S3 · 14:30–14:45

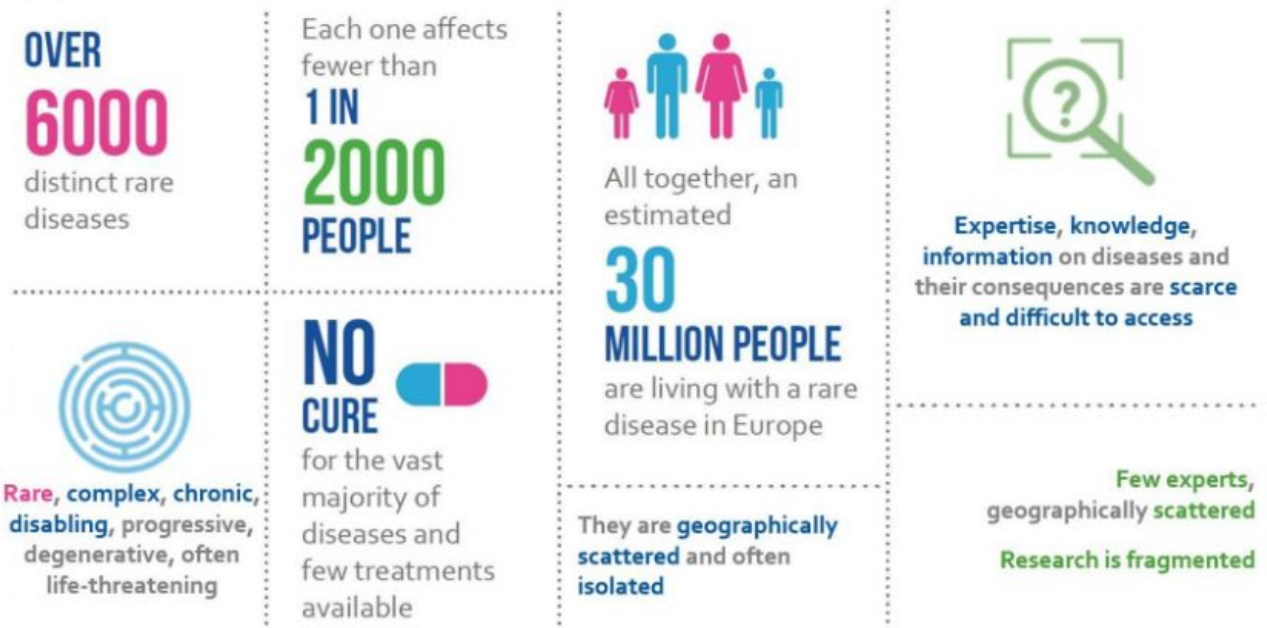
# A Prediction Tool for the Inference of Rare Misfolding Variants Responsive to Pharmacological Chaperones

**Damiano Parrone**

*Sapienza University of Rome*

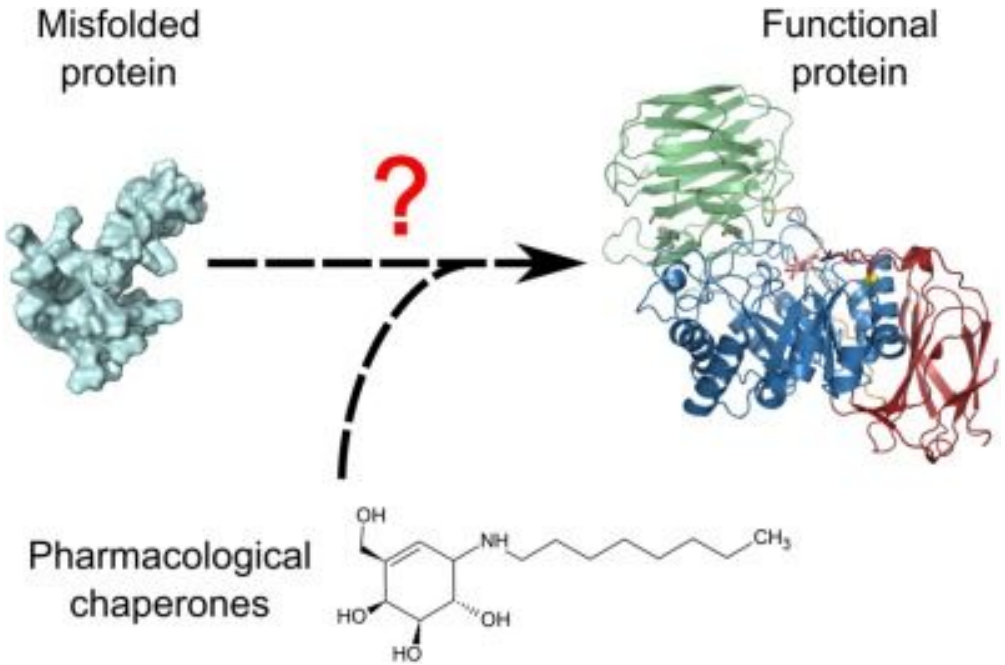
# Rare diseases: an overview





# Pharmacological chaperones (PCs)






First generation: **inhibitory** (e.g. Migalastat, Ambroxol, Miglitol)

Second generation: **non-inhibitory** (e.g. Ciclopirox, Ebselen)

# Aim of the project



Development of a **machine learning**-based model to **predict** the susceptibility of variants associated with rare misfolding diseases to pharmacological chaperones

# Data collection



| Data | Source |
|---|---|
| 180 pharmacochaperones | Literature, Drugbank, PubChem, ChemSpider |
| 30 Proteins | Uniprot |
| 30 Diseases | Orphanet, Uniprot |
| 6683 Variants | Uniprot |
| 960 Efficacy Tests | Literature |

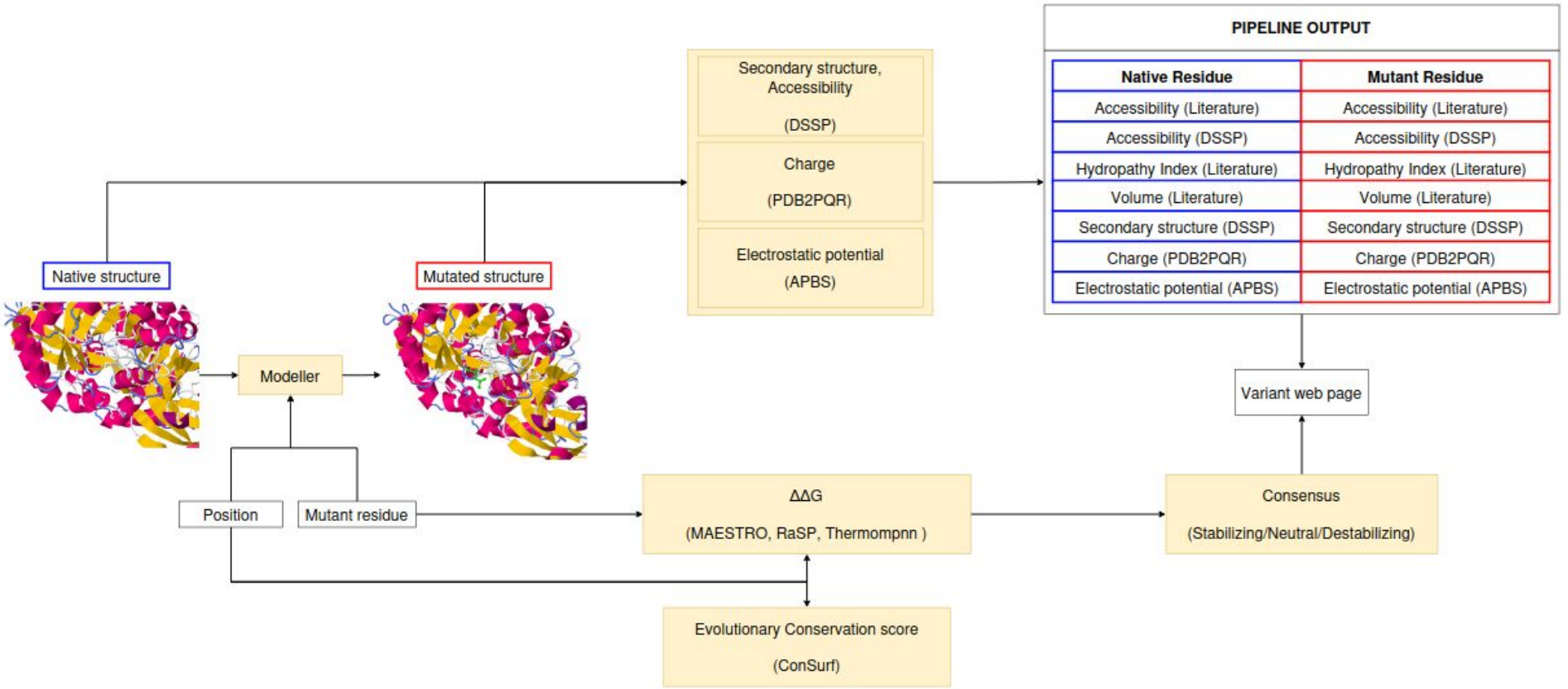
# Data collection



| Data | Source |
|---|---|
| 180 pharmacochaperones | Literature, Drugbank, PubChem, ChemSpider |
| 30 Proteins | Uniprot |
| 30 Diseases | Orphanet, Uniprot |
| **6683 Variants** | Uniprot |
| 960 Efficacy Tests | Literature |

# Variant annotation and database construction





# Web interfacing




https://bioinfo.bio.uniroma1.it/circle

# Machine learning: definition



- Machine learning (ML) is a statistical technique used to **identify patterns** in data and **make predictions** without explicit, deterministic programming
- It’s a data-based approach
- All non-numerical data must be converted into **numbers**

The output of ML is a **model** capable of predicting outcomes on the basis of the patterns observed during the training phase

# Machine Learning: dataset extraction and pre-processing



- Our ML dataset has been extracted from CIRCLE
- Each data point is an **efficacy test** of a PC performed on a variant associated with a rare conformational disease
- The data has been cleaned of duplicates and missing values

In total, the dataset amounts to **960** data points: 699 **effective** (variant activity is restored by the PC) and 261 **ineffective** efficacy tests

# Machine learning: dataset description

A ML dataset is a table of rows (data points) and columns (features or variables)

| | Variant data | | Chaperone | … | Outcome |
|---|---|---|---|---|---|
| | **Sequence** | **Substitution** | **SMILES** | **…** | **Efficacy** |
| **Row 1** | VMLAC… | A594T | C[C@@H]... | … | Effective |
| **Row 2** | MLLYT… | M30V | C1=CC=C… | … | Not Effective |
| **Row 3** | PCSGI… | C164R | [H][C@@]... | … | Effective |
| **…** | … | … | … | … | … |

# Machine Learning: types of data and encoding



1. **Numerical variables**: e. g. residue volume, residue solvent accessibility

2. **Categorical variables**: e.g.  secondary structure, chaperone class

3. **Symbolical variables**: protein variant sequence and pharmacochaperone

# Machine Learning: the “curse” of dimensionality

The issue with encoding is the **increase** in the number of variables

For instance

Before encoding: **38** variables

The issue with encoding is the **increase** in the number of variables

For instance

Before encoding: **38** variables

After encoding: 2000 variables

# Machine Learning: drug encoding

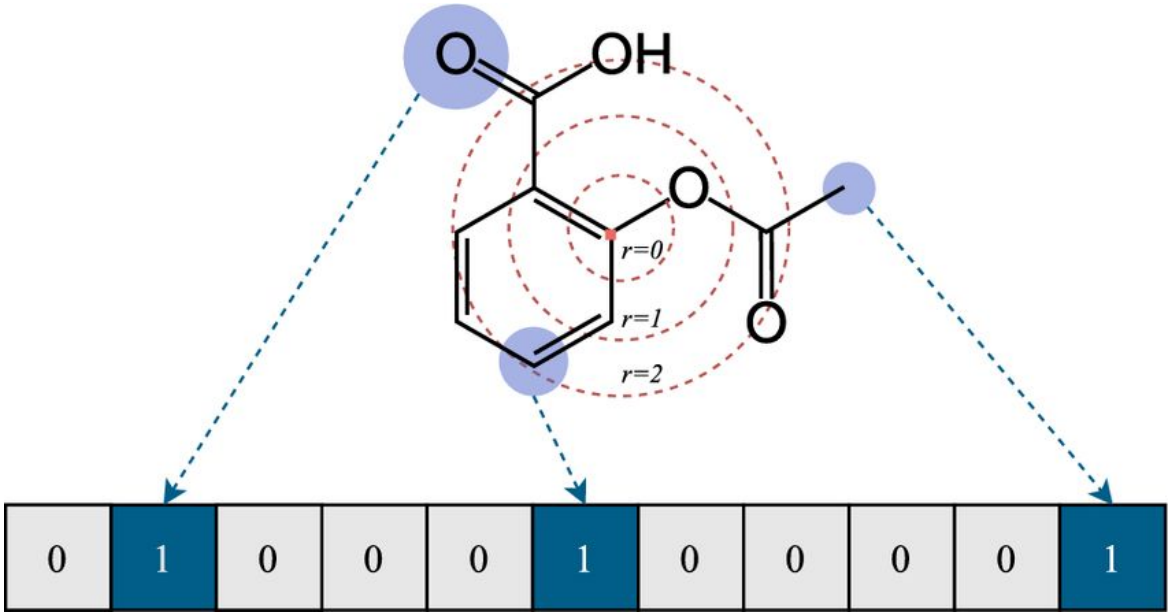


- In computational chemistry, molecular structures are usually represented in **SMILES** format
- SMILES (Simplified Molecular Input Line Entry System) is a compact string-based representation of a compound
- A SMILES must be **converted** into numbers for ML applications

Example: Aspirin ➤ CC(=O)OC1=CC=CC=C1C(O)=O

# Machine Learning: drug encoding

Morgan Fingerprint: a **binary** vector representing the absence (0s) or presence (1s) of chemical substructures



# Machine Learning: protein sequence encoding



- Protein sequence ➤➤➤ ProtBERT and ESM **embeddings**
- Embedding = vector of numbers representing a single amino acid residue
- Average of all amino acid embeddings ➤➤➤ **compact** sequence embeddings of 1024 (ProtBERT) and 1280 (ESM) values

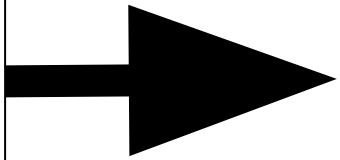
# Machine Learning: dataset versions



# Machine Learning: double cross-validation



24 Datasets



| Tested algorithms |
| --- |
| eXtreme Gradient Boosting |
| Gaussian Naïve Bayes |
| K-Nearest Neighbour |
| Logistic Regression |
| Support Vector Machine |
| Decision Tree |
| Random Forest |

# Machine Learning: results



Best sequence encoding: **ProtBERT**; Best drug encoding: **64 bit** fingerprint

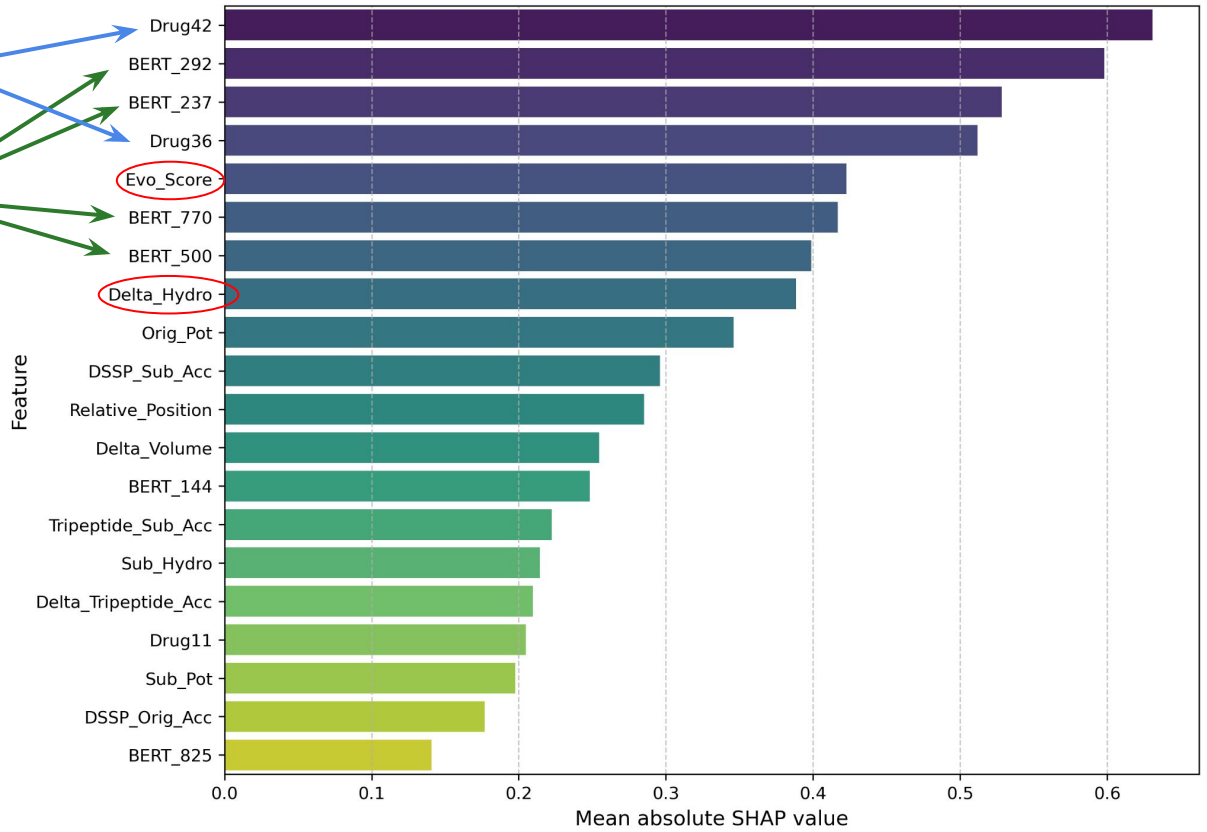
| Algorithm | Accuracy | F1 Score |
|---|---|---|
| **eXtreme Gradient Boosting** | **84%** | 0.89 |
| Random Forest | 82% | 0.88 |
| Decision Tree | 79% | 0.86 |
| Logistic Regression | 78% | 0.86 |
| K-Nearest Neighbour | 76% | 0.85 |
| Support Vector Machine | 74% | 0.85 |
| Gaussian Naïve Bayes | 70% | 0.77 |

# Machine Learning: feature importance



Chemical substructures

Sequence embeddings



# Future prospects



1. **Further increase** ML model performance

2. **Make available** ML models to the scientific community through the CIRCLE web platform

3. **Use** of developed resources and tools to study second generation PCs binding pockets and mechanisms of action

# THANK YOU

**Allegra Via**

**Claudio Carta**

**Federico Mari**

**Anna Marabotti**

**Stefano Pascarella**