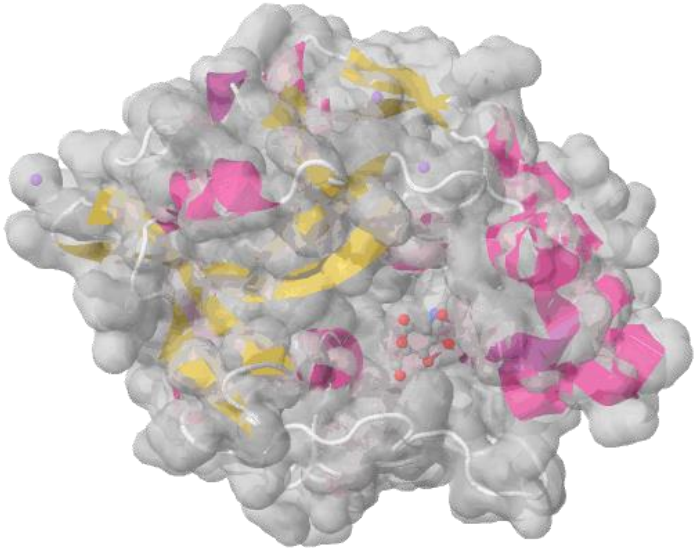
Využití strojového učení k identifikaci protein-ligand aktivních míst



David Hoksza, Radoslav Krivák

SIRET Research Group

Katedra softwarového inženýrství,

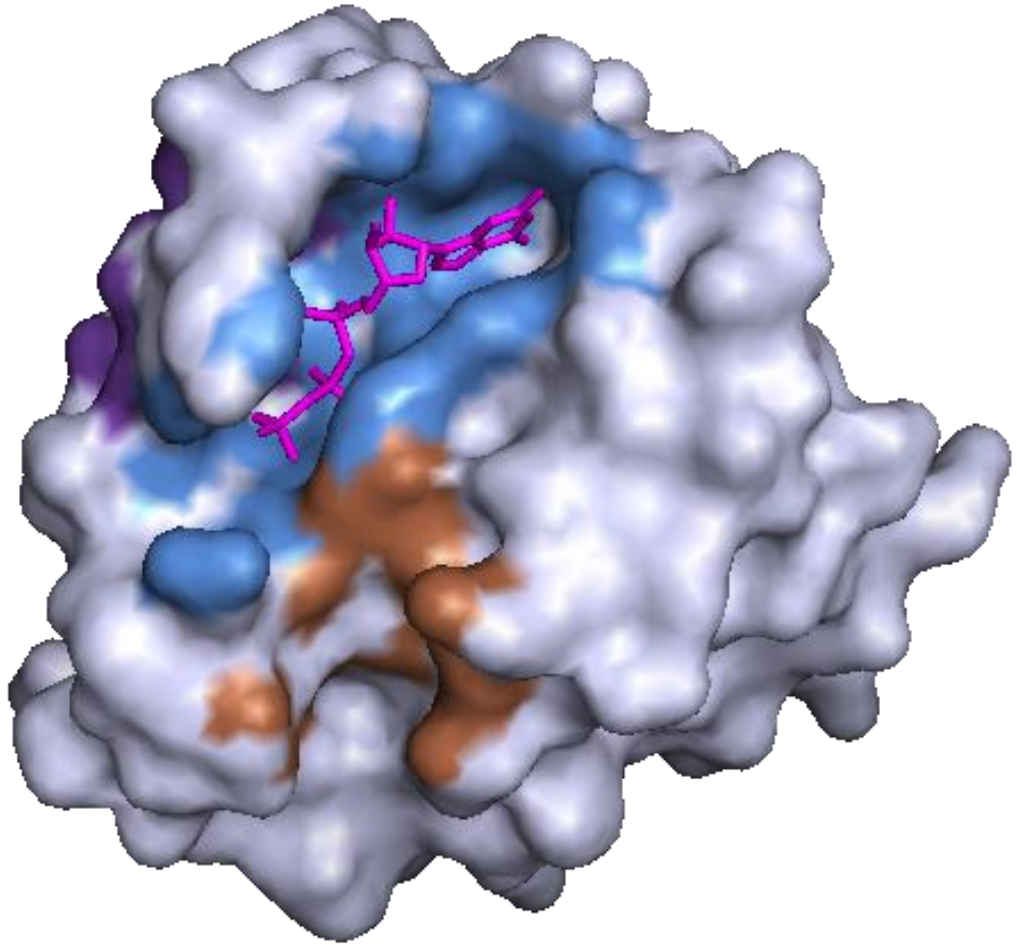
Matematicko-fyzikální fakulta

Karlova Univerzita v Praze

Funkce proteinu

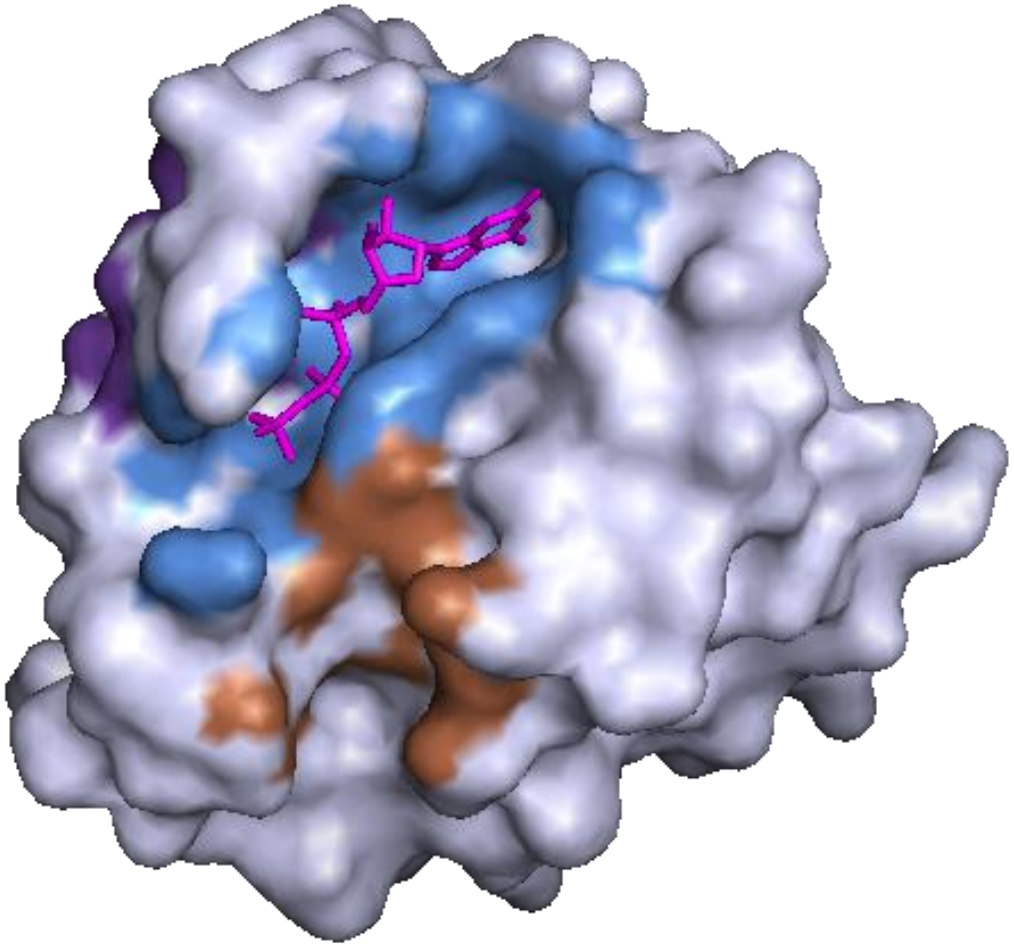
- Interakce s dalšími molekulami
  - DNA
  - RNA
  - Proteiny
  - **Malé molekuly (ligandy)**

- Určena strukturou
  - Distribuce fyzikálně-chemických charakteristik v prostoru → **princip zámku a klíče**



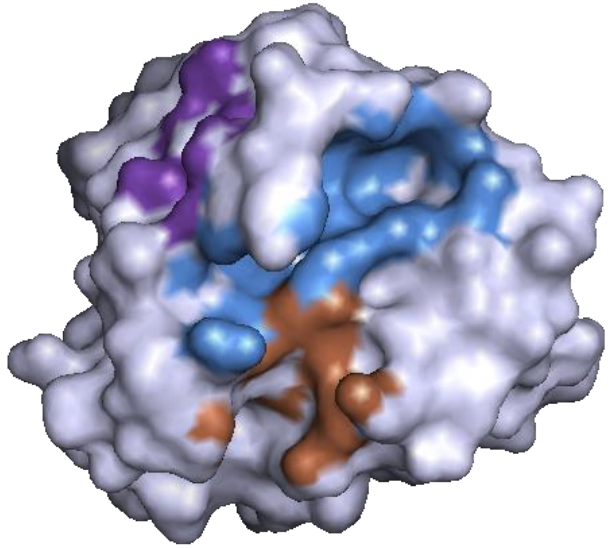
Protein-ligand interakce

- **Aktivní místa (kapsy)**
- **Motivace**
  - **Predikce funkce** neznámého proteinu
  - Identifikace potenciálních **cílů pro léčiva**
  - Predikce **vedlejších účinků** léčiv



P2RANK

- Počítačová metoda (algoritmus) schopný **identifikovat místa na povrchu proteinu**, na které se může s vysokou pravděpodobností **vázat nespecifikovaný ligand**



- **Vstup**: počítačová reprezentace proteinové struktury
- **Výstup**: seznam míst na povrchu proteinu pravděpodobně schopných vázat ligand

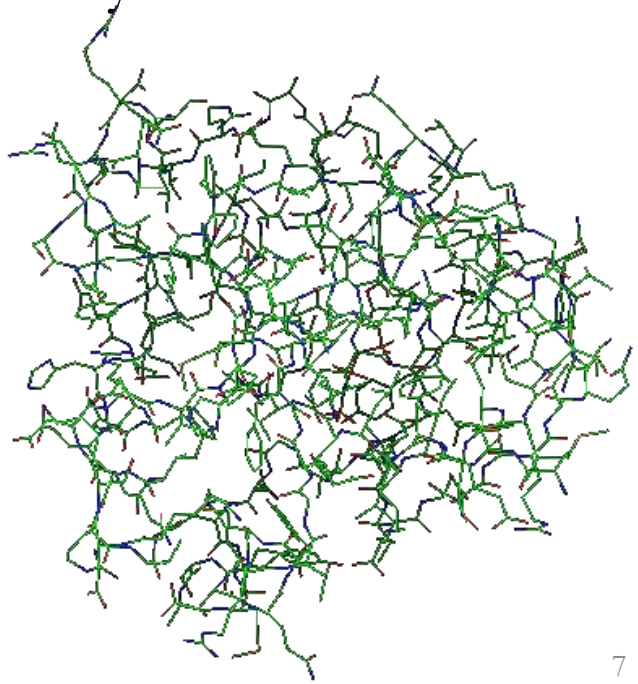
Informatický pohled na protein

Sekvence

- **Řetěz** aminokyselin → **lineární sekvence** písmen (slovo)
  - Písmena reprezentují aminokyseliny (ARNDCEQGHILKMFPSTWYV)

Struktura

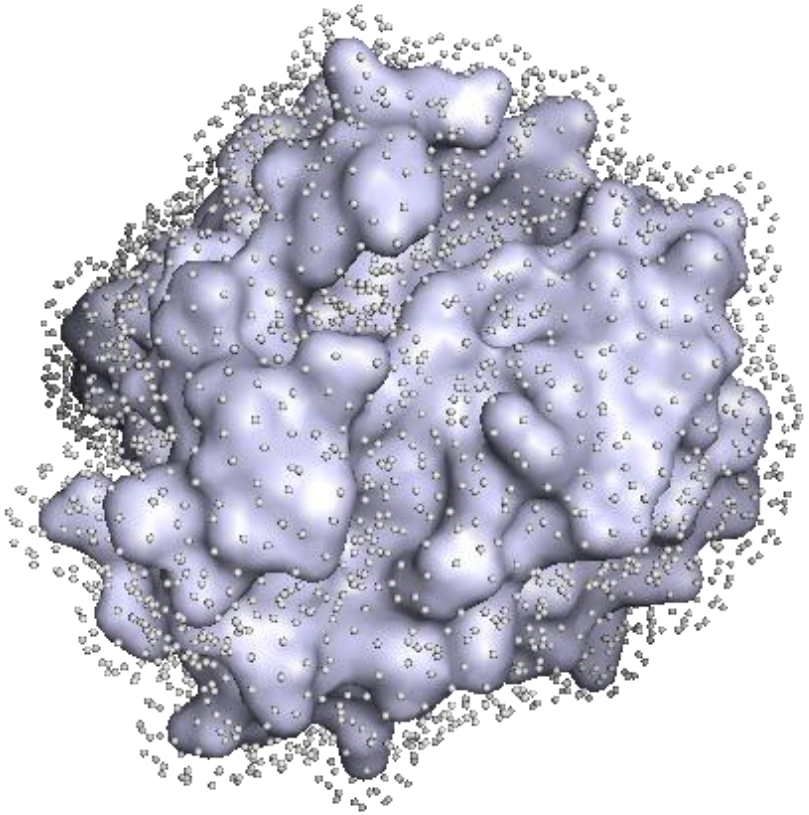
- **Pozice** jednotlivých atomů v 3D prostoru



P2RANK princip

- Využití informací o existujících aktivních místech pro rozpoznání typově podobných míst na neznámém proteinu
  - **Jak popsat rysy povrchu proteinu?**
    - Projekce fyzikálně chemických vlastností aminokyselin na povrch proteinu
  - **Jak určit kapsu na dosud neviděném proteinu?**
    - Strojové učení (s učitelem)
      - Fáze učení: naučení modelu pro rozpoznání rysů bodů aktivních míst
      - Fáze rozpoznávání: aplikace modelu na povrch neznámého proteinu

Algoritmus – učící fáze



1. Získání známých protein-ligand komplexů

2. Potažení povrchů proteinů **sítí bodů**

3. **Extrakce vektoru fyzikálně-chemických vlastností** pro každý z bodů každého proteinu

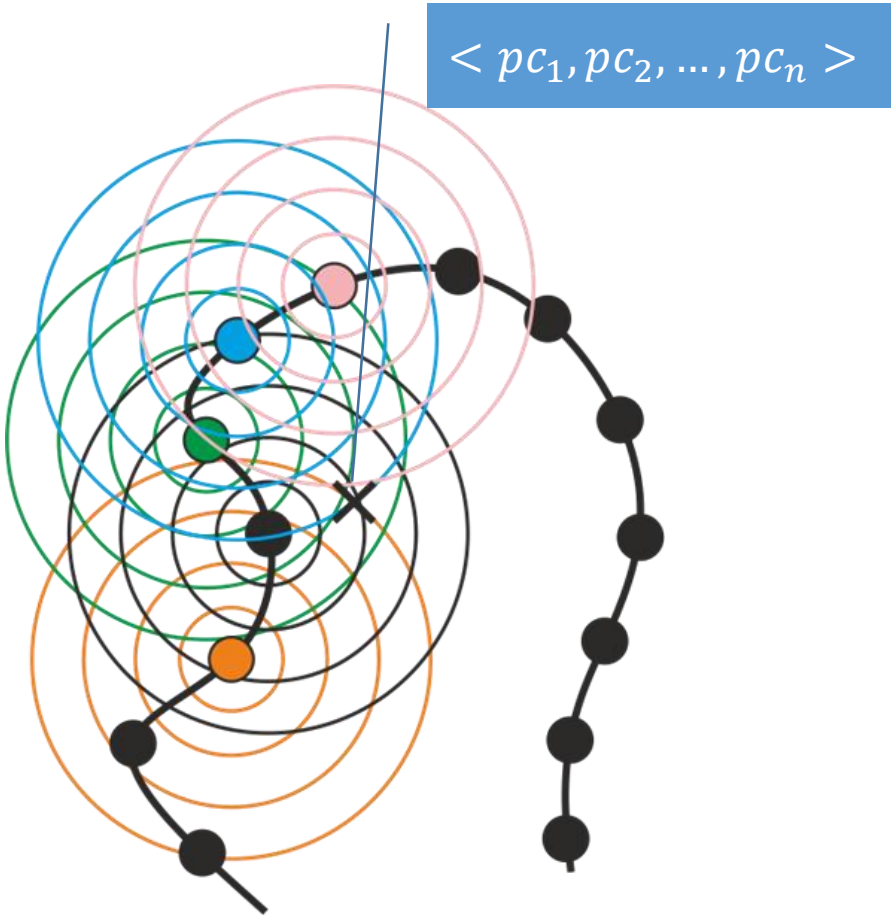
4. **Vybudování modelu**, který pro daný bod (vektor) bude schopný určit, s jakou pravděpodobnostní je tento součástí kapsy = **strojové učení**

P2RANK - extrakce vlastností

- Okolo **30 atributů** popisující fyzikálně-chemické vlastnosti aminokyselin a lokálního okolí daného bodu

$$\mathrm{PC}(V) = \frac{1}{m} \sum_{\mathrm{A}_i \in \mathrm{A}(V)}^{m} \mathrm{PC}(\mathrm{A}_i) \cdot w(dist(V, \mathrm{A}_i))$$

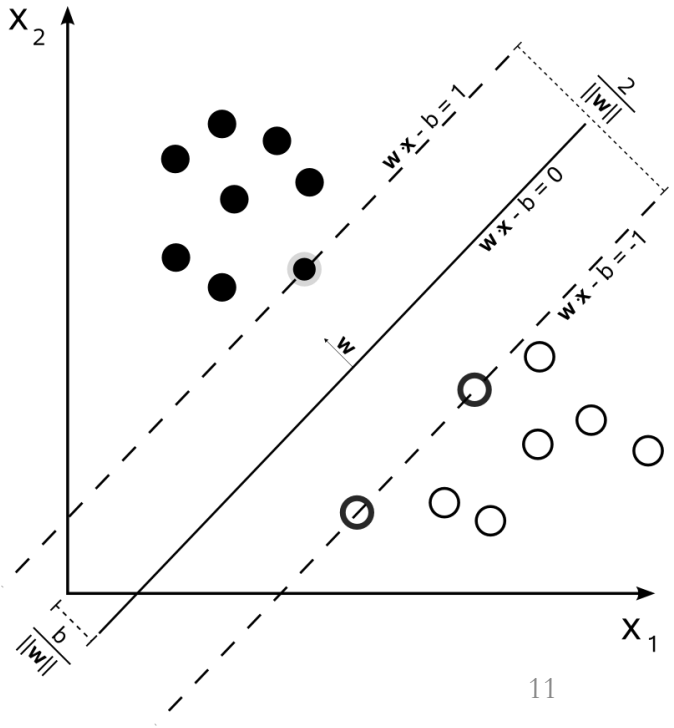
$$w(d) = \begin{cases} 1, & d \leq 4\ \text{Å} \\ (4/d)^2 & d > 4\ \text{Å} \end{cases}$$

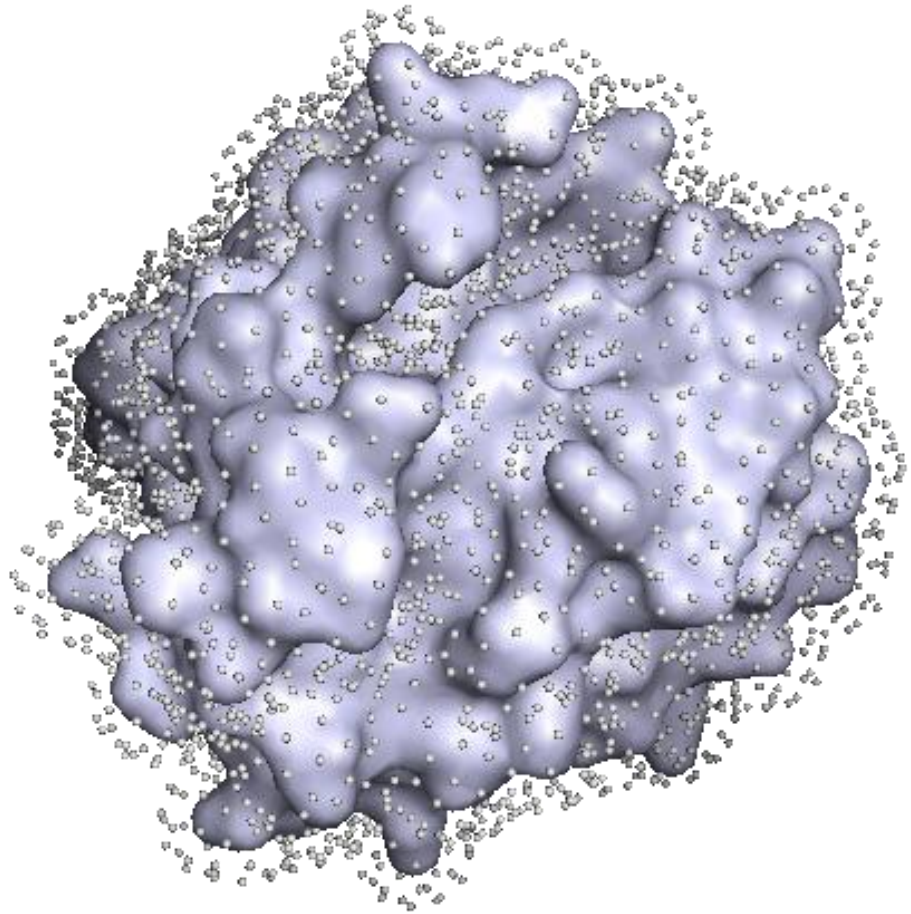


Strojové učení

- Skupina algoritmů schopných **identifikovat (naučit se) vzory v datech** a a tuto znalost **aplikovat na dosud neviděných příkladech**
  - **Klasifikace**
  - Regrese
  - Shlukování

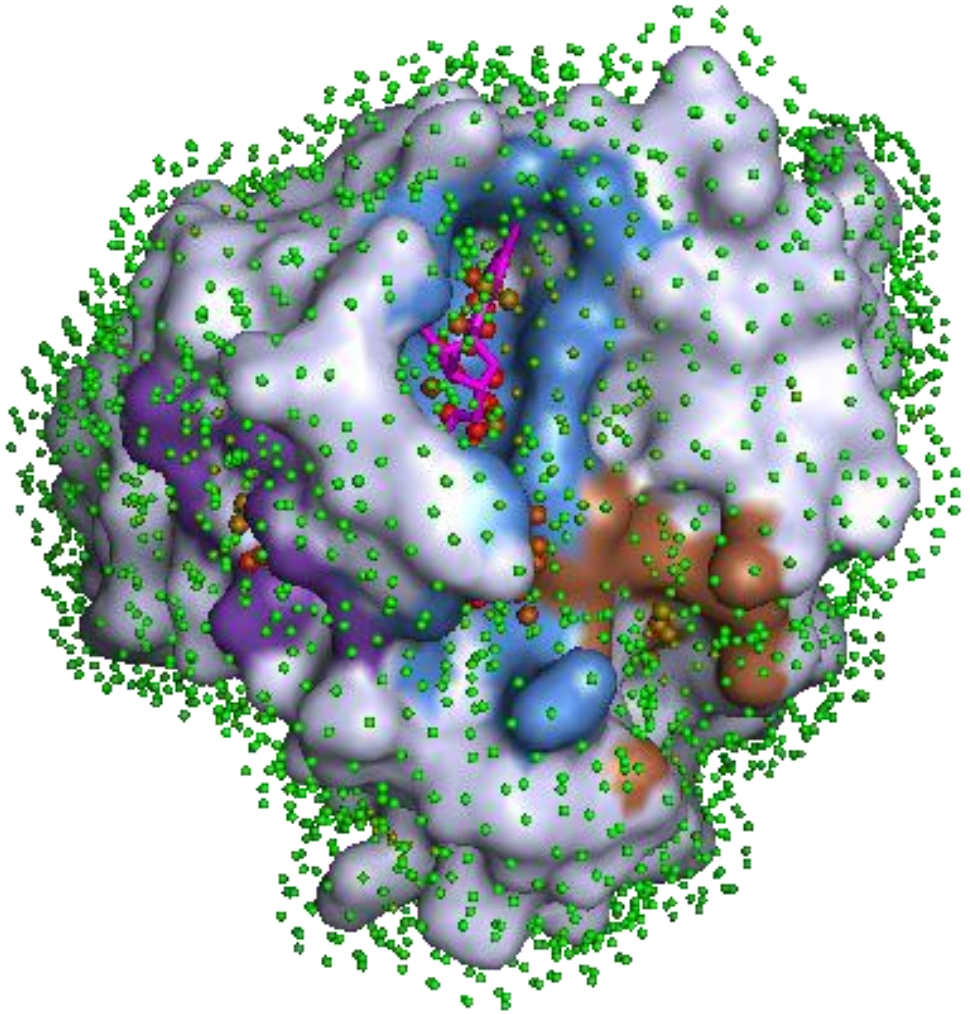
- Typy
  - **Strojové učení s učitelem (supervised learning)**
  - Strojové učení bez učitele (unsupervised learning)





Algoritmus – fáze rozpoznávání

1. **Potažení povrchu** neznámého proteinu **sítí bodů**
2. **Aplikace modelu** pro každý bod sítě → vazebné skóre bodu
3. Vypuštění bodů s nízkým vazebným skórem
4. **Identifikace shluků** vysoce skórujících bodů → **kapsa**
5. **Ohodnocení** kapes součtem skór jejich bodů



Jaké znalosti jsou třeba pro vývoj P2RANKu?

- Základy biologie, proteomiky
- Znalost zdrojů biologických dat (PDB)
- Programování
- Pokročilá algoritmizace (strojové učení)
- Statistika

Bioinformatické projekty na KSI MFF UK

- **Proteiny**
  - Podobnostní hledání proteinových struktur – P3S
  - Identifikací protein-ligand interakcí – P2RANK
  - Identifikace protein-protein interakcí
  - Identifikace proteinových sekvencí ze spektrometrických dat - SIMTANDEM
- **RNA**
  - Podobnost RNA struktur – SETTER
  - Predikce sekundární struktury RNA - rPredictor
- **Malé molekuly**
  - Identifikace biologicky aktivních molekul explorací chemického prostoru - Molpher

Dotazy