S3 · 14:45–15:00

Computational Framework for the Design of Second-Generation Pharmacochaperones Targeting Allosteric Pockets

Supervisor: Allegra Via

Alessandra Robello

Sapienza University of Rome

What is protein misfolding?

Missense mutations substitute a single amino acid, destabilising the protein's native conformation and impairing its biological function.
This underlies several rare diseases:

- Phenylketonuria (PKU)
- Alkaptonuria
- Hereditary TTR amyloidosis (hATTR)
- Lysosomal storage disorders

> 7000

Rare diseases known worldwide

Pharmacological Chaperones: 1st vs 2nd Generation

1st Generation

- Bind directly to the active site;
- Compete with the natural substrate;
- Risk of interference with enzyme activity;
- Lower selectivity profile;
- Potential off-target effects.

VS

2nd Generation

- Bind to allosteric pockets;
- Do NOT compete with the substrate;
- Preserve native enzyme activity;
- Higher selectivity profile;
- Reduce off-target effects.

Current Research Challenges

No generation-specific classifier for pharmacological chaperones

Existing tools do not distinguish 1st-generation (pharmacological rescue) from 2nd-generation chaperones (allosteric/non-competitive) — no validated ML predictor exists for this classification.

Binding pocket often unknown — no co-crystallised structure available

When the target is not co-crystallised with a ligand, the allosteric or orthosteric binding site is not experimentally defined — structure-based approaches become unreliable and blind docking carries high uncertainty.

Limited data on allosteric pockets of pharmacological chaperones

The number of experimentally characterised allosteric sites in chaperone targets remains small — limiting ML model training, pocket-descriptor extraction, and the statistical power of any comparative analysis.

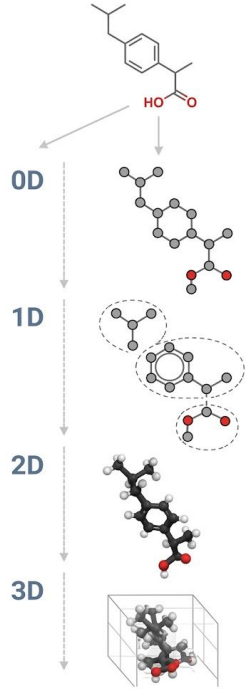
Step 1-2: Descriptor calculation & normalization

Molecular representations

Morgan fingerprints
2048-bit, radius 2

RDKit descriptors
~200 physico-chemical properties

PaDEL descriptors
1800+ advanced descriptors



Normalization

StandardScaler
zero mean, unit variance

MinMaxScaler
range [0,1]

No scaling
applied when not needed

Step 3: Feature selection to select only features with significant differences between classes

Remove constant columns

Remove any constant value: it does not provide information.

X (all features)

	f_1	f_2	f_3	...	f_n
mol 1	1.2	5.0	7.1	...	3.3
mol 2	0.8	5.0	2.4	...	1.2
mol 3	1.5	5.0	0.3	...	4.7
...	...	...	...	...	...
mol N	2.1	5.0	1.8	...	2.9

f_2 is constant (unique value = 5.0) → removed

Normality test

For each feature, test whether the distribution is normal.

$p > 0.05$

if n molecules ≥ 5000: D'Agostino - Pearson

if n molecules < 5000: Shapiro - Wilk test

Differential statistical test

Distribution	Number of classes	Test applied
Normal	2 (binary)	Welch's t-test (unequal variances)
Normal	> 2 (multiclass)	One-way ANOVA
Not normal (any)	any	Kruskal–Wallis

Select top k by p-value

Features are sorted by increasing p-value.

Ranking by p-value (ascending)

Rank	Feature	p-value
1	f_3	1.2e-06
2	f_{17}	3.4e-05
3	f_8	6.7e-05
...	...	...
k	f_{56}	2.1e-03
k+1	f_{23}	4.8e-02
...	...	...
p	f_{102}	7.2e-01

Classification Pipeline

Step 4: Model training

Phase 1
100 combinations
On all 5 algorithms

Phase 2
1000 combinations
Top 3 algorithms only

Phase 3
10000 combinations
Single winning algorithm

5 Algorithms

Random Forest	Gradient Boosting	SVM	Decision Tree	KNN
Ensemble of distinct models	Optimised sequential learning	Maximises classification margin	Hierarchical decisions	Instance-based clustering





Random Sampling

- **Model hyperparameters:**
  estimators, neighbors, lr, C, max_depth...
- **Preprocessing params:**
  scaling (Standard/MinMax), PCA variance

Selection Phase

Phase 1 →	top 3 algorithms selected
Phase 2 →	top 1 algorithm selected
Phase 3 →	optimal model & threshold

Score: Full Score

Harmonic mean of Pred Score and CV Score. Penalises high test score with low cross-validation (and vice versa).

Hyperparameter optimization: Optuna

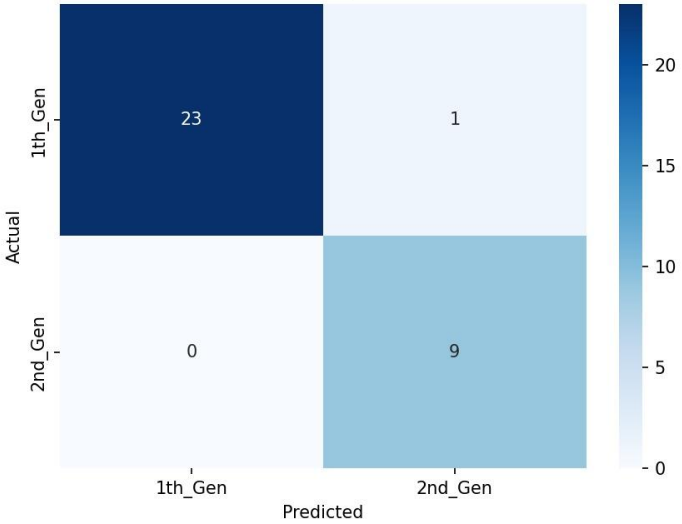
Bayesian optimisation (TPE) over ±50% range of Phase 3 best hyperparameters → a new model is trained with the refined parameters → best Full Score wins.

Results

Test

MCC: 0.93

APC 1th Gen: 95.8%, 2nd Gen: 100%



Dataset: 166 molecules

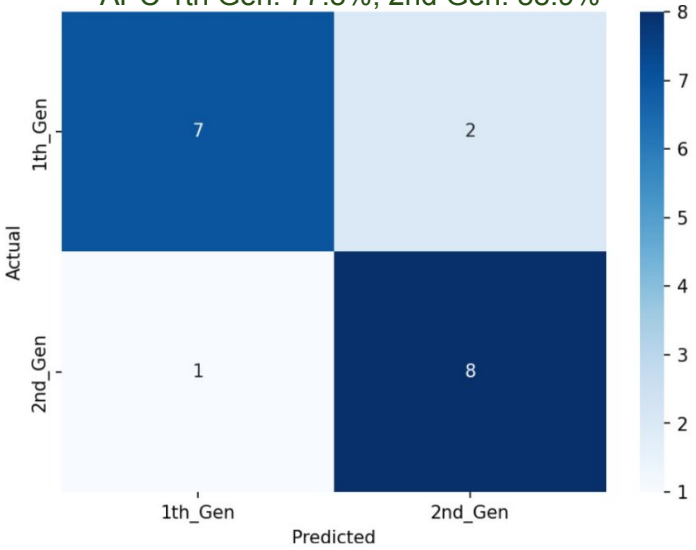
Gradient Boosting

RDKit

Validation

MCC: 0.67

APC 1th Gen: 77.8%, 2nd Gen: 88.9%



Target Identification Pipeline

01 Input

PDB Retrieval

Structures co-crystallised with an allosteric ligand are retrieved from the Protein Data Bank. Primary amino acid sequence is extracted for downstream analysis.

02 Genomics

BLAST Homologue Search

The sequence is submitted to BLAST to identify proteins with sequence and functional similarity — broadening the target space beyond the query protein.

03 Structural

Pocket Characterisation

Allosteric pockets are described geometrically and chemically using Graphein (protein graphs) and STING descriptors encoding hydrophobicity, electrostatic potential, and solvent accessibility.

04 Output

AlloBench Query

Pocket descriptors are used to query the AlloBench allosteric protein database. Proteins sharing analogous pocket features are retrieved as candidate targets.

No co-crystallised structure → fpocket predicts cavities via Voronoi spheres and compares them against known target pockets (Track 2).

Target Identification Pipeline

CSB Meeting 2026

Molecular Similarity Search

1. **Known allosteric molecules as queries**

   A curated set of experimentally validated allosteric compounds is used as the starting point for similarity-based virtual screening.

2. **Morgan fingerprints**

   Circular fingerprints with radius 2 encode the local chemical environment of each atom, capturing substructural features relevant to binding.

3. **RDKit descriptors**

   Physicochemical descriptors (MW, LogP, H-bond donors/acceptors, rotatable bonds…) complement the topological fingerprints for a richer representation.

4. **Compound database screening**

   Similarity scores are computed against large compound libraries to identify structurally related candidates as potential new 2nd-generation pharmacological chaperones.

Pocket Prediction (no crystal)

fpocket

Voronoi sphere-based algorithm that identifies and scores surface and internal cavities on the protein structure.

Cavity scoring & filtering

Pockets are ranked by druggability score, volume, and hydrophobicity. Only top-ranked cavities are retained.

Cross-comparison

Predicted pockets are compared against characterised allosteric sites from Track 1, extending the search to poorly annotated proteins.

Integration of both tracks: structural, genomic + cheminformatics strategies define a reproducible, scalable pipeline for computational prioritisation of targets and molecules.

Future prospective

1

Computational Pipeline

(Previous phase)

- ML/DL classification (GNN)
- Allosteric target identification
- Molecular candidate screening

2

Molecular Dynamics

Integration of results

- MD simulations of protein-ligand complexes
- Allosteric pocket stability (RMSD, RMSF)
- Prioritisation by binding affinity and stability

3

In vitro Validation

Selected target protein

- Selection of optimal target protein
- Thermal stability assays (TSA/DSF)
- Binding experiments (SPR, ITC)
- Functional activity assessment

Final goal: identify a 2nd-generation allosteric chaperone with demonstrated stabilising activity in vitro

Thank you for the attention!

- Prof. Allegra Via
- Dr. Damiano Parrone
- RCMD lab

ROME CENTER FOR MOLECULAR DESIGN

SAPIENZA
UNIVERSITÀ DI ROMA