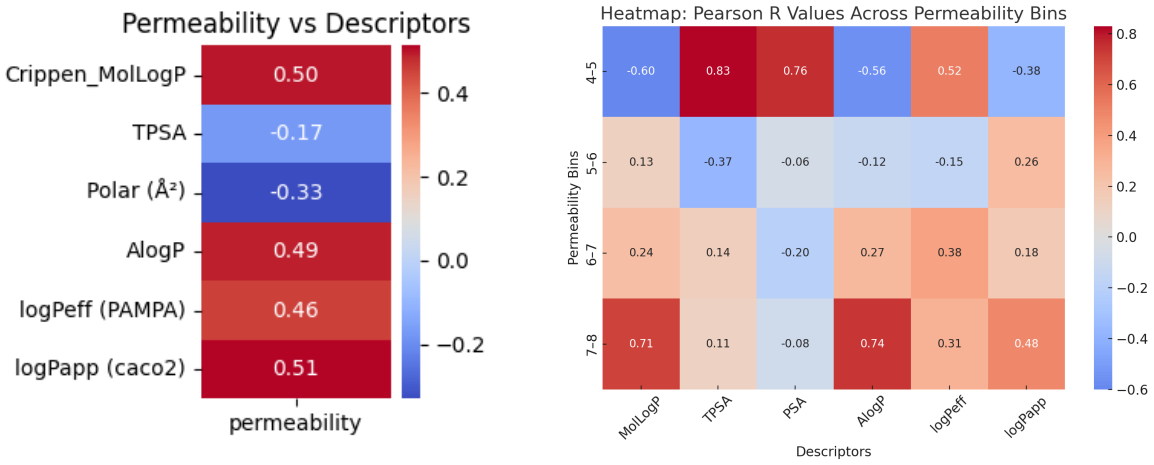
This part of the study focused on a manually curated subset of 60 MCPs, selected from the full CycPeptMPDB dataset to explore descriptor–permeability relationships in greater detail. The peptides were assembled as 30 matched pairs based on small structural variations (e.g., stereochemical inversions, N-methylations) and known differences in experimentally measured

permeability. This preparatory step required extensive manual inspection, monomer-level comparison, and structural reconstruction, making it one of the most time-intensive components of the project. To ensure data quality, only peptides with continuous, high-confidence permeability values from cell-based assays were retained.

The curated peptides were processed using the established computational workflow. Modified monomers not present in the pyPept library were manually built using ChemDoodle and integrated using HELM notation. For each MCP, descriptors related to lipophilicity and polarity were calculated, including MolLogP, AlogP, TPSA, and solvent-accessible PSA. The full table with all calculated descriptors are available in Table 6.1 in the Appendix.

Pearson correlation analysis was used to examine how well each descriptor matched the experimentally measured permeability values. These results are shown in Figure 3.3(a). MolLogP and AlogP showed moderate positive correlations (R = 0.50/0.49), while polarity-related descriptors such as TPSA and PSA were inversely correlated, with PSA yielding R = –0.33 and TPSA only –0.17. Additionally, permeability predictions from external small-molecule-trained models, $\log P_{\text{eff}}$ (PAMPA) and $\log P_{\text{app}}$ (Caco-2), were included and yielded R values of 0.46 and 0.51, respectively. Notably, MolLogP was among the most correlated descriptors in both the full dataset and the 60-MCP subset, suggesting consistency across scales.

To better understand how descriptor performance varies across permeability levels, the 60-MCP dataset was divided into bins (–8 to –7, –7 to –6, etc.), and are visualized in Figure 3.3(b). R^2 values were calculated within each interval, revealing that descriptor–permeability correlation peaked in the –5 to –4 range and declined at both lower and higher extremes.



(a) Correlation plot of descriptors and permeability.

(b) Heat map of R-values for intervals.

Figure 3.3: (a) Pearson correlation coefficients (R) between selected molecular descriptors and experimentally measured permeability values for 60 MCPs. Positive values indicate direct correlation, while negative values indicate inverse correlation. (b) Heatmap showing R^2 values for each descriptor within different permeability intervals. This binning highlights how descriptor–permeability relationships vary across the dynamic range of permeability. Higher R^2 values indicate stronger linear correlation within each interval.

3.1.3 Part 3 – Selected Matched Pairs for Synthesis

To build on the correlation analysis, a smaller group of matched peptides was selected to explore whether calculated free energies differences and lipophilicity values align with experimental permeability data, and to evaluate how well computational methods can predict these properties in more controlled cases.

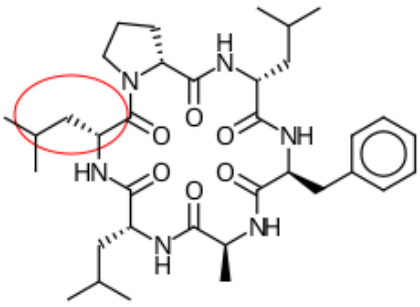
From the initial set of 30 matched MCP pairs, five were selected for detailed computational analysis of their free energy differences. For each peptide, the free energy difference for transfer from water to 1-octanol ($\Delta G_{\mathrm{w}\to\mathrm{o}}$) was computed using the *ABSOLV* framework, along with predicted logP values based in the free energy differences. These values were compared to experimental permeability data (logPeff), and their correlations are seen in Table 3.1.

Descriptor	**Correlation (R)**
logP	-0.79
$\Delta G_{\mathrm{w}\to\mathrm{o}}$	0.77

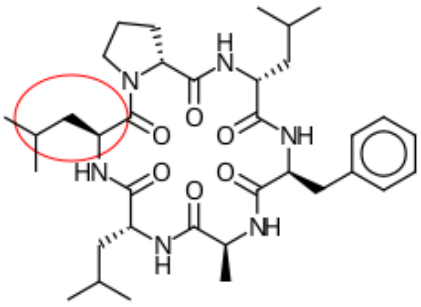
Table 3.1: Correlation between predicted free energy transfer and logP values and the experimental permeabilities from CycPeptMPDB

Two of the matched MCP pairs were selected for synthesis based on material availability and project timelines, and are visualized in Figure 3.4. The resulting four peptides were further

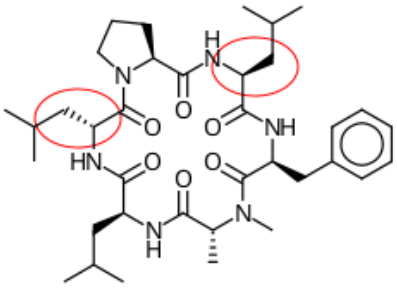
analyzed using both molecular descriptors and MD simulations. Herein, representative results for one of the synthesized peptides are presented below.



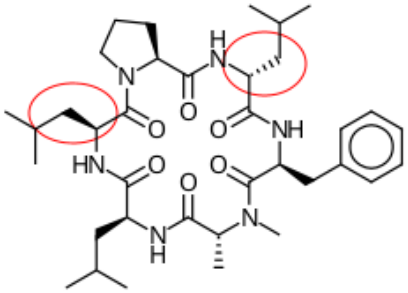
(a) Peptide 37



(b) Peptide 39



(c) Peptide 48



(d) Peptide 49

Figure 3.4: Chemical structures of the four synthesized macrocyclic peptides selected from matched pairs: peptide 37, 39, 48, and 49. Peptides 37/39 and 48/49 differ by stereochemistry, allowing structure–permeability relationships to be assessed.

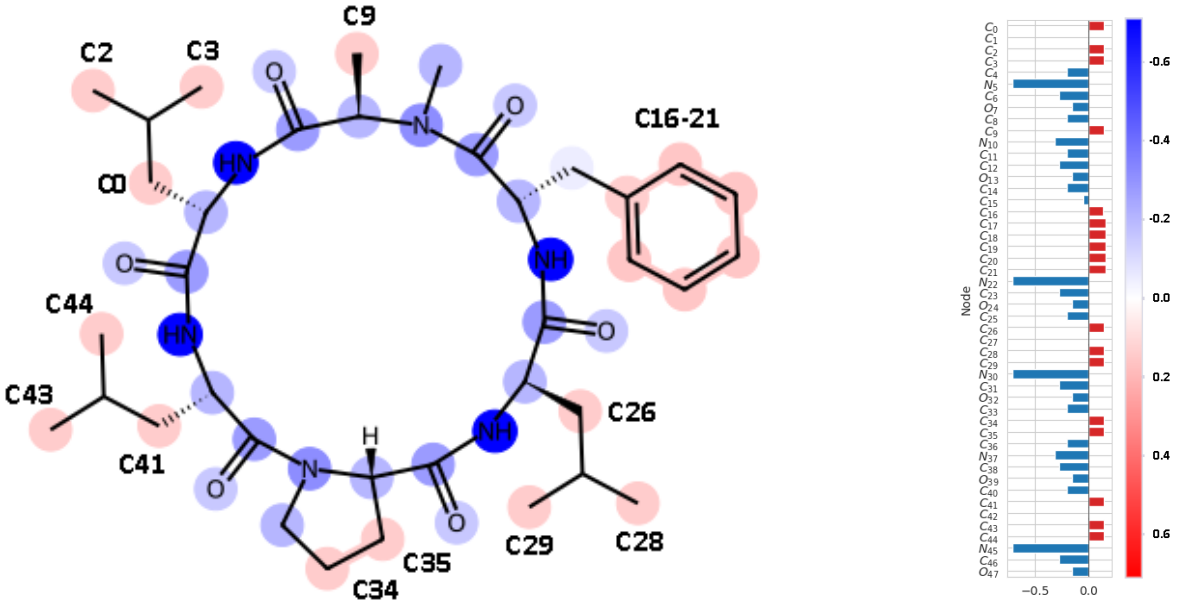
Peptide	$\boldsymbol{P}_{exp}$	**MolLogP**	**TPSA**	**fPSA**	**ALogP**	**logP**$_{eff}$	**logP**$_{app}$	**cPSA**	**%T**
37	-6.67	1.816	165.81	139.23	3.47	-3.92	0.17	75.7	2.65
39	-4.82	1.816	165.81	150.37	3.47	-3.92	0.14	78.2	85.3
48	-5.08	2.158	157.02	141.05	3.68	-3.97	0.43	74.6	64.9
49	-6.36	2.158	157.02	119.49	3.68	-3.97	0.26	77.8	5.33

Table 3.2: Descriptor and permeability data for the two selected matched peptide pairs.

Conformational flexibility was assessed using *Melodia_py* by calculating the standard deviation of curvature, torsion, arc length and writhing number across the peptide backbone from NOE-distances MD simulations. This approach provides an initial measure of conformational variability.

To further explore descriptor relevance, the contributions of selected 2D descriptors such as MolLogP, VSA_EState9, and EState_VSA6 were visualized using the explainable_qsar and vsa-explainer frameworks to better understand the atomic contribution of individual atoms to logP and VSA/EState descriptors. The contribution of MollogP and the atomic-level-interpretation is visualized for peptide 49 in Figure 3.5. The atomic contributions for the

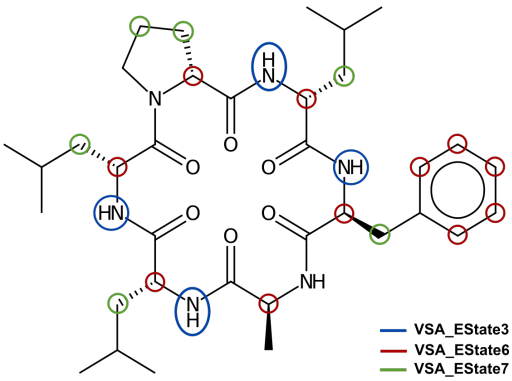
VSA_EState descriptor for peptide 39 and 48 are visualized in Figure 3.6



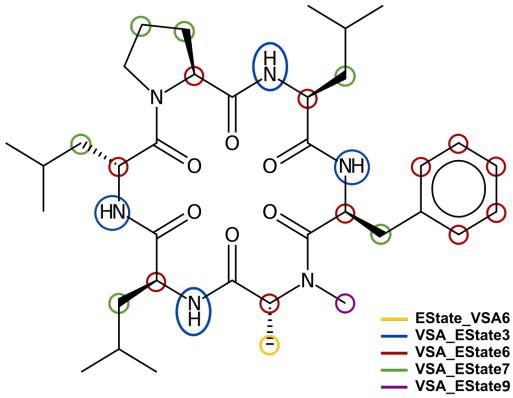
(a) Visualization of MolLogP contributions for molecule 49

(b) Atomic-level interpretation of logP contribution

Figure 3.5: (a) Atom-level contributions to MolLogP for peptide 49, visualized using explainable_qsar. (b) Detailed breakdown of atomic influences, where red regions increase logP and blue regions decrease it. Blue indicate negative contributions. Obtained with explainable_qsar.



(a) Atomic contributions for VSA_EState descriptors in Peptide39



(b) Atomic contributions for VSA_EState descriptors in Peptide48

Figure 3.6: Atomic contributions to VSA_EState descriptors for peptides 39 (a) and 48 (b), highlighting backbone regions and N-methylations as key structural determinants. Obtained with vsa-explorer (*github.com/srijitseal/vsa_explainer*)

3.2 Synthesis and Validation

All synthesis, purification and NMR experiments and characterization steps were performed by the teams at the company. My role was limited to the computational comparison and analysis.

3.2.1 Synthesis of the compounds

Two matched MCP pairs (four molecules) were selected from the initial subset of 30 pairs for synthesis, based on raw material availability and project timeline constraints. All compounds were successfully synthesized and purified, by the peptide group at the company. Details of the synthesis and purification protocols, along with compound information, are summarized in Table 3.3.

Table 3.3: Details on synthesis and purification of the four selected MCPs that were synthesized.

Peptide ID	**Theoretical MW**	***m/z* [M+H]$^+$**	**% yield**	**% purity**
37	654	655	≈ 25-30	99.71
39	654	655	≈ 25-30	99.64
48	668	669	≈ 25-30	99.41
49	668	669	≈ 25-30	99.82

3.2.2 logD determination

To evaluate the accuracy of the computational predictions, four selected MCPs were synthesized and experimentally tested for logD by colleagues at the company. In addition to

permeability estimates, the assay provided experimental logD values for each peptide, which were compared to the predicted values obtained through *ABSOLV*-based calculations or ML models trained from small molecules experimental data. As mentioned in the introduction, in this case LogD is equivalent to LogP due to the absence of ionizable groups in the MCPs. The results are summarized in Table 3.4

Table 3.4: Experimental logD values obtained for the four synthesized MCPs.

Peptide ID	***ABSOLV* logP**	**ML logD**	**Experimental logD**
37	5.10	2.27	3.06
39	1.46	2.44	3.35
48	2.78	2.81	4.16
49	2.21	2.76	4.00

3.2.3 Conformer Generation using NMR NOE-derived Distances

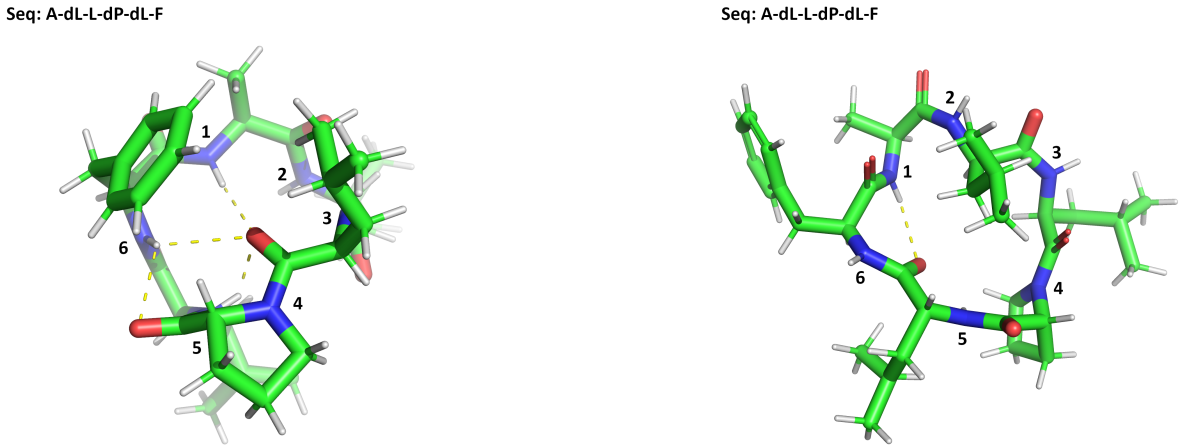
Each synthesized compound was prepared in $CDCl_3$ and DMSO-d_6 for NMR analysis. Spectra were recorded using NOESY, TOCSY, and HSQC experiments to evaluate the solution-state conformations by colleagues at the company. From each NOESY spectra, 15 and 11 proton-proton distances were calculated for $CDCl_3$ and DMSO, respectively, and used to generate NOE-restrained conformers, which were then compared to computational predictions.
Table 3.5 summarizes the root-mean-square deviation (RMSD) values between all conformers generated for peptide 39 across different tools and solvent conditions.

Table 3.5: Root-Mean Square Deviation values between all generated conformers of Pep39.

	DMSO	$CDCl_3$	pyPept	DataWarrior	$ABSOLV_{wat}$	$ABSOLV_{oct}$
DMSO		4.46	4.82	2.99	3.76	3.12
$CDCl_3$	5.27		2.79	4.00	2.94	3.38
pyPept	6.18	6.27		4.22	3.44	4.12
DataWarrior	7.33	6.34	6.28		3.05	2.47
$ABSOLV_{wat}$	7.04	5.96	6.83	5.92		2.39
$ABSOLV_{oct}$	5.87	6.61	6.87	9.48	7.84	

Top rows, minimized RMSD values; bottom rows, RMSD values from centering molecules at origin

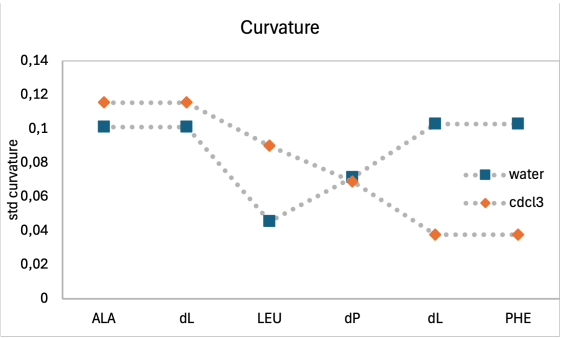
Figure 3.7 shows the conformers generated with *noeETKDG* that had the lowest number of NOE distance violations, defined as deviations from expected interproton distances calculated using a reference distance and the inverse-sixth power relationship from NOESY cross-peak intensities. 4% for the DMSO structure and 37% for the chloroform structure. The radius of gyration decreased by approximately 8% in DMSO and by 2% in chloroform. According to the MD simulations, the average number of IMHBs was 0.5 ± 0.6 per timeframe in DMSO.



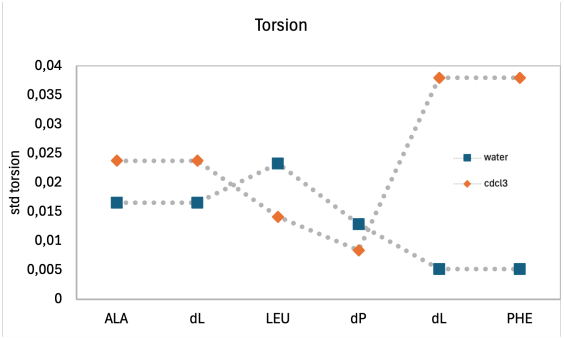
(a) DMSO (b) CHCl_3

Figure 3.7: NOE-restrained 3D conformers of peptide 39 generated in (a) DMSO and (b) chloroform. Images represent the conformations with the lowest number of NOE distance violations.

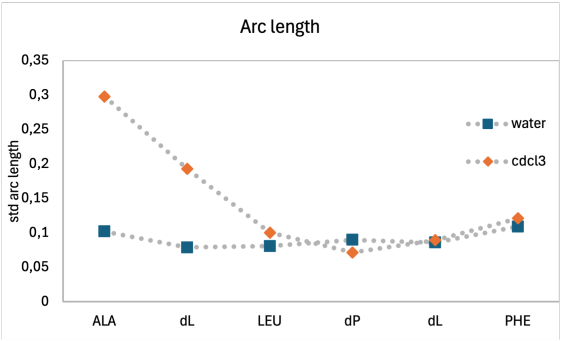
To further quantify structural differences between the solvent conditions, additional geometric descriptors, including curvature, torsion, arc length, and writhing, were computed using *Melodia_py*. These metrics are shown in Figure 3.8. Across all descriptors, the values differed between DMSO and chloroform, with DMSO consistently showing lower radius of gyration and higher curvature.



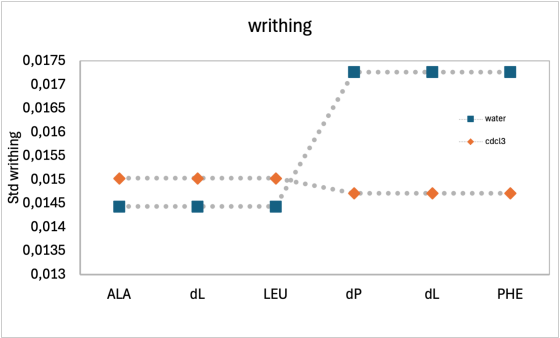
(a) curvature



(b) torsion



(c) arc length



(d) writhing

Figure 3.8: Per-residue structural metrics computed for peptide 39 in DMSO and chloroform: (a) curvature, (b) torsion, (c) arc length, and (d) writhing number.

Taken together, the results demonstrate that descriptor-based machine learning can identify key permeability-related features, but dynamic modeling and experimental validation are essential for capturing the full conformational complexity of MCPs. The agreement between predicted and observed differences across matched peptide pairs highlights the value of combining ML, MD, and NMR data in permeability prediction.

Chapter 4

Discussion

This thesis has explored how the permeability of macrocyclic peptides (MCPs) is influenced by their structural and physicochemical properties, with a focus on computational methods for predicting passive membrane permeability. This chapter analyzes the main findings, highlights key descriptor–permeability relationships, and evaluates their broader implications. The discussion is organized according to the three main workflows applied in the study: (1) full dataset modeling, (2) correlation analysis of a curated subset, and (3) experimental validation of selected matched pairs.

4.1 Part 1 - whole dataset

In the first part of this study, a large-scale machine learning approach was used to evaluate the extent to which molecular descriptors can predict passive membrane permeability in MCPs. A dataset of 6,986 MCPs from CycPeptMPDB was used to calculate both 2D and 3D descriptors, which were then used to train regression models using PyCaret. Multiple algorithms were benchmarked, and CatBoost consistently outperformed the others, producing the highest R^2 values and stable residual distributions; it was therefore selected for further analysis. The 2D-descriptor-based model achieved a test $R^2 \approx 0.639$ and training $R^2 \approx 0.790$, while the 3D-descriptor-based model showed a higher training $R^2 \approx 0.910$ but a lower test $R^2 \approx 0.543$, suggesting a degree of overfitting. This discrepancy may reflect the limitations of representing flexible macrocylic systems using only one static 3D conformer per peptide. In terms of predictive performance, the prediction error plot showed that predicted permeability values broadly followed the experimental trend along the diagonal, though some deviations were observed. The residual plot showed errors centered around zero with no major skew or heteroscedasticity, suggesting a relatively stable and unbiased model fit. These observations indicate that the model captures general permeability trends across the dataset, even if local accuracy varies.

In general, 2D descriptors offer broad chemical coverage and are less affected by conformational uncertainty, making them robust for large-scale modeling. Meanwhile, 3D descriptors encode spatial features and intramolecular interactions that can be critical for capturing local effects, such as steric or solvation-specific behaviors. However, their accuracy depends on reliable conformer generation, which still remains a challenge for macrocyclic systems. Together, these descriptor types provide complementary insights into the features that might govern permeability.

Feature importance analysis showed that MolLogP, VSA_EState9, and EState_VSA6 were the most influential 2D descriptors. These are related to the lipophilicity of the molecule and how its electronic properties are distributed—factors that are commonly linked to membrane permeability. For the 3D model, GRAV-1, DB1e, and TDB1m stood out as key descriptors. These capture how mass and atomic properties are distributed in space, and suggest that permeability in flexible macrocycles is influenced by how atoms are packed and arranged within the molecule. This supports the idea that both the amino acid type and its position within the macrocycle can strongly affect permeability, something that has also been observed experimentally [53].

One limitation of the models is that they could not properly tell the difference between stereoisomers, such as peptides containing L- versus D-amino acids. Although some studies suggest that a single change in stereochemistry does not always affect permeability [53], such changes can still impact the shape of the molecule. Because the descriptors were calculated from a single static 3D conformer per peptide, they could not capture conformational variability introduced by stereochemical changes. Even though the 3D model had high training accuracy, it may still miss subtle effects caused by conformational changes. A better approach for future work would be to use several conformers per peptide or include dynamic features from molecular simulations.

Another example of this limitation is *N*-methylation, which increases permeability by eliminating backbone hydrogen-bond donors, favoring intramolecular hydrogen bonding and compact conformations that shield polarity. This behavior was not clearly picked up by the models, likely because *N*-methylation affects the molecule's flexibility, which single static structures do not fully reflect. To better understand the role of such modifications, future studies could use MD simulations or examine how hydrogen bonds form and change over time.

The CatBoost model trained on the full dataset identified descriptors such as MolLogP, VSA_EState9, and EState_VSA6 as important for predicting permeability. While this model was not directly applied to the smaller 60-MCP subset, to avoid circular validation, as these peptides were part of the training data, several of the same descriptors, particularly MolLogP, also appeared as relevant in this smaller set, suggesting consistency across scales. Although the model effectively captured broad permeability trends, future work could focus on more narrowly defined subsets of peptides to improve interpretability and predictive accuracy. This could involve grouping molecules by structural class, sequence length, or permeability range to better resolve localized structure-property relationships.
These initial findings, while informative, highlighted both strengths and limitations in the descriptor-based models, particularly regarding stereochemistry and conformational dynamics, that motivated a more focused investigation on a curated subset of peptides, as described in Part 2.

4.2 Part 2 – 60 MCPs

The next step in the workflow aimed to investigate whether descriptor–permeability relationships identified in the full dataset also hold true at a more detailed level. To do this, a curated subset of 60 MCPs with experimental data was selected for detailed analysis. This part was done to be able to do a direct comparison between predicted molecular descriptors and measured permeability values using correlation plots and binning strategies. To ensure robustness of the correlation analysis, peptides with poor transport properties (e.g., <10 %recovery or <2.0 %transport) were excluded due to the risk of measurement error. These filters helped minimize false correlations from noisy experimental values. Descriptors such as MolLogP and PSA maintained reasonable correlations, consistent with observations from the full dataset. However, when the data were stratified into intervals based on permeability, the R^2 values varied across bins. This suggests that different structural features may dominate in distinct permeability regimes, and that descriptor utility can be context-dependent.

Among the five descriptors analyzed, the strongest correlations were observed in the –5 to –4 and –7 to –8 permeability intervals. Outside these ranges, the R^2 values declined noticeably, indicating that descriptor performance is not consistent across the full permeability spectrum. Specifically, descriptors like PSA and TPSA appeared more predictive for peptides with moderate or very low permeability, but less reliable in the intermediate ranges. This may reflect experimental variability or structural influences not captured by static, single-conformer descriptors. These findings highlight the value of stratified or context-aware analysis when evaluating structure-permeability relationships in flexible compounds like MCPs. They also underscore the importance of curating experimental permeability data, as measurement accuracy can vary depending on assay sensitivity and methodological consistency. The binning approach helped uncover these localized trends, which would likely be masked in a global correlation.
One reason for these differences may lie in how well each descriptor captures the structural complexity of MCPs. For instance, MolLogP showed consistent performance across the dataset, likely due to its sensitivity to overall lipophilicity, which strongly influences membrane permeability. In contrast, PSA and TPSA rely on static conformations and cannot account for IMHBs and shielding of polar atoms. This may explain their weaker and more variable correlations across permeability bind. Additionally, descriptors like ALogP often fail to distinguish stereoisomers, which is a critical limitation when analyzing matched pairs that have these subtle structural changes.

While the 2D-based model performed well on the full dataset (test $R^2 \approx 0.639$), the correlations between descriptors and permeability in the 60-MCP subset were noticeably weaker. This difference can be explained by several factors. Firstly, the larger dataset used binary labels (high or low permeability), which made it easier for the model to separate classes. The 60-peptide subset, on the other hand, used continuous permeability values and relied on correlation analysis, which is more sensitive to variability. Secondly, the full dataset covered a wide range of structures, while the smaller set was carefully selected to highlight subtle changes, such as stereochemistry or N-methylation, that are harder to detect using standard descriptors, especially when only one 3D conformer is used. Finally, the smaller sample size and possible differences in experimental conditions may also have reduced the strength of the correlations.

These differences show how important it is to match the model type, data format, and molecular diversity when comparing results across datasets.

Building on these differences, several limitations became evident in the analysis. The descriptors were calculated from single, static conformers and do not fully account for stereochemistry, flexibility, or solvent effects. For example, RDKit and related tools compute topological descriptors by fragmenting the molecule and summing atomic contributions, without accounting for 3D arrangement. As a result, descriptors such as XlogP, AlogP, or Crippen's logP yield identical values for matched pairs that differ only in stereochemistry (e.g., L- versus D-amino acids), even when such differences substantially affect permeability.

Additional challenges arose during conformer generation. pyPept, used for 3D structure generation, constructs conformers from distance matrices, but in some cases produced strained geometries including strained peptide bonds in the *cis* configuration., which are energetically unfavorable. To resolve this, the pyPept source code was slightly modified to enforce torsion angle constraints consistent with established peptide backbone geometries. This modification in the code resulted in predicted structures that more closely resembled expected geometries and showed reduced artificial strain. The modified code can be found in the Appendix. These observations highlights both the complexity of conformer generation for MCPs and the need for tailoring computational tools to accommodate that complexity.

To overcome the limitations associated with static, topological descriptors, trajectory-based methods were also explored. Tools such as ABSOLV and Melodia_py enabled the extraction of time-averaged structural features, providing insight into how peptides behave dynamically in both polar and apolar environments. These tools offer a more realistic representation of the relationship between structure and permeability, especially for MCPs which are known for their conformational flexibility and environment-dependent behavior.

Due to the scope of the project, trajectory-based analysis was performed only on the subset of matched pairs selected for synthesis, rather than the full 60-peptide set.

4.3 Part 3 – Selected Matched Pairs

From the original set of 30 matched pairs, five were selected for free energy difference analysis using the ABSOLV framework. Free energies differences from water to 1-octanol ($\Delta G_{\mathrm{w}\to\mathrm{o}}$) and predicted logP values were calculated and compared to experimental permeability data, as summarized in Table 3.1. The results revealed a strong positive correlation between $\Delta G_{\mathrm{w}\to\mathrm{o}}$ and permeability ($R = 0.77$), and a strong negative correlation between predicted logP and permeability ($R = –0.79$). Notably, higher (less negative) free energies differences were associated with increased permeability, suggesting more favorable partitioning into hydrophobic environments. Conversely, lower logP values may reflect decreased membrane integration or retention. Together, these findings support solvation energetics as a key determinant of passive permeability in MCPs.

Two of these matched pairs were selected for synthesis, based on both scientific interest and practical constraints such as raw material availability. The selected peptides exhibited no-

table differences in predicted permeability despite similar overall structures. While statistical validation was not the aim, the intention was to explore whether observable structural or conformational differences could explain the predicted permeability gaps through a combination of computational and experimental tools.

Explainable QSAR techniques where applied to visualize atomic-level contributions to predicted MolLogP values for the four selected MCPs. In many cases, the side chains positioned on the MCPs outer surface had the most pronounced influence, either enhancing or reducing lipophilicity depending on their properties. Although not sufficient to fully explain all permeability differences, these visualizations offer insight into which structural features are most influential in modulating predicted logP.

Further analysis revealed that the peptide backbone generally contributes negatively to MolLogP, particularly in regions with methylated amines. These consistently appeared as strong blue regions, indicating strong negative contributions. In contrast, hydrophobic side chains frequently appeared in red, reflecting a positive contribution on overall lipophilicity. These trends suggest that backbone polarity may reduce passive permeability, while peripheral side chains can offset this effect by increasing lipophilic surface exposure.

Similarly, visualization of VSA/EState descriptors also identified key structural regions contributing to their importance in predicting permeabilities. As with MolLogP, the backbone was found to contribute most strongly, typically with a negative influence. Notably, N-methylation of the amide groups was also identified to play an important role, aligning well with previous studies linking this modification to enhanced permeability.

Altogether, these complementary methods, thermodynamic modeling, structural visualization, and explainable machine learning, offer a more comprehensive understanding of membrane permeability than descriptor values alone. Although only four peptides were experimentally validated, the combined use of computational predictions and NMR data supports the notion that passive membrane permeability is governed by a combination of environment-dependent conformational flexibility and physicochemical features.

4.3.1 Experimental results from logD

Following the computational predictions and descriptor analysis in Part 1-3, this section presents experimental validation of permeability-relevant properties for four synthesized MCPs. The aim was to assess whether predicted physicochemical values and conformations align with solution-state experimental data. This includes (i) logD measurements compared to machine learning predictions and (ii) structural characterization using NMR and geometric descriptors. Although the scope was limited to four MCPs, the results offer important insight into how well current computational tools perform in realistic chemical environments.

Firstly, experimental logD values were measured for four peptides and are presented in Table 3.4 in the Results part. Although the dataset is too small for statistical validation, clear trends can still be observed. The experimental logD values showed a strong correlation with machine learning predictions (R^2=0.99), with both methods correctly identifying which MCP of each

pair that had higher lipophilicity.

4.3.2 NMR

After synthesis and purification of two matched peptide pairs, conformational analysis was performed using NMR spectroscopy. The full spectral assignment, including NOE peak extraction and structure generation, was carried out by the NMR team at the company. My contribution focused on computationally comparing the NOE-derived structures to predicted conformers generated using different modeling tools. This allowed for an assessment of how well computational methods capture experimentally observed conformational preferences in different solvent environments.

To assess how well computational methods capture experimentally observed conformations, NOE-restrained NMR structures assigned by the NMR-team, were compared to predicted conformers of peptide 39. Root-mean-square deviation (RMSD) analysis revealed that conformers generated by ABSOLV in water and octanol were similar to each other (2.39 Å and 2.47 Å), respectively. This indicates that the MD-based methods produced consistent structures across different solvent conditions. In contrast, conformers generated by pyPept and DataWarrior deviated more substantially from the NOE-based structures, with RMSD values typically exceeding >4 Å. This indicates limited accuracy in reproducing experimental conformations. Notably, pyPept conformers aligned more closely with NOE_{CDCl3} than with NOE_{DMSO}, suggesting that the method performs better in apolar environments. These findings suggest that static or geometry based conformer generators may struggle to accurately model solvent-specific conformational changes. Given that conformation influences properties such as PSA and hydrogen bonding, these structural inaccuracies may propagate to errors in predicted permeability. This highlights the importance of integrating experimental data into the validation of computational tools and workflows.

Figure 3.7 illustrates the NOE-restrained 3D conformations of peptide 39 in DMSO and $CDCl_3$, highlighting solvent-specific structural differences. The DMSO structure is visibly more compact and twisted, exhibiting greater torsion compared to the more extended conformation observed in chloroform. Using the NOE-derived distances, residue level geometric descriptors, including curvature, torsion, arc length and writhing, were computed and visualized in Figure 3.6. The analysis suggests that Leu3 has the greatest influence on conformational variability, as evidenced by a marked increase in standard deviation in curvature at this position.

Overall, RMSD comparisons between NOE-restrained structures and computationally generated conformers revealed consistent patterns. Most predicted structures deviated more than 2.0 Å, suggesting that further refinement is needed for them to reflect the true conformation in solution. pyPept conformers were more similar to NOE_{CDCl3} structures, indicating potential optimization for apolar environments. In contrast, DataWarrior predictions aligned more closely to NOE_{DMSO}, suggesting better performance in polar solvents. The lowest RMSD values between DataWarrior and $ABSOLV_{oct}$ (2.47 Å) and between $ABSOLV_{wat}$ and $ABSOLV_{oct}$ (2.39 Å) indicate high internal consistency between solvent-sensitive MD-based conformers. These differences highlight how the choice of conformer generation method can significantly

affect subsequent property predictions, and emphasize the importance of accounting for solvent effects in computational tools and workflows.

Differences in radius of gyration and curvature supports the idea that MCPs adopt more compact and twisted conformations in polar solvents such as DMSO. This conformational compaction likely facilitates more intramolecular hydrogen bonding, as confirmed by the increased count in them. In contrast, the chloroform conformer was more extended, with a wider ring and the D-Leucine residue positioned centrally, while other side-chains extended outwards, consistent with reduced environmental polarity.

To extend the RMSD analysis, additional residue-level geometric descriptors such as curvature, torsion, arc length, and writhing were calculated using Melodia_py. This provided detailed insight into how individual residues contribute to conformational changes changes across solvents. The data showed that residues influence the conformation of the MCP to varying degree. Leu3 in particular, had a strong effect on curvature in DMSO, marking it as a key contributor to solvent-dependent folding. Torsion values were generally higher in chloroform, suggesting that the lack of stabilizing IMHB in in apolar solvent induces greater conformational flexibility or strain. Smaller residues like Ala1 contributed more to conformational flexibility in chloroform, leading to wider geometries. In contrast, in DMSO, the MCP adopted a more compact and twisted structure. This twisting was especially apparent in the curvature and writhing data, where Leu1 and D-Pro4 emerged as primary drivers of backbone distortion in polar environments.

However, even with these promising results, there are several limitations to this. Most computational models here rely on static descriptors that do not fully capture the conformational diversity of MCPs. Although this was partially addressed with ABSOLV and MD simulations fora selected subset of peptides, the majority of descriptors, including 3D ones, were still based on a single conformer per molecule. Additionally, standard logP descriptors such as MollogP and AlogP do not account for stereochemistry, leading to indistinguishable values for matched pairs with opposite chirality. Future work could benefit from integrating ensemble-based features or data driven 3D representations that reflect solvent-specific behavior more broadly across the dataset.

These findings reinforce the importance of solvent-aware conformational sampling in computational tools and workflows aimed at predicting permeability in MCPs.

Chapter 5

Conclusion

This thesis set out to evaluate whether computational methods can accurately predict the permeability and conformational landscape of macrocyclic peptides (MCPs), a class of compounds with growing pharmaceutical interest. By integrating machine learning, molecular descriptors, and molecular dynamics simulations, the study examined both large-scale trends across nearly 7,000 MCPs and detailed structural differences in selected matched pairs. The modeling results showed that both 2D and 3D descriptors capture relevant features associated with membrane permeability. While 2D descriptors are less affected by conformational variability and well-suited for high-throughput screening, 3D descriptors offer spatial and electronic detail that can be essential for capturing local interactions and conformational effects. Each approach brings complementary strengths, and their combined use provides a more comprehensive framework for understanding structure–permeability relationships.

In relation to the first research question, the results showed that computational tools, particularly machine learning models using 2D and 3D descriptors, can predict permeability with moderate to strong accuracy. However, accuracy is constrained by the use of static conformers and lack of stereochemical sensitivity. The second question, concerning the relevance of traditional permeability descriptors, revealed that while parameters like MolLogP and TPSA correlated reasonably with permeability in many cases, they failed to distinguish matched pairs differing only by chirality or local folding. To address the third question, modifications to PyPept and the integration of solvent-aware MD simulations using NOE-derived distance restraints and ABSOLV proved essential for improving structural realism and predictive performance in selected cases.

A focused analysis on 30 matched pairs revealed that minor structural changes can significantly influence permeability, though traditional descriptor-based models often struggle to capture stereochemical nuances. Experimental synthesis and characterization of two selected pairs further validated computational predictions, particularly in terms of conformational preferences and solvation energetics. The results support the well-established hypothesis that the ability to form internal hydrogen bonds is critical for passive membrane permeability and can be the distinguishing factor among closely related compounds. These findings highlight the potential of refined computational pipelines to aid in the rational design of permeable MCPs. Nevertheless, improvements are needed in handling stereochemistry and conformational ensembles. Future work should continue to integrate enhanced dynamic descriptors, expand experimen-

tally validated datasets, and refine conformer generation protocols to better reflect the flexible nature of macrocycles.

Overall, this work demonstrates that integrating machine learning with both static and dynamic molecular modeling can significantly advance the predictive design of bioavailable MCPs. For future work, there is a clear need to establish a more streamlined and integrated workflow that combines structure generation, conformational validation, and permeability prediction. In this study, tools such as `pyPept`, NOE-based distance extraction, and `ABSOLV` were used in sequence, requiring manual coordination and format conversion between steps. A unified platform would greatly improve efficiency, reproducibility, and accessibility for similar studies on macrocyclic peptides. This study provides a foundation for developing integrated, experimentally grounded modeling frameworks to accelerate the discovery of orally available MCPs.

Bibliography

[1] Alexander Alex et al. "Intramolecular hydrogen bonding to improve membrane permeability and absorption in beyond rule of five chemical space". In: *Med. Chem. Commun.* 2 (7 2011), pp. 669–674. DOI: 10.1039/C1MD00093D. URL: http://dx.doi.org/10.1039/C1MD00093D.

[2] M. Bermejo et al. "PAMPA–a drug absorption in vitro model 7. Comparing rat in situ, Caco-2, and PAMPA permeability of fluoroquinolones". In: *European Journal of Pharmaceutical Sciences* 21.4 (2004), pp. 429–441. DOI: 10.1016/j.ejps.2003.10.009.

[3] Patrick Bleiziffer, Kay Schaller, and Sereina Riniker. "Machine Learning of Partial Charges Derived from High-Quality Quantum-Mechanical Calculations". In: *Journal of Chemical Information and Modeling* 58.3 (2018). PMID: 29461814, pp. 579–590. DOI: 10.1021/acs.jcim.7b00663. eprint: https://doi.org/10.1021/acs.jcim.7b00663. URL: https://doi.org/10.1021/acs.jcim.7b00663.

[4] Simon Boothroyd. *absolv: A Python package for solvation free energy calculations*. Accessed: 2025-03-27. 2025. URL: https://simonboothroyd.github.io/absolv/latest/.

[5] Richard B van Breemen and Yongmei Li and. "Caco-2 cell permeability assays to measure drug absorption". In: *Expert Opinion on Drug Metabolism & Toxicology* 1.2 (2005). PMID: 16922635, pp. 175–185. DOI: 10.1517/17425255.1.2.175. eprint: https://www.tandfonline.com/doi/pdf/10.1517/17425255.1.2.175. URL: https://www.tandfonline.com/doi/abs/10.1517/17425255.1.2.175.

[6] Laura K. Buckton and Shelli R. McAlpine. "Improving the Cell Permeability of Polar Cyclic Peptides by Replacing Residues with Alkylated Amino Acids, Asparagines, and d-Amino Acids". In: *Organic Letters* 20.3 (2018). PMID: 29364690, pp. 506–509. DOI: 10.1021/acs.orglett.7b03363. eprint: https://doi.org/10.1021/acs.orglett.7b03363. URL: https://doi.org/10.1021/acs.orglett.7b03363.

[7] M.C. Burger. "ChemDoodle Web Components: HTML5 toolkit for chemical graphics, interfaces, and informatics". In: *Journal of Cheminformatics* 7 (2015), p. 35. DOI: 10.1186/s13321-015-0085-3.

[8] John Cavanagh et al. *Protein NMR Spectroscopy: Principles and Practice*. 2nd ed. San Diego: Academic Press, 2007. ISBN: 978-0-12-164491-8. DOI: 10.1016/B978-0-12-164491-8.X5000-3.

[9] P. L. Chang, A. W. Rinne, and T. G. Dewey. "Structure alignment based on coding of local geometric measures". In: *BMC Bioinformatics* 7 (2006), p. 346. DOI: 10.1186/1471-2105-7-346.

[10] Vito Digiesi et al. "Permeability prediction in the beyond-Rule-of 5 chemical space: Focus on cyclic hexapeptides". In: *European Journal of Pharmaceutics and Biopharmaceutics* 165 (2021), pp. 259–270. ISSN: 0939-6411. DOI: https://doi.org/10.1016/j.ejpb.2021.05.017. URL: https://www.sciencedirect.com/science/article/pii/S0939641121001478.

[11] Patrick G. Dougherty, Ashweta Sahni, and Dehua Pei. "Understanding Cell Penetration of Cyclic Peptides". In: *Chemical Reviews* 119.17 (2019). PMID: 31083977, pp. 10241–10287. DOI: 10.1021/acs.chemrev.9b00008. eprint: https://doi.org/10.1021/acs.chemrev.9b00008. URL: https://doi.org/10.1021/acs.chemrev.9b00008.

[12] Eric Durham et al. "Solvent accessible surface area approximations for rapid and accurate protein structure prediction". In: *Journal of molecular modeling* 15.9 (2009), pp. 1093–1108. DOI: 10.1007/s00894-009-0454-9.

[13] Peter Eastman et al. "OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials". In: *The Journal of Physical Chemistry B* 128.1 (2024). PMID: 38154096, pp. 109–116. DOI: 10.1021/acs.jpcb.3c06662. eprint: https://doi.org/10.1021/acs.jpcb.3c06662. URL: https://doi.org/10.1021/acs.jpcb.3c06662.

[14] Urban Fagerholm, Sven Hellberg, and Ola Spjuth. "Advances in Predictions of Oral Bioavailability of Candidate Drugs in Man with New Machine Learning Methodology". In: *Molecules* 26.9 (2021), p. 2572. DOI: 10.3390/molecules26092572.

[15] Justin H. Faris et al. "Membrane Permeability in a Large Macrocyclic Peptide Driven by a Saddle-Shaped Conformation". In: *Journal of the American Chemical Society* 146.7 (2024). PMID: 38330910, pp. 4582–4591. DOI: 10.1021/jacs.3c10949. eprint: https://doi.org/10.1021/jacs.3c10949. URL: https://doi.org/10.1021/jacs.3c10949.

[16] Thomas Fox et al. "BILN: A Human-Readable Line Notation for Complex Peptides". In: *Journal of Chemical Information and Modeling* 62.17 (2022). PMID: 35984937, pp. 3942–3947. DOI: 10.1021/acs.jcim.2c00703. eprint: https://doi.org/10.1021/acs.jcim.2c00703. URL: https://doi.org/10.1021/acs.jcim.2c00703.

[17] J. Fralliciardi, J. Melcr, P. Siginou, et al. "Membrane thickness, lipid phase and sterol type are determining factors in the permeability of membranes to small solutes". In: *Nature Communications* 13 (2022), p. 1605. DOI: 10.1038/s41467-022-29272-x. URL: https://doi.org/10.1038/s41467-022-29272-x.

[18] Arup K. Ghose, Vellarkad N. Viswanadhan, and John J. Wendoloski. "Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods". In: *The Journal of Physical Chemistry A* 102.21 (1998), pp. 3762–3772. DOI: 10.1021/jp980230o. eprint: https://doi.org/10.1021/jp980230o. URL: https://doi.org/10.1021/jp980230o.

[19] Colin A Grambow. "CREMP: Conformer-rotamer ensembles of macrocyclic peptides for machine learning". en. In: *Scientific Data* (2024). DOI: 10.1038/s41597-024-03698-y.

[20] Lowell H. Hall and Lemont B. Kier. "Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information". In: *Journal of Chemical Information and Computer Sciences* 35.6 (1995), pp. 1039–1045. DOI: 10.1021/ci00028a014. eprint: https://doi.org/10.1021/ci00028a014. URL: https://doi.org/10.1021/ci00028a014.

[21] Esther Heid et al. "Chemprop: A Machine Learning Package for Chemical Property Prediction". In: *Journal of Chemical Information and Modeling* 64.1 (2024). PMID: 38147829, pp. 9–17. DOI: 10.1021/acs.jcim.3c01250. eprint: https://doi.org/10.1021/acs.jcim.3c01250. URL: https://doi.org/10.1021/acs.jcim.3c01250.

[22] William Humphrey, Andrew Dalke, and Klaus Schulten. "VMD – Visual Molecular Dynamics". In: *Journal of Molecular Graphics* 14 (1996), pp. 33–38.

[23] Kin-Kai Hwang et al. "Permeation prediction of M100240 using the parallel artificial membrane permeability assay". en. In: *J Pharm Pharm Sci* 6.3 (Sept. 2003), pp. 315–320.

[24] Melanie Kah and Colin D. Brown. "LogD: Lipophilicity for ionisable compounds". In: *Chemosphere* 72.10 (2008), pp. 1401–1408. ISSN: 0045-6535. DOI: https://doi.org/10.1016/j.chemosphere.2008.04.074. URL: https://www.sciencedirect.com/science/article/pii/S004565350800605X.

[25] Alan R. Katritzky et al. "Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics". In: *The Journal of Physical Chemistry* 100.24 (1996), pp. 10400–10407. DOI: 10.1021/jp953224q. eprint: https://doi.org/10.1021/jp953224q. URL: https://doi.org/10.1021/jp953224q.

[26] bp-kelley. *Descriptastorus*. https://github.com/bp-kelley/descriptastorus. Accessed: 2025-04-23. 2025.

[27] Taeho Kim and Hwangseo Park. "Computational prediction of octanol–water partition coefficient based on the extended solvent-contact model". In: *Journal of Molecular Graphics and Modelling* 60 (2015), pp. 108–117. ISSN: 1093-3263. DOI: https://doi.org/10.1016/j.jmgm.2015.06.004. URL: https://www.sciencedirect.com/science/article/pii/S1093326315300103.

[28] Christian T. Klein, Dominik Kaiser, and Gerhard Ecker. "Topological Distance Based 3D Descriptors for Use in QSAR and Diversity Analysis". In: *Journal of Chemical Information and Computer Sciences* 44.1 (2004). PMID: 14741029, pp. 200–209. DOI: 10.1021/ci0256236. eprint: https://doi.org/10.1021/ci0256236. URL: https://doi.org/10.1021/ci0256236.

[29] Grant Koch et al. "Chromatographic Determination of Permeability-Relevant Lipophilicity Facilitates Rapid Analysis of Macrocyclic Peptide Scaffolds". In: *Journal of Medicinal Chemistry* 67.21 (2024). PMID: 39453819, pp. 19612–19622. DOI: 10.1021/acs.jmedchem.4c01956. eprint: https://doi.org/10.1021/acs.jmedchem.4c01956. URL: https://doi.org/10.1021/acs.jmedchem.4c01956.

[30] Paul Labute. "A widely applicable set of descriptors". In: *Journal of Molecular Graphics and Modelling* 18.4 (2000), pp. 464–477. ISSN: 1093-3263. DOI: https://doi.org/10.1016/S1093-3263(00)00068-1. URL: https://www.sciencedirect.com/science/article/pii/S1093326300000681.

[31] Albert Leo, Corwin Hansch, and David Elkins. "Partition coefficients and their uses". In: *Chemical Reviews* 71.6 (1971), pp. 525–616. DOI: 10.1021/cr60274a001. eprint: https://doi.org/10.1021/cr60274a001. URL: https://doi.org/10.1021/cr60274a001.

[32] Jianan Li et al. "CycPeptMPDB: A Comprehensive Database of Membrane Permeability of Cyclic Peptides". In: *Journal of Chemical Information and Modeling* 63.7 (2023). PMID: 36930969, pp. 2240–2250. DOI: 10.1021/acs.jcim.2c01573. eprint: https://doi.org/10.1021/acs.jcim.2c01573. URL: https://doi.org/10.1021/acs.jcim.2c01573.

[33] Christopher A. Lipinski et al. "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings". In: *Advanced Drug Delivery Reviews* 23.1 (1997). In Vitro Models for Selection of Development Candidates, pp. 3–25. ISSN: 0169-409X. DOI: https://doi.org/10.1016/S0169-409X(96)00423-1. URL: https://www.sciencedirect.com/science/article/pii/S0169409X96004231.

[34] Spiros Liras and Kim F. Mcclure. "Permeability of Cyclic Peptide Macrocycles and Cyclotides and Their Potential as Therapeutics". In: *ACS Medicinal Chemistry Letters* 10.7 (2019), pp. 1026–1032. DOI: 10.1021/acsmedchemlett.9b00149. eprint: https://doi.org/10.1021/acsmedchemlett.9b00149. URL: https://doi.org/10.1021/acsmedchemlett.9b00149.

[35] "Macrocyclic peptides: Aiming for the perfect fit". In: *C&EN Global Enterprise* 102.19 (2024), pp. 20–24. DOI: 10.1021/cen-10219-cover. eprint: https://doi.org/10.1021/cen-10219-cover. URL: https://doi.org/10.1021/cen-10219-cover.

[36] Favour Danladi Makurvet. "Biologics vs. small molecules: Drug costs and patient access". In: *Medicine in Drug Discovery* 9 (2021), p. 100075. ISSN: 2590-0986. DOI: https://doi.org/10.1016/j.medidd.2020.100075. URL: https://www.sciencedirect.com/science/article/pii/S2590098620300622.

[37] Bojan D. Markovic et al. "A PAMPA Assay as Fast Predictive Model of Passive Human Skin Permeability of New Synthesized Corticosteroid C-21 Esters". In: *Molecules* 17.1 (2012), pp. 480–491. ISSN: 1420-3049. DOI: 10.3390/molecules17010480. URL: https://www.mdpi.com/1420-3049/17/1/480.

[38] Robert T. McGibbon et al. "MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories". In: *Biophysical Journal* 109.8 (2015), pp. 1528–1532. DOI: 10.1016/j.bpj.2015.08.015.

[39] Rocco Meli and Philip C. Biggin. "spyrmsd: symmetry-corrected RMSD calculations in Python". In: *J. Cheminform.* 12 (2020). DOI: 10.1186/s13321-020-00455-2.

[40] Rinaldo W Montalvão et al. "Melodia: a Python library for protein structure analysis". In: *Bioinformatics* 40.7 (July 2024), btae468. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btae468. eprint: https://academic.oup.com/bioinformatics/article-pdf/40/7/btae468/58704525/btae468.pdf. URL: https://doi.org/10.1093/bioinformatics/btae468.

[41] Hirotomo Moriwaki et al. "Mordred: a molecular descriptor calculator". In: *Journal of Cheminformatics* 10.1 (2018), p. 4. DOI: 10.1186/s13321-018-0258-y. URL: https://doi.org/10.1186/s13321-018-0258-y.

[42] Matthew R Naylor et al. "Cyclic peptide natural products chart the frontier of oral bioavailability in the pursuit of undruggable targets". In: *Current Opinion in Chemical Biology* 38 (2017). Next Generation Therapeutics, pp. 141–147. ISSN: 1367-5931. DOI: https://doi.org/10.1016/j.cbpa.2017.04.012. URL: https://www.sciencedirect.com/science/article/pii/S1367593117300480.

[43] Rodrigo Ochoa, J. B. Brown, and Thomas Fox. "pyPept: a python library to generate atomistic 2D and 3D representations of peptides". In: *Journal of Cheminformatics* 15.1 (Sept. 2023), p. 79. ISSN: 1758-2946. DOI: 10.1186/s13321-023-00748-2. URL: https://doi.org/10.1186/s13321-023-00748-2 (visited on 01/14/2025).

[44] Vasanthanathan Poongavanam et al. "Molecular chameleons in drug discovery". en. In: *Nature Reviews Chemistry* 8.1 (Dec. 2023), pp. 45–60. ISSN: 2397-3358. DOI: 10.1038/s41570-023-00563-1. URL: https://www.nature.com/articles/s41570-023-00563-1 (visited on 01/13/2025).

[45] S. Prasanna and R. J. Doerksen. "Topological polar surface area: a useful descriptor in 2D-QSAR". In: *Current Medicinal Chemistry* 16.1 (2009), pp. 21–41. DOI: 10.2174/092986709787002817.

[46] RDKit. *RDKit: Open-source cheminformatics*. 2024. DOI: 10.5281/zenodo.591637. URL: https://www.rdkit.org.

[47] Sereina Riniker. "Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data To Predict Free-Energy Differences". In: *Journal of Chemical Information and Modeling* 57.4 (2017). PMID: 28368113, pp. 726–741. DOI: 10.1021/acs.jcim.6b00778. eprint: https://doi.org/10.1021/acs.jcim.6b00778. URL: https://doi.org/10.1021/acs.jcim.6b00778.

[48] Daniel R. Roe and Thomas E. III Cheatham. "PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data". In: *Journal of Chemical Theory and Computation* 9.7 (2013). PMID: 26583988, pp. 3084–3095. DOI: 10.1021/ct400341p. eprint: https://doi.org/10.1021/ct400341p. URL: https://doi.org/10.1021/ct400341p.

[49] Daniel R. Roe and III Thomas E. Cheatham. "PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data". In: *Journal of Chemical Theory and Computation* 9.7 (2013), pp. 3084–3095.

[50] Gabriela B. Santos, A. Ganesan, and Flavio S. Emery. "Oral Administration of Peptide-Based Drugs: Beyond Lipinski's Rule". In: *ChemMedChem* 11.20 (2016), pp. 2245–2251. DOI: https://doi.org/10.1002/cmdc.201600288. eprint: https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/cmdc.201600288. URL: https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cmdc.201600288.

[51] LLC Schrödinger and Warren DeLano. *PyMOL*. Version 2.5.0. May 20, 2020. URL: http://www.pymol.org/pymol.

[52] Dan Sindhikara et al. "Automated Design of Macrocycles for Therapeutic Applications: From Small Molecules to Peptides and Proteins". en. In: *Journal of Medicinal Chemistry* 63.20 (Oct. 2020), pp. 12100–12115. ISSN: 0022-2623, 1520-4804. DOI: 10.1021/acs.jmedchem.0c01500. URL: https://pubs.acs.org/doi/10.1021/acs.jmedchem.0c01500 (visited on 01/16/2025).

[53] C. Townsend et al. *The Passive Permeability Landscape Around Geometrically Diverse Hexa- and Heptapeptide Macrocycles*. ChemRxiv. This content is a preprint and has not been peer-reviewed. 2020. DOI: 10.26434/chemrxiv.13335941.v1.

[54] Alexander A. Vinogradov, Yizhen Yin, and Hiroaki Suga. "Macrocyclic Peptides as Drug Candidates: Recent Progress and Remaining Challenges". In: *Journal of the American Chemical Society* 141.10 (2019). PMID: 30768253, pp. 4167–4181. DOI: 10.1021/jacs.8b13178. eprint: https://doi.org/10.1021/jacs.8b13178. URL: https://doi.org/10.1021/jacs.8b13178.

[55] Renxiao Wang, Ying Fu, and Luhua Lai. "A New Atom-Additive Method for Calculating Partition Coefficients". In: *Journal of Chemical Information and Computer Sciences* 37.3 (1997), pp. 615–621. DOI: 10.1021/ci960169p. eprint: https://doi.org/10.1021/ci960169p. URL: https://doi.org/10.1021/ci960169p.

[56] Shuzhe Wang et al. "Incorporating NOE-Derived Distances in Conformer Generation of Cyclic Peptides with Distance Geometry". In: *Journal of Chemical Information and Modeling* 62.3 (2022). PMID: 35029985, pp. 472–485. DOI: 10.1021/acs.jcim.1c01165. eprint: https://doi.org/10.1021/acs.jcim.1c01165. URL: https://doi.org/10.1021/acs.jcim.1c01165.

[57] Mark C. Wenlock et al. "A Method for Measuring the Lipophilicity of Compounds in Mixtures of 10". In: *Journal of Biomolecular Screening* 16.3 (2011). PMID: 21343602, pp. 348–355. DOI: 10.1177/1087057110396372. eprint: https://doi.org/10.1177/1087057110396372. URL: https://doi.org/10.1177/1087057110396372.

[58] Chun Wei Yap. "PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints". In: *Journal of Computational Chemistry* 32.7 (2011), pp. 1466–1474. DOI: https://doi.org/10.1002/jcc.21707. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.21707. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21707.

[59] Tianhong Zhang et al. "HELM: A Hierarchical Notation Language for Complex Biomolecule Structure Representation". In: *Journal of Chemical Information and Modeling* 52.10 (2012). PMID: 22947017, pp. 2796–2806. DOI: 10.1021/ci3001925. eprint: https://doi.org/10.1021/ci3001925. URL: https://doi.org/10.1021/ci3001925.

[60] Alessandro Zorzi, Kaycie Deyle, and Christian Heinis. "Cyclic peptide therapeutics: past, present and future". In: *Current Opinion in Chemical Biology* 38 (2017). Next Generation Therapeutics, pp. 24–29. ISSN: 1367-5931. DOI: https://doi.org/10.1016/j.cbpa.2017.02.006. URL: https://www.sciencedirect.com/science/article/pii/S1367593116302010.

Chapter 6

Appendix

Table 6.1 lists the selected molecular descriptors used to characterize the 60 macrocyclic peptides investigated in Part 2 of this thesis.

Table 6.1: Descriptor values for 60 selected macrocyclic peptides.

Peptide ID	**Experimental Permeability**	**Crippen_MolLogP**	**TPSA**	**PSA**	**AlogP**	**logPeff**	**logPapp**
8	-7.3	2.548	186.04	169.82	4.438	-4.422	0.087
2	-6.2	2.548	186.04	176.81	4.438	-4.41	0.327
966	-6.0	3.174	179.9	162.66	4.047	-4.613	0.371
967	-5.1	3.174	179.9	117.2	4.047	-4.733	0.381
970	-5.7	2.832	188.69	156.18	3.841	-4.513	0.202
1003	-5.1	2.548	186.04	175.39	4.438	-4.255	0.366
1004	-7.35	2.548	186.04	176.66	4.438	-4.248	0.343
1005	-6.32	2.548	186.04	194.35	4.438	-4.257	0.354
1078	-8.0	0.303	186.04	210.21	1.921	-5.203	-0.762
1079	-7.15	0.303	186.04	216.53	1.921	-5.276	-0.555
1129	-6.89	-1.013	165.81	183.84	0.255	-5.055	-0.356
1130	-8.06	-0.625	165.81	166.69	0.632	-4.709	-0.378
1137	-5.04	2.301	139.44	142.91	3.29	-4.409	0.332
1138	-7.27	1.148	148.67	138.53	1.521	-4.52	0.014
1167	-6.58	4.677	139.44	107.7	5.755	-4.277	1.107
1168	-5.23	3.524	148.67	106.94	3.985	-4.371	1.17
1217	-4.96	2.937	139.44	108.95	4.133	-4.165	0.589
1218	-6.26	1.784	148.67	156.53	2.364	-4.233	0.321
1237	-6.82	4.16	139.44	97.44	5.689	-4.084	0.844
1238	-5.07	3.006	148.67	109.15	3.92	-4.128	0.726
1247	-8.71	5.313	139.44	105.44	6.598	-4.097	1.125
1248	-5.53	4.16	148.67	137.52	4.828	-4.137	1.213
1357	-8.61	5.302	139.44	84.84	6.654	-4.052	0.636
1358	-6.0	4.148	148.67	80.19	4.884	-4.119	0.865
1607	-5.32	2.301	139.44	135.94	3.29	-4.479	0.274
1608	-7.56	1.148	148.67	152.14	1.521	-4.593	-0.018
1679	-5.01	1.959	148.23	131.0	3.084	-4.529	0.015
1680	-6.93	0.805	157.46	155.32	1.315	-4.643	-0.263
2691	-4.93	2.55	157.02	89.77	4.174	-4.207	0.292
2692	-5.9	2.55	157.02	152.49	4.174	-4.183	0.271
3220	-6.33	2.158	157.02	114.82	3.679	-4.06	0.149
3251	-4.98	2.5	148.23	135.19	3.885	-4.139	0.784
3325	-5.08	2.158	157.02	141.05	3.679	-3.974	0.428
3324	-6.36	2.158	157.02	119.49	3.679	-3.975	0.259
3411	-5.06	2.158	157.02	158.05	3.679	-3.96	0.353

Peptide ID	Experimental Permeability	Crippen_MolLogP	TPSA	PSA	AlogP	logPeff	logPapp
3412	-6.11	2.158	157.02	110.35	3.679	-3.966	0.29
3420	-4.99	2.158	157.02	120.06	3.679	-3.96	0.353
3423	-6.18	2.5	148.23	105.96	3.885	-4.059	0.786
3496	-6.12	2.158	157.02	131.08	3.679	-3.892	0.372
3498	-7.2	2.158	157.02	131.27	3.679	-3.911	0.259
3858	-6.07	1.807	186.12	171.01	3.442	-3.93	0.02
3859	-4.79	2.149	177.33	125.04	3.648	-3.985	0.531
3893	-5.34	1.807	186.12	177.92	3.442	-3.925	0.004
3891	-7.85	1.807	186.12	152.75	3.442	-3.927	-0.024
4629	-6.17	2.149	177.33	105.71	3.648	-4.142	0.182
4630	-7.49	2.149	177.33	156.44	3.648	-4.146	0.097
4767	-7.21	2.647	168.54	166.41	3.828	-4.271	0.561
4768	-6.02	2.647	168.54	138.47	3.828	-4.271	0.561
4814	-6.75	2.149	177.33	114.48	3.648	-4.012	0.353
4815	-5.71	2.149	177.33	114.8	3.648	-3.996	0.348
5435	-5.29	1.807	186.12	163.45	3.442	-4.091	0.051
5436	-6.86	2.149	177.33	122.84	3.442	-4.16	0.52
2773	-6.08	2.55	157.02	121.64	4.174	-4.091	0.465
2770	-5.08	2.55	157.02	140.7	4.174	-4.091	0.313
3026	-6.67	1.816	165.81	139.23	3.473	-3.918	0.171
3029	-4.82	1.816	165.81	150.37	3.473	-3.919	0.143
4477	-6.69	1.807	186.12	166.46	3.442	-4.082	0.008
4479	-5.84	1.807	186.12	151.04	3.442	-4.078	0.046
5427	-7.08	1.807	186.12	152.06	3.442	-4.091	0.051
5429	-6.15	1.807	186.12	181.69	3.442	-4.094	0.055

Master's Theses in Mathematical Sciences 2025:E55
ISSN 1404-6342

LUNFTB-3001-2025

Applied Computational Science
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/