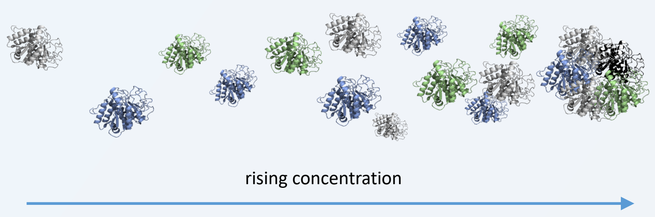
Protein Learning for Solubility Design

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

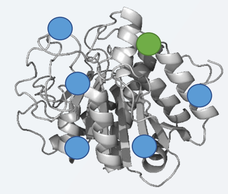
LOSCHMIDT LABORATORIES

Motivation



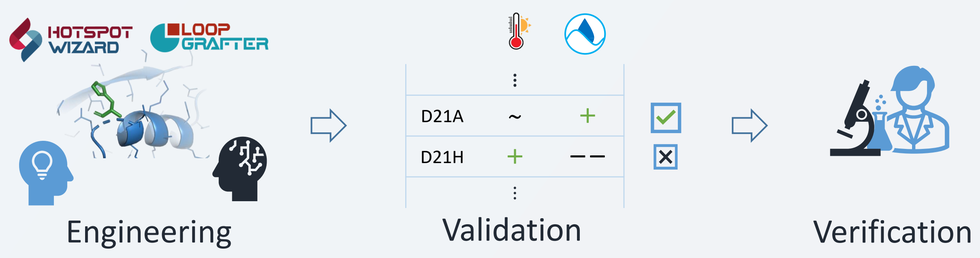
Proteins usually occur in low concentrations **on the edge of their solubility** limits in nature. **High concentrations** thus cause aggregations or molecular crowding. Insoluble protein **loses its function**. Lower solubility also means lower yields in **protein production** in chemical or pharmaceutical industries.

Solubility & its importance



- **Ability to freely dissolve in a solution**
- Evolutionary stabilized solubility of natural proteins
- Design of protein properties = **choice of mutations**

Protein design



Engineering — Validation — Verification

Solubility predictors

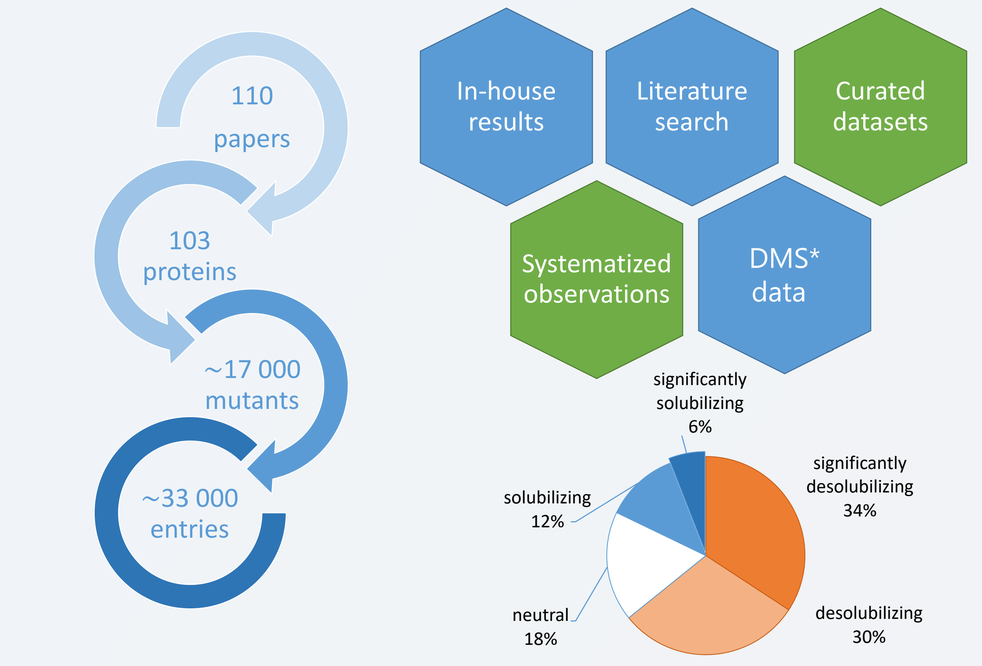
Solubility-change predictors:
- OptSolMut (2010)
- CamSol (2014)
- PON-Sol (2016)
- SODA (2017)
- PON-Sol2 (2021)

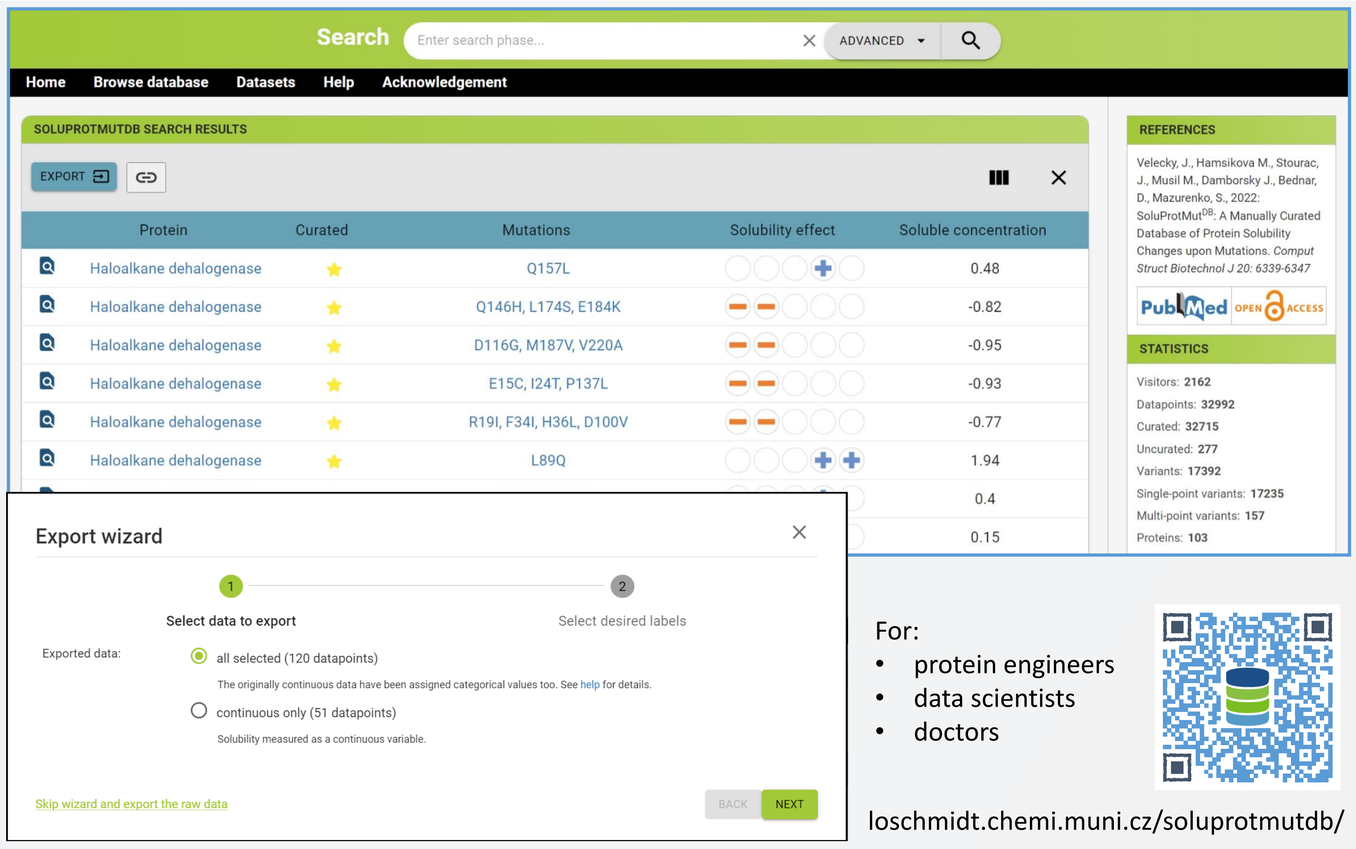
Reported **accuracy of 60–80 %**. Even lower solubilizing precision.

Soluble protein predictors:
- Either soluble, or insoluble
- SoA* balanced accuracy up to 59 %
- Mutational prediction **up to 57 %**

Data: the database of mutational solubility data

SOLUPROTMUT DB



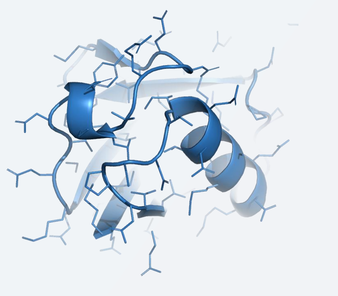


For:
- protein engineers
- data scientists
- doctors

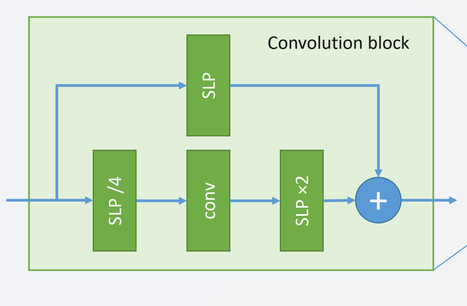
loschmidt.chemi.muni.cz/soluprotmutdb/

Model: protein convolutional neural network

Representation



Convolution

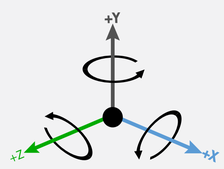


$$\mathrm{conv}(\vec{x}) = \sum_{\vec{y} \in B_r(\vec{x})} \boldsymbol{F}_y \times \boldsymbol{\kappa} \circ (d, \tau_C, \tau_N)(\vec{x}, \vec{y})$$

Architecture

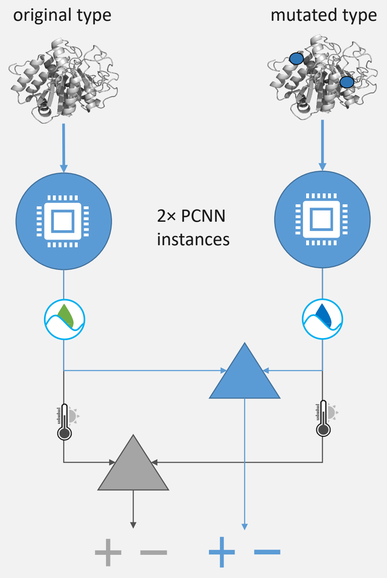
- Graph-based convolutional network for **protein learning**
- Vertex-weighted multi-graph representation of protein structure
- **SoA* prediction** of EC* numbers
- Rewritten, simplified and optimized under the author's supervision
- Expertise brought from *Dr. Pedro Hermosillia* from *Visual Computing Group, Ulm University, Germany*

- Protein Convolutional Neural Network (PCNN) displays perceptivity to **local spatial features**
- Meanwhile, is **invariant to rotations & translations** of a protein structure
- Appears to be a perfect model for mutational predictions: stability, solubility, **dual-objective**



Action plan

Model training



dual-objective prediction

Augmenting data

Reverse mutations		two-fold
Stability changes		two-fold
Soluble/insoluble		two-fold
PDBs*		30-fold
EC* numbers		three-fold

Protein design:
Vašina M, **Velecký J**, Planas-Iglesias J et al. (2022) *Adv. Drug Delivery Rev. 183:114143*

Data – SoluProtMutDB:
Velecký J et al. (2022) *Comput. Struct. Biotechnol. J. 20:*6339-6347

Machine learning model:
Hermosilla P et al. (2021) *ICLR 2021 Conference*

* **Abbreviation glossary:**
DMS: deep mutational scanning
EC number: Enzyme Commission number
MLP/SLP: multi-/single layer perceptron
PDB: Protein Data Bank entry
SoA: state of the art