Student Project Proposal

ProbeMap-AI: Mixed-Solvent Molecular Dynamics and AI-Ready FragMaps for Protein Ligandability Mapping

Master area	AI for Drug Discovery; molecular simulation; structure-based drug design; cheminformatics
Project type	Computational project with Python implementation, molecular dynamics analysis and AI-ready feature generation
Duration	2-3 months
Supervision	One tutor or co-tutor; weekly or bi-weekly progress meetings
Recommended level	Basic Python and data analysis; basic knowledge of molecular modelling is useful
Main output	A documented notebook/workflow, ligandability maps, a technical report and a short final presentation
Possible datasets	A small protein structure or short MD trajectory; tutor-provided protein system; public PDB structures suitable for pocket analysis

Short description. The project asks the student to prototype a simplified mixed-solvent molecular dynamics workflow for identifying ligandable regions, cryptic pockets and chemically interpretable interaction hotspots on a protein surface. The student will use small molecular probes, such as xenon, benzene, isopropanol, acetonitrile and acetamide, to generate density maps and simplified energetic FragMaps. The final representation should be compact, interpretable and suitable for downstream AI analysis, such as clustering, ranking, feature engineering or graph-based comparison of binding sites.

1. Background and rationale

Ligandability assessment is a central step in early drug discovery. Before investing in virtual screening, de novo design or medicinal chemistry, it is useful to determine whether a protein region can sustain favourable interactions with drug-like chemical fragments. Static structures provide important information on pocket geometry, but they often underestimate dynamic behaviour, transient cavities and cryptic binding sites that appear only through protein breathing motions or solvent competition.

Mixed-solvent molecular dynamics provides a practical computational strategy for exploring these effects. In this approach, the protein is simulated in water containing small probe molecules with different chemical properties. Apolar probes can highlight hydrophobic cavities and internal channels, aromatic probes can reveal pi-compatible or hydrophobic hotspots, and polar probes can detect dipolar or hydrogen-bonding environments. Regions repeatedly visited by probes are converted into occupancy maps and then into simplified energy-like maps.

Within AI for drug discovery, the main interest is not only the molecular dynamics simulation itself, but the transformation of a large trajectory into reusable descriptors. Probe densities, hotspot clusters, pocket rankings and local chemical labels can become machine learning-ready features. They can support comparison between proteins, prioritization of binding pockets, interpretation of docking results and construction of graph or voxel representations for downstream models.

The project is designed as a realistic Master-level proof of concept. It does not require the student to reproduce a full proprietary SILCS protocol or to run very long enhanced-sampling simulations. Instead, the goal is to implement a transparent, open, SILCS-like educational workflow based on OpenMM and Python, with clear assumptions and controlled scope.

2. Project objective

The objective of the project is to implement and test a simplified ProbeMap-AI workflow that converts mixed-solvent molecular dynamics or probe-flooding simulations into interpretable ligandability maps and AI-ready pocket descriptors.

The project should remain realistic for a 2-3 month student activity. The minimal successful version may focus on xenon flooding and density-map analysis. A recommended version can include two or three organic probes. A more advanced version can integrate all selected probes and produce a ranked list of protein hotspots suitable for feature engineering or machine learning.

3. Specific aims

- Select a small protein system and prepare it for probe-based molecular dynamics analysis.
- Implement a simplified OpenMM workflow for restrained protein simulations in water with selected molecular probes.
- Run or analyse at least three independent replicas with different initial probe positions and random velocity seeds.
- Extract probe coordinates over time and convert them into three-dimensional occupancy and density maps.
- Transform normalized densities into simplified FragMap-like energetic maps using a transparent statistical formula.
- Cluster local maxima and generate a ranked table of chemically annotated ligandability hotspots.
- Convert the final output into AI-ready descriptors, such as pocket-level feature tables, voxel maps or probe-labelled graphs.
- Critically evaluate the strengths, assumptions and limitations of the approach in comparison with docking-only or static pocket analysis.

4. Proposed methodology

4.1 Input system and preparation

The student will start from a protein structure, preferably a small soluble protein or a protein domain with a known or suspected ligandable region. The structure will be cleaned by removing irrelevant crystallographic molecules, adding missing atoms and hydrogens, assigning protonation states and solvating the system with physiological ionic strength. The initial educational setup can use OpenMM with an AMBER-compatible protein force field and TIP3P water.

4.2 Probe selection

The probes will be chosen to represent complementary interaction types. Xenon can be used as a simple apolar monoatomic probe for hydrophobic cavities and cryptic internal regions. Benzene represents aromatic and hydrophobic interactions. Isopropanol captures amphiphilic environments. Acetonitrile reports on dipolar or acceptor-rich regions. Acetamide is useful for hydrogen-bond donor and acceptor environments. The student may begin with xenon only and progressively add organic probes if time and parametrization are feasible.

4.3 Molecular dynamics protocol

The core protocol will include energy minimization, restrained equilibration and production dynamics. Backbone atoms may be restrained with a harmonic positional restraint to maintain the global fold while allowing local side-chain and pocket rearrangements. Three replicas should be generated with different probe positions and velocity seeds. For a short Master project, the simulations can be short demonstration trajectories or precomputed trajectories provided by the tutor.

4.4 Density-map and FragMap construction

For each probe, coordinates will be extracted frame by frame and accumulated on a regular three-dimensional grid. Raw occupancy will be converted into a density map and normalized with respect to a bulk or reference density. A simplified free-energy-like map can be calculated using G(r) = -RT ln[rho(r)/rho0], where rho(r) is the local probe density and rho0 is a reference density. The result is not intended as an absolute binding free energy, but as a transparent hotspot score.

4.5 Hotspot clustering and pocket ranking

Low-energy or high-density regions will be thresholded and clustered. Each cluster will be described by size, mean score, minimum score, dominant probe type and approximate protein region. The final ranking should highlight hydrophobic, aromatic, amphiphilic, dipolar and hydrogen-bonding hotspots. When possible, clusters from different probes will be combined to define composite ligandability regions.

4.6 AI-ready representation

The final output will be converted into a machine-readable representation. A minimal representation is a CSV table containing one row per hotspot and columns for probe type, density, energy score, size, location and nearby residues. A richer representation may include a graph in which nodes are protein residues or hotspot clusters and edges represent spatial proximity or chemical complementarity. These descriptors can then be used for clustering, visualization, ranking or comparison between protein states.

5. Example probe-to-feature framework

Probe or map	Chemical role	Possible interpretation	AI-ready feature
Xenon	Apolar monoatomic probe	Hydrophobic cavities, cryptic pockets, internal channels	Hydrophobic pocket score; cavity occupancy; buried hotspot flag
Benzene	Aromatic hydrophobic probe	Aromatic-compatible and hydrophobic surface regions	Aromatic hotspot score; pi/hydrophobic feature label
Isopropanol	Amphiphilic probe	Mixed polar/apolar regions and solvent-exposed ligandable patches	Amphiphilic balance descriptor; local polarity transition feature
Acetonitrile	Small polar dipolar probe	Dipolar regions and acceptor-compatible environments	Polar density score; dipolar pocket annotation
Acetamide	Hydrogen-bond donor/acceptor probe	Hydrogen-bonding regions and polar anchoring points	H-bond hotspot label; donor/acceptor complementarity feature
Composite map	Combination of probe maps	Chemically diverse ligandability regions	Pocket-level feature vector for clustering or ML ranking

This framework is only a starting point. The tutor may narrow the scope to one protein, one probe class, or one simplified trajectory. The important requirement is that each map or descriptor remains linked to a chemical interpretation and to a reproducible computational step.

6. Indicative work plan

Period	Main task	Expected output
Weeks 1-2	Problem definition and system selection	Short technical note describing the selected protein, available files, probes, force field choices and expected outputs.
Weeks 3-4	Protein preparation and baseline workflow	Prepared structure, solvated system and initial OpenMM or analysis notebook; definition of restraints and simulation parameters.
Weeks 5-6	Probe setup and replica generation	First probe-containing systems or tutor-provided trajectories; at least three replicas or three initial probe configurations.

Period	Main task	Expected output
Weeks 7-8	Coordinate extraction and density maps	Python functions for extracting probe positions and generating occupancy or density maps for each probe.
Weeks 9-10	FragMap scoring and hotspot clustering	Energy-like maps, clustered hotspots and preliminary pocket ranking table.
Weeks 11-12	AI-ready representation and reporting	Final notebook, hotspot feature table, visualizations, short report and presentation summarizing methods, results and limitations.

7. Expected deliverables

- A documented Python notebook or small code repository implementing the simplified ProbeMap-AI workflow.
- A prepared example protein system or a clearly documented analysis of tutor-provided trajectories.
- Probe occupancy or density maps for at least one probe, preferably xenon; recommended versions include multiple probes.
- Simplified FragMap-like energetic maps and a clustered hotspot ranking table.
- A machine-readable pocket descriptor table suitable for clustering, visualization or downstream machine learning.
- At least one visualization of probe densities, hotspot clusters, pocket rankings or AI-ready feature space.
- A final technical report of approximately 3-5 pages, including aims, methods, results, limitations and possible extensions.
- A final presentation of approximately 10-15 minutes.

8. Skills acquired by the student

- Practical handling of protein structures, molecular dynamics trajectories and simulation-derived features.
- Basic implementation of OpenMM-based workflows for restrained molecular simulations.
- Design of chemically interpretable probe-based descriptors for ligandability analysis.
- Construction of occupancy maps, density maps and simplified energy-like molecular interaction maps.
- Clustering and ranking of protein hotspots from simulation-derived spatial data.
- Preparation of AI-ready feature tables or graph representations for structure-based drug discovery.
- Critical evaluation of methodological assumptions in molecular simulation and AI-assisted pocket analysis.

9. Suggested software and resources

The project can be implemented in Python using standard scientific and molecular modelling packages. Useful tools may include OpenMM for simulation, PDBFixer for structure preparation, MDTraj or MDAnalysis for trajectory processing, NumPy and pandas for data analysis, SciPy for clustering, scikit-learn for baseline analysis, NetworkX for optional graph construction and matplotlib for visualization. RDKit and the OpenFF Toolkit may be used for handling and parametrizing organic probes, while Packmol can be used to place probe molecules in the simulation box if a full mixed-solvent setup is attempted.

The tutor should provide, or help the student identify, a small protein structure and a feasible computational scope. The project should initially avoid very large proteins, membrane systems or long production simulations unless preprocessed trajectories and computational resources are already available.

10. Evaluation criteria

Criterion	What should be assessed
Scientific clarity	The project question, ligandability concept and chemical interpretation of probe maps are clearly defined.
Technical implementation	The code is readable, documented and reproducible on the selected protein or

Criterion	What should be assessed
	trajectory example.
Map construction	Occupancy, density and energy-like maps are generated using transparent assumptions and appropriate normalization.
AI relevance	The final representation can support clustering, ranking, graph analysis, feature engineering or machine learning.
Interpretability	Probe-specific hotspots are connected to meaningful chemical interaction types and protein regions.
Critical discussion	The student discusses limitations, sampling issues, parametrization assumptions and possible validation strategies.

11. Scope control for a 2-3 month project

The project should remain focused on a reproducible proof of concept rather than a complete production-grade ligandability platform. The student does not need to perform long simulations or implement a fully validated SILCS method. The central objective is to show that probe-based molecular dynamics data can be transformed into interpretable maps and reusable AI-ready descriptors.

- Minimum scope: xenon flooding or xenon trajectory analysis, density map generation, simple hotspot clustering and one pocket ranking table.
- Recommended scope: xenon plus one or two organic probes, three replicas, simplified FragMap scoring and visual comparison of probe-specific hotspots.
- Advanced scope: full mixed-solvent setup with xenon, benzene, isopropanol, acetonitrile and acetamide, composite pocket descriptors and unsupervised clustering of hotspots.

12. Optional extensions

For students with stronger computational skills, the project may be extended by comparing apo and ligand-bound protein states, testing different restraint strengths, integrating docking poses with probe-derived hotspots, generating voxel descriptors for convolutional models, building a residue-hotspot graph, or training a simple classifier on synthetic pocket labels. Another useful extension is to compare static pocket detection with probe-based dynamic ligandability maps to highlight the added value of molecular dynamics.