Machine Learning in Protein Solubility Engineering

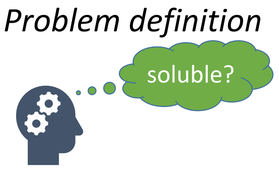
Jan Velecký | velda@mail.muni.cz | live:vvelda

LOSCHMIDT LABORATORIES

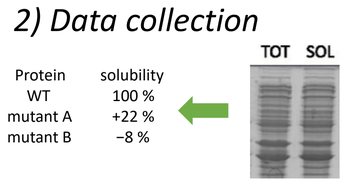
Machine learning

Not only does ML prove a hypothesis but it also finds this hypothesis

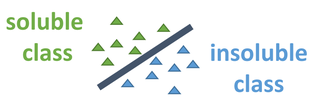
1) Problem definition



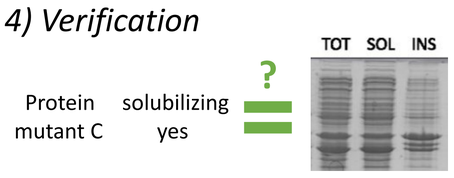
2) Data collection



3) Model training



4) Verification



Use cases:

- **traditional methods are insufficient**
- possession of **big data** about the problem
- **high-dimensionality** of the problem

Solubility engineering

Where solubility matters:

- Proteins manufacturing – the higher solubility, the higher yield
  Insoluble protein is usually aggregated or having other abnormality and **is short of the** intended **function** (bioenzyme), alternatively may even turn **harmful** (drug)
  Insoluble fraction = waste
- Crystallization experiments
  For proteins without known structures, too **low/high solubility impedes crystallization** experiments
- Disease prediction
  Revealing genetic mutations causing **low-solubility-related diseases**

Simplistic definition:
"Degree to which a substance dissolves in a solvent to make a solution"

Multiple usages:

- **soluble fraction** [%]
- soluble expression [g/l] = yield
- expression/expressibility [g/l]
- aggregation propensity
- binary solubility
  above/below a specific level

⇩

ambiguous data

The state of data

Old & overlapping **datasets** for predictions: CamSol, OptSolMut, PON-Sol, A3D:

- **tens** of proteins
- **hundreds** of mutations
- often low-soluble proteins

Deep mutational scanning (high-throughput) data:

- **units** of proteins
- **thousands** of mutations

Common characteristics:

- solubility change upon mutation
- mostly desolubilizing

SOLUPROTMUT DB

The 1st database of protein solubility upon mutation

UNDER DEVELOPMENT: the preview version expected in June 2021

Jan Velecký, Stanislav Mazurenko, Marie Jankůjová, Jan Štourač, David Bednář

Highlights:

- Solubility-prediction tools could perform better
- Low solubility affects yields in industry
- No big data regarding solubility present at single place
- Loosely defined data make machine learning challenging

The state of tools

Several overall-solubility predictors – usually **cannot predict effects of mutations**.

A few solubility-change predictors:

- SOLPro (2009)
- OptSolMut (2010)
- CamSol (2014)
- PON-Sol (2016)
- SODA (2017)

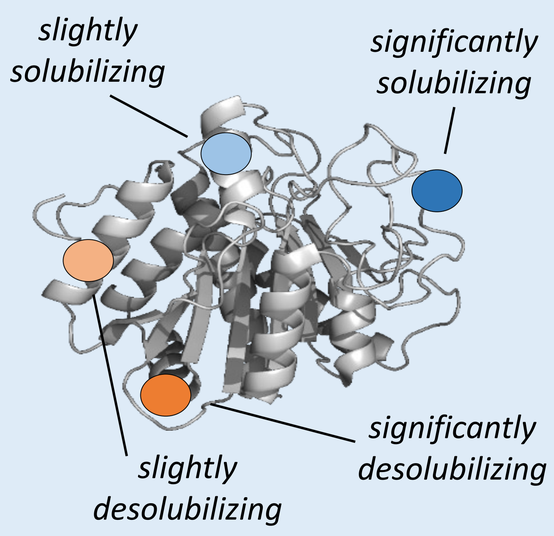
Accuracy of 60–80 %

None utilize big data from deep mutational scanning experiments.

Mutational solubility data

Effect of a mutation(s) on the protein's solubility:

- mutation site & AA* substitution
- effect classification
  −− − N + ++ (≤ 5 bins)
  for ML classification models
- conditions of the experiment
- pre-computed (HotSpot Wizard) per-AA features, like:
  residue conservation, accessible surface area, …



Systemization of reported values in the literature:

- usually, **discreet, loose values** are reported
- nonetheless, reports follows one of the schemes (columns) below
- the table also defines **comparability** between different schemes

	reported change					real change
	unipolar	2-value	3-value	4-value	5-value	
++	enhancing			significantly enhancing		++
+				slightly enhancing		+
N	non-enhancing (NE)		neutral		neutral	neutral
−		deteriorating		slightly deteriorating		−
−−				significantly deteriorating		−−

Objectives

- Single and complete source for solubility mutagenesis
- Ready-made for ML* training with labels and features
- Error-free data
- Enhance accuracy of solubility prediction
- Invite out-of-the-field ML experts to try their models on these data

Functionality

- On-the-web browsable DB
  advanced search filters
- Export tool for data scientists
  values conversion to a target scheme, data augmentation using symmetrisation

Preview in June 2021

Outlook

- Target the special database issue in Nucleic Acid Research
- Conduct high-throughput experiments in Loschmidt Laboratories
- Train our own predictive models on these data
- Combine solubility and other pressing tasks in protein engineering

* AA: amino acid | DB: database | ML: machine learning

MUNI | RECETOX