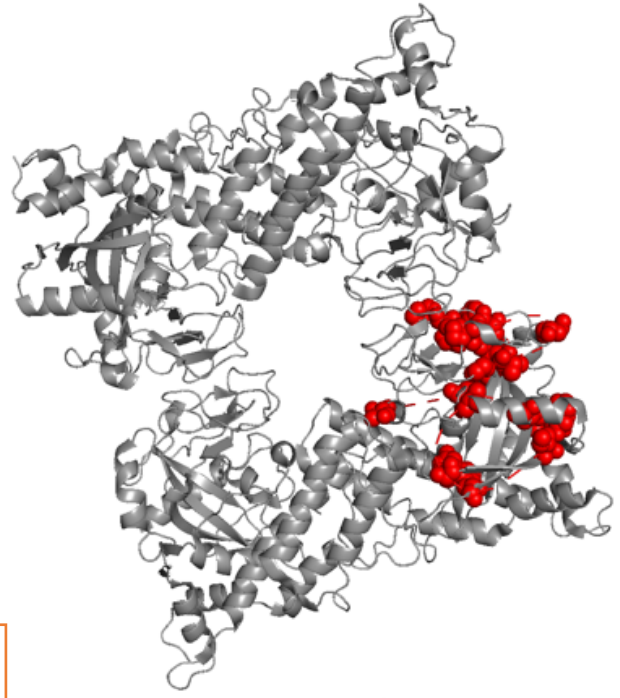
# Identifikácia proteínových sektorov ako evolučných jednotiek bielkovín

Martina Babinská
Školiteľ: prof. RNDr. Ľubomír Tomáška, DrSc.
Konzultant: doc. Mgr. Bronislava Brejová, PhD.

# Úvod do problematiky

- Proteíny
- Konzervovanosť
- Koevolúcia
- Proteínový sektor
- PARP proteín z kvasiniek
  - Katalytická doména



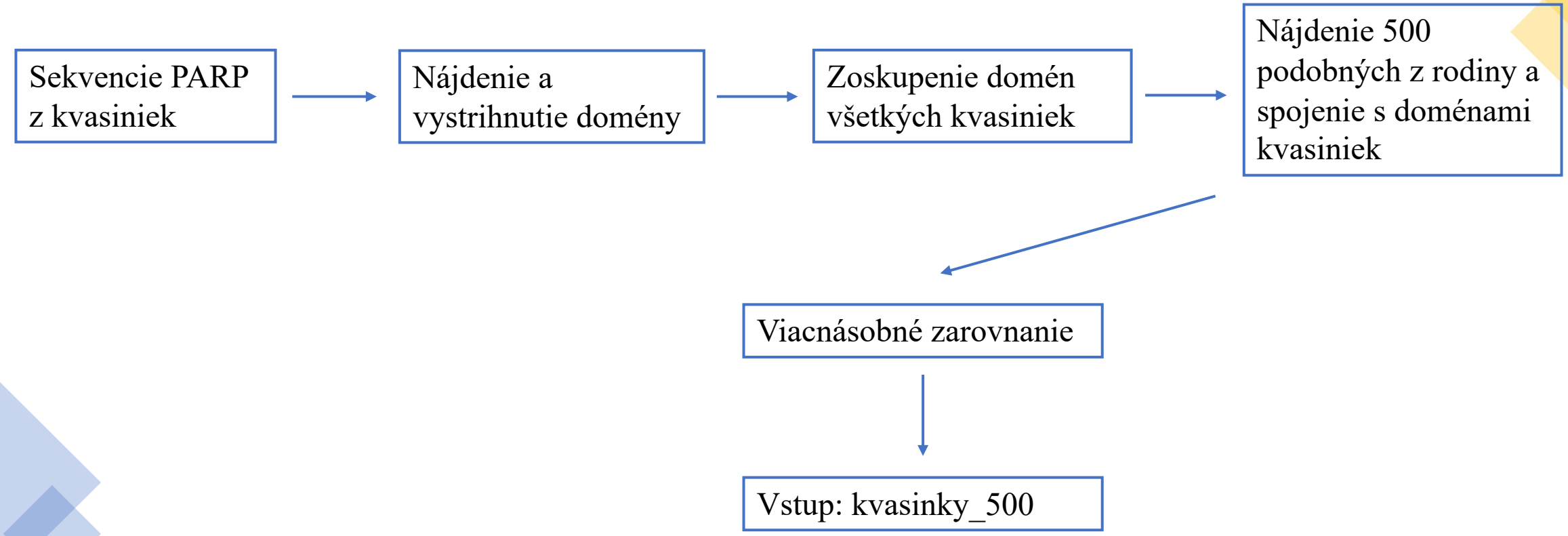
| i | | | | | j | | k | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | B | C | D | - | F | G | H | N | J | K | L |
| A | C | D | E | F | F | H | N | J | - | L | M |
| M | G | H | N | - | Y | L | N | - | - | L | M |
| A | G | H | N | - | F | L | N | F | L | Y | M |
| M | S | H | B | - | Y | G | N | S | G | W | T |
| M | S | A | N | - | Y | - | N | K | A | R | T |

# Metódy

- Štatistická analýza prepojení - **SCA** (z angl. *Statistical Coupling Analysis*)
  - Hľadanie nezávislých komponentov a proteínových sektorov

- **GREMLIN** (z angl. *Generative REgularized ModeLs of proteINs*)
  - Hľadanie fyzických kontaktov aminokyselín v terciárnej štruktúre proteínu

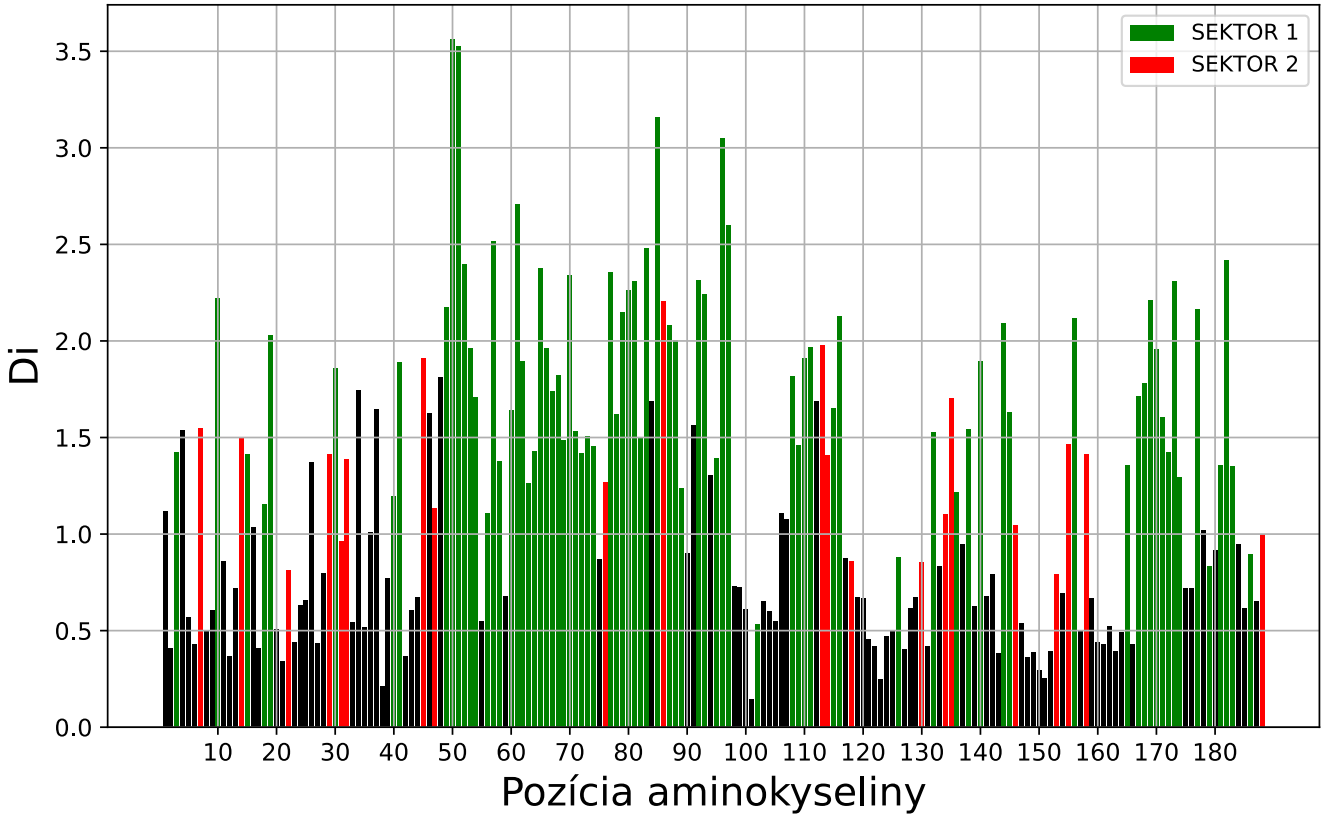
# Príprava vstupných zarovnaní



# SCA analýza

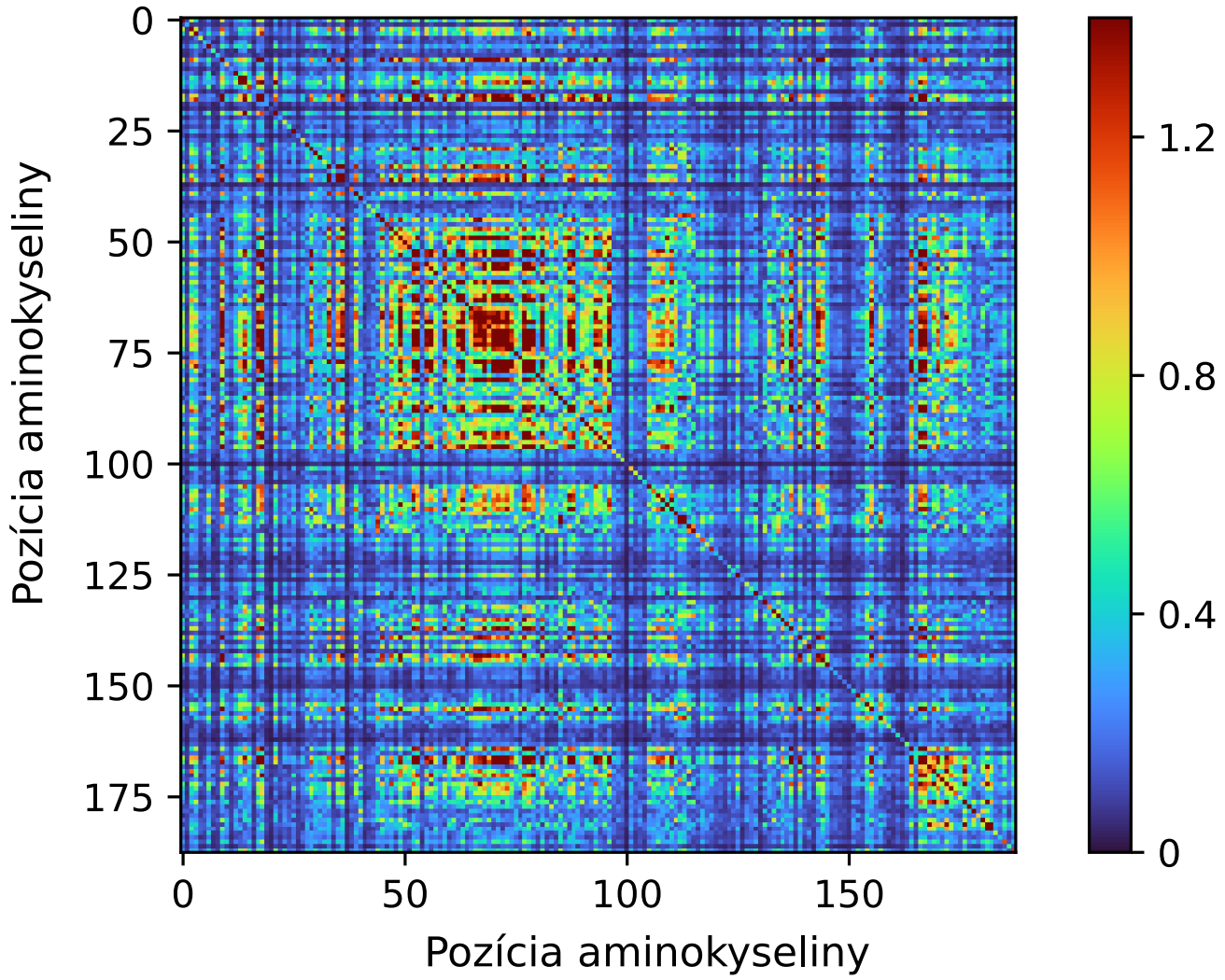

- Štúdium fungovania analýzy
- Vizualizácia celkových aj čiastkových výsledkov – *printResultsSK.py*
  - Vzťahy medzi konzervovanosťou pozície a príslušnosťou do komponentu
- Hľadanie pozícií významných pre kvasinky – porovnávanie vstupov
  - Porovnanie pozícií príslušných do nezávislých komponentov – *compareICs.py*

# SCA výsledky: Konzervovanosť

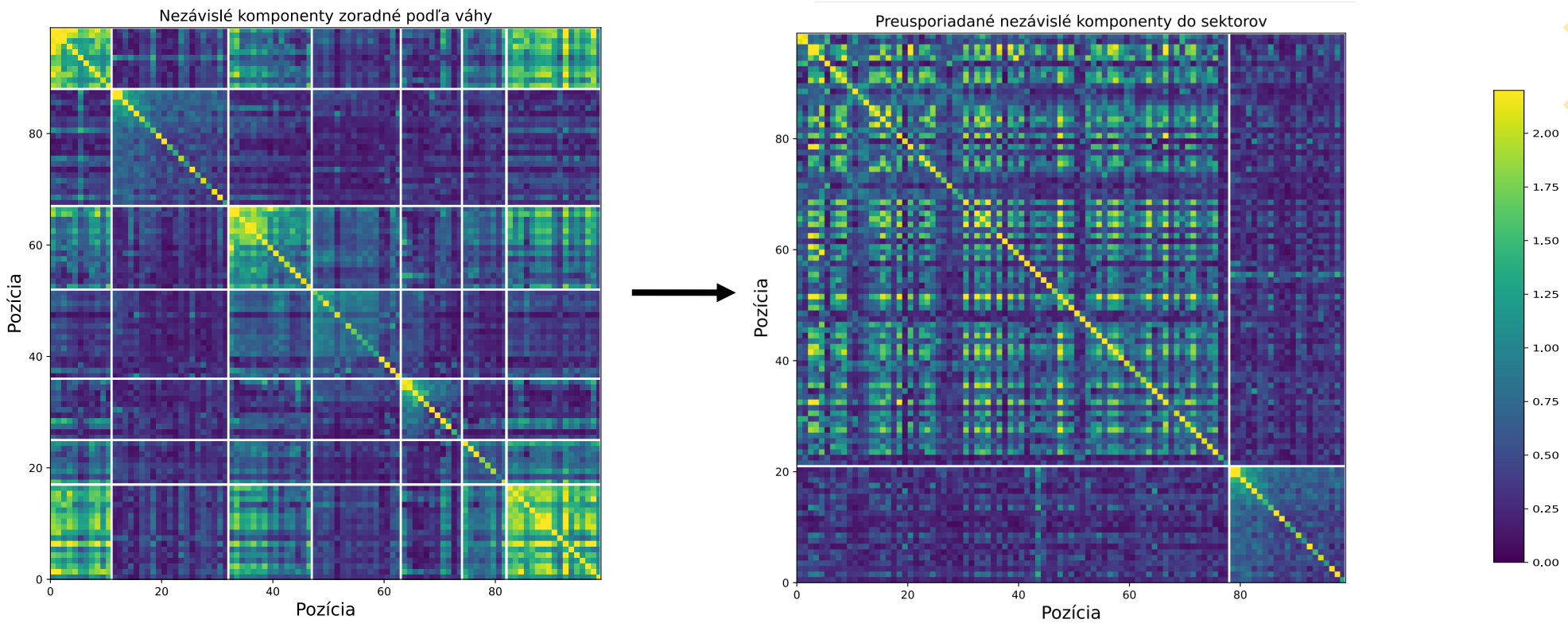



Konzervovanosť jednotlivých pozícií katalytickej domény proteínu PARP

# SCA výsledky: Koevolúcia – korelačná matica

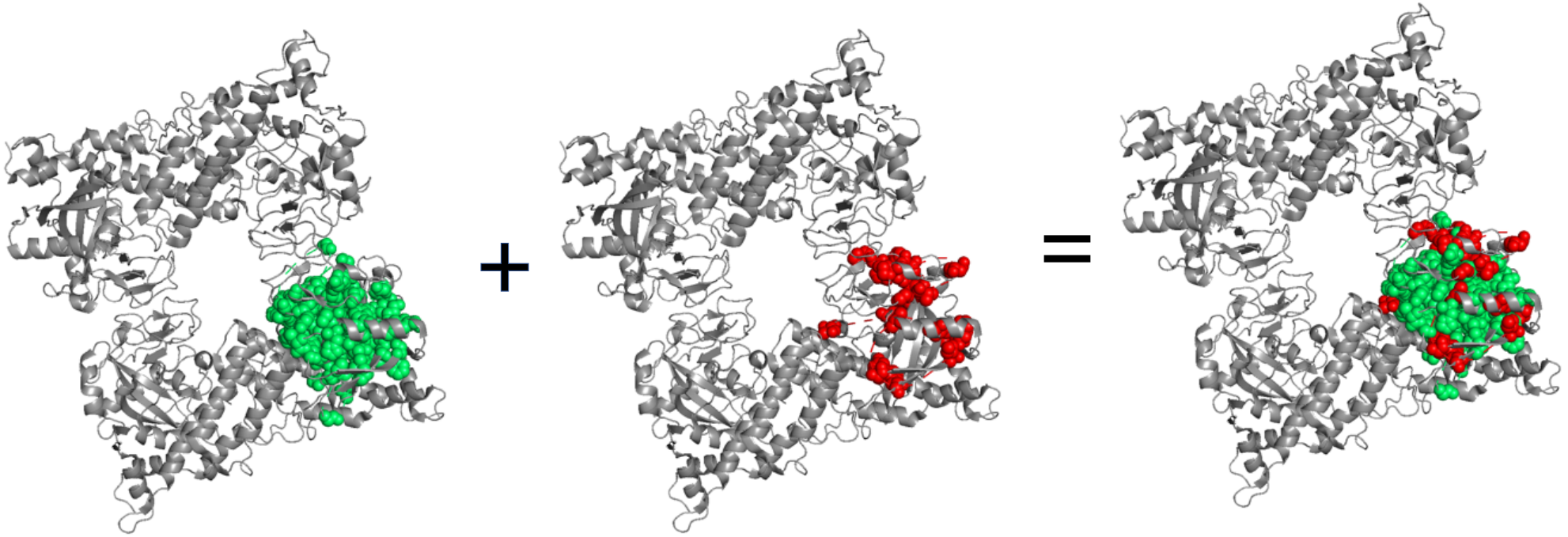


# SCA výsledky: Nezávislé komponenty → Sektory



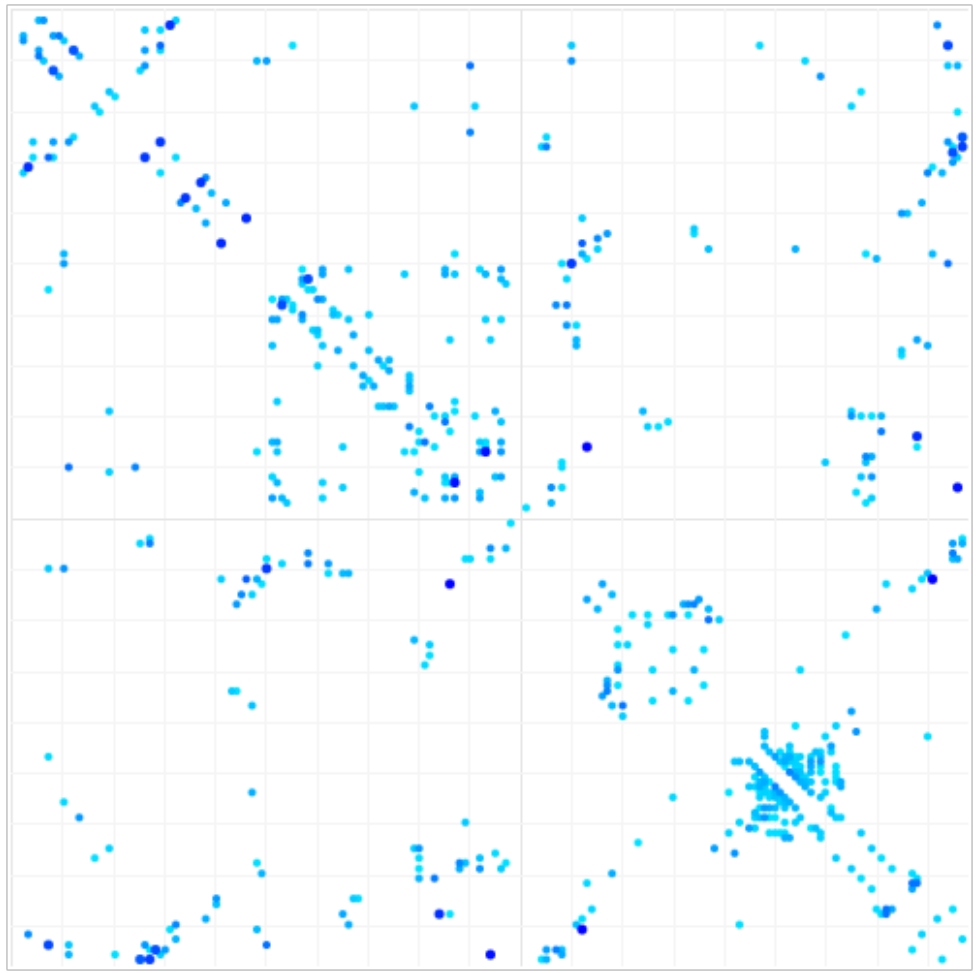

Matice zoskupovania nezávislých komponentov

# SCA výsledky: 3D vizualizácia proteínových sektorov

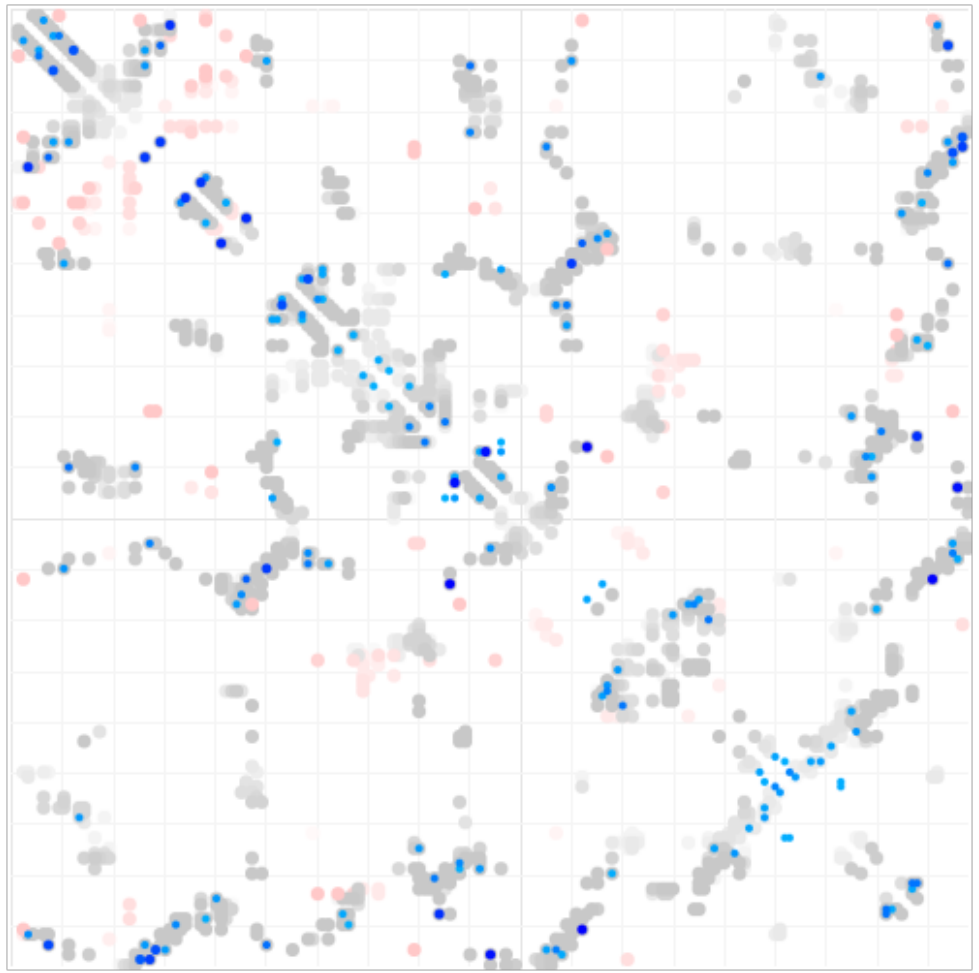


3D mapovanie pozícií na terciárnu štruktúru katalytickej PARP domény – 1WOK

# GREMLIN analýza



Obr. 1: Koevolúcia vybraných dvojíc pozícií



Obr. 2: Prekryv predikovaných kontaktov s informáciami z existujúcich štruktúr

# Porovnanie: SCA vs. GREMLIN

**SCA**

- ✓ Nájdenie proteínových sektorov
- ✓ Algoritmická predikcia nezávislých komponentov – subštruktúr proteínu

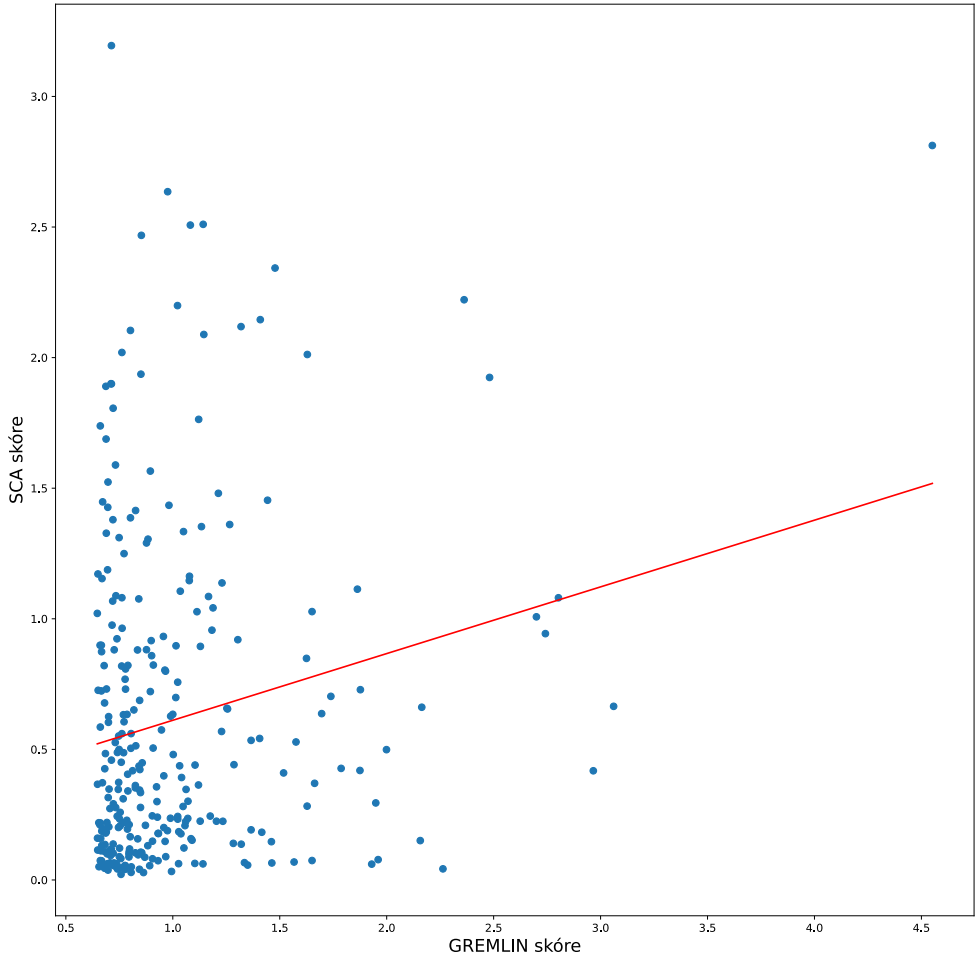
- ✓ Štatistické nástroje a práca s maticami
- ✓ Konzervovanosť a koevolúcia
- x Vyhýbanie sa nepriamym koreláciám
- ✓ Vzdialené korelujúce pozície

**GREMLIN**

- ✓ Predikcia fyzických kontaktov
- ✓ Porovnanie výsledkov s existujúcimi štruktúrami
- ✓ Zlepšovanie komparatívnych modelov
- ✓ Internetové rozhranie

- ✓ Využitie učiacich sa algoritmov a pravdepodobnostných modelov
- ✓ Konzervovanosť a koevolúcia
- ✓ Vyhýbanie sa nepriamym koreláciám
- x Vzdialené korelujúce pozície

# Porovnanie: SCA vs. GREMLIN




- Hľadanie GREMLIN korelácií v sektoroch nájdených v SCA analýze
- Obr: Porovnanie hodnôt korelácií vybraných dvojíc pozícií
  - *SCA_GREM_corrCompare.py*

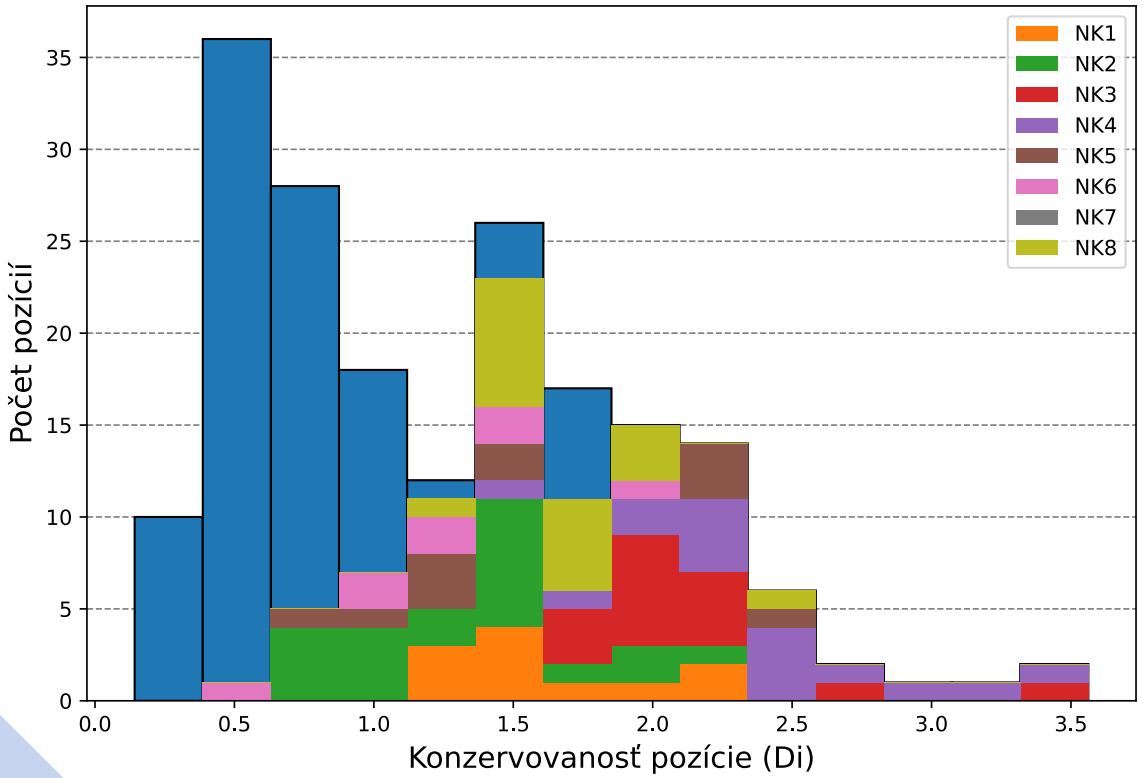
# Prínos práce a vízie do budúcnosti

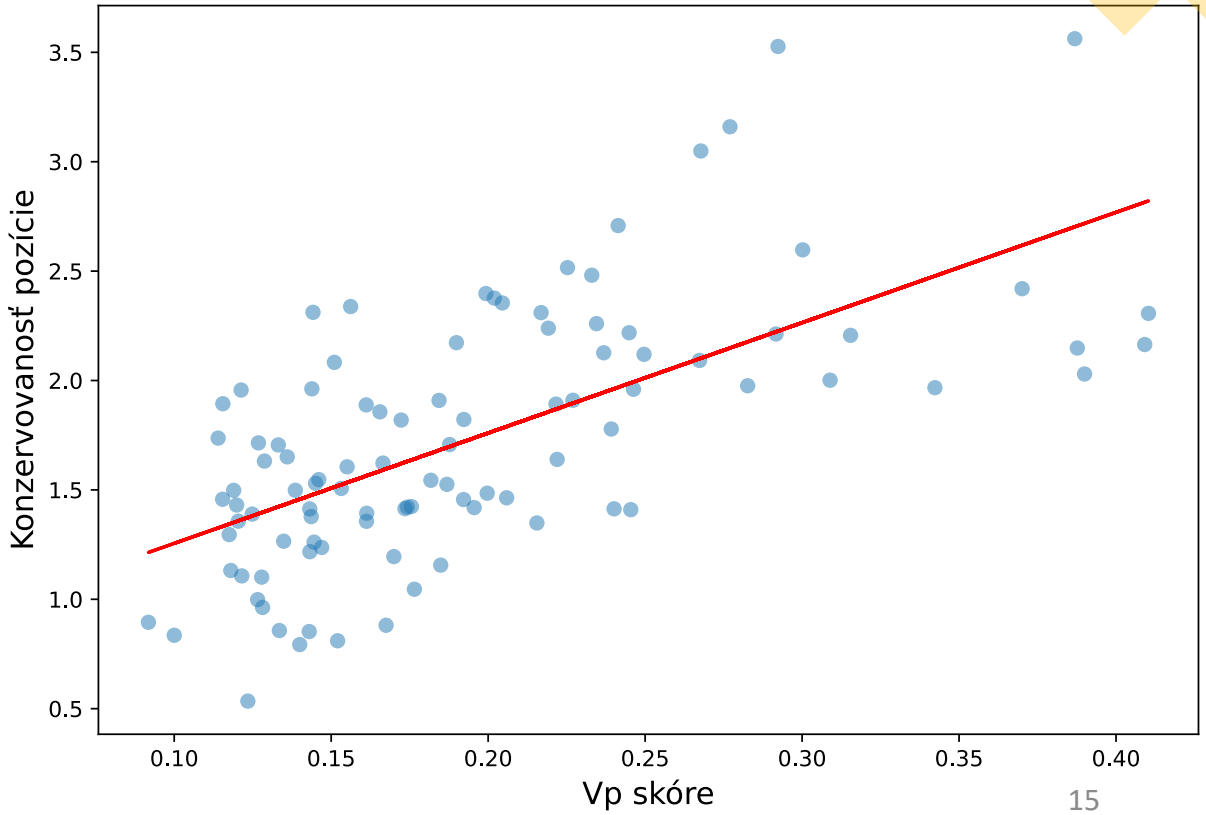
- Prínos práce
  - Neintuitívne pozície proteínu potenciálne zaujímavé pre experimentálne pozorovanie – výrazné zúženie biologického problému
  - Hodnoty korelácií nápomocné pre hľadanie spriahnutých dvojíc pozícií
  - Zjednotená vizualizácia medzivýsledkov
  - Porovnávanie jednotlivých výstupov v rámci SCA, porovnanie s výsledkami GREMLIN analýzy

- Vízie do budúcnosti
  - Návrh biologického experimentu, získanie informácií o biochemických vlastnostiach sektorov
  - Automatické spúšťanie programu a získanie štatisticky významných pozícií práve pre kvasinky

Ďakujem za pozornosť.

# Dodatok 1: SCA výsledky: Konzervovanosť vs. komponenty



# Dodatok 2: SCA výsledky: Porovnanie

| 1.vstup | NK/P | 2.vstup | NK/P | Zhody | Jaccard Index |
|---|---|---|---|---|---|
| kvasinky_500 | 7/99 | podobne_kvasinky | 9/112 | 96 | 83,48 |
| kvasinky_500 | 7/99 | nahodny_vyber | 5/91 | 80 | 72,73 |
| yarrowia_500 | 8/104 | podobne_yarrowia | 8/106 | 99 | 89,19 |
| yarrowia_500 | 8/104 | nahodny_vyber | 5/91 | 87 | 80,56 |
| kvasinky_500 | 7/99 | yarrowia_500 | 8/104 | 91 | 81,25 |

# Odpovede na oponentské otázky

- Aký vplyv by malo na vašu analýzu, keby ste ku kvasinkovým proteínom pridali výrazne menej alebo výrazne viac nekvasinkových proteínov?
- Je 500 správny počet?

→ príliš málo: analýza padne

→ málo efektívnych sekvencií, nepresné výsledky

→ príliš vela: strata významnosti kvasinkových sekvencií

# Odpovede na oponentské otázky

- Podľa vašich analýz, filtrovanie stĺpcov a riadkov môže mať veľký vplyv na výsledky analýz (čo sa prejavuje napríklad tým, že v rôznych data setoch vám “vypadávajú” rôzne stĺpce). Analýza tak dosť môže závisieť od toho, ktorých 500 nekvasinkových proteínov sa vyberie ku jadru kvasinkových proteínov. Vedeli by ste navrhnúť nejaký postup, ktorý by urobil analýzu robustnejšiu / menej závislú od výberu konkrétnych nekvasinkových proteínov?

→ automatizované spúšťanie  s rôznymi podmnožinami podrodiny – štatistické vyhodnotenie najviac opakujúcich sa ”nových pozícií”

→ pridanie väčších váh (napr. zdvojnásobenie) kvasinkovým sekvenciám