Hledání léku v kupce molekul, aneb počítačový vývoj léčiv

David Hoksza
http://bioinformatics.cuni.cz/hoksza
https://ksi.mff.cuni.cz

Definice hřiště

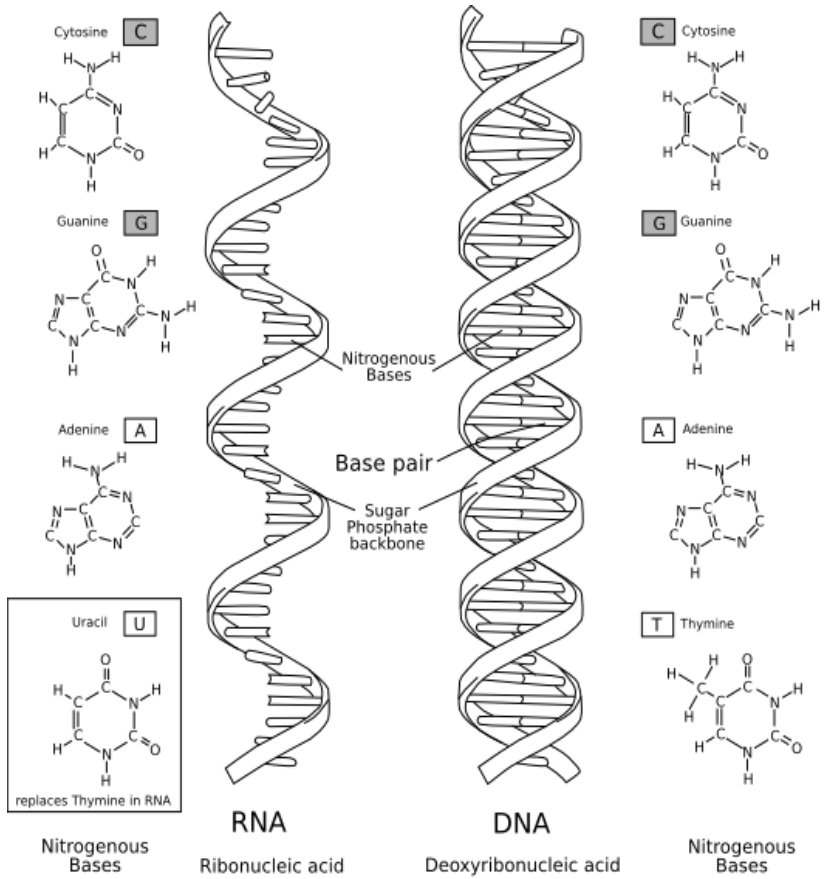
Molekulární hráči

SYSTÉMOVÁ BIOLOGIE

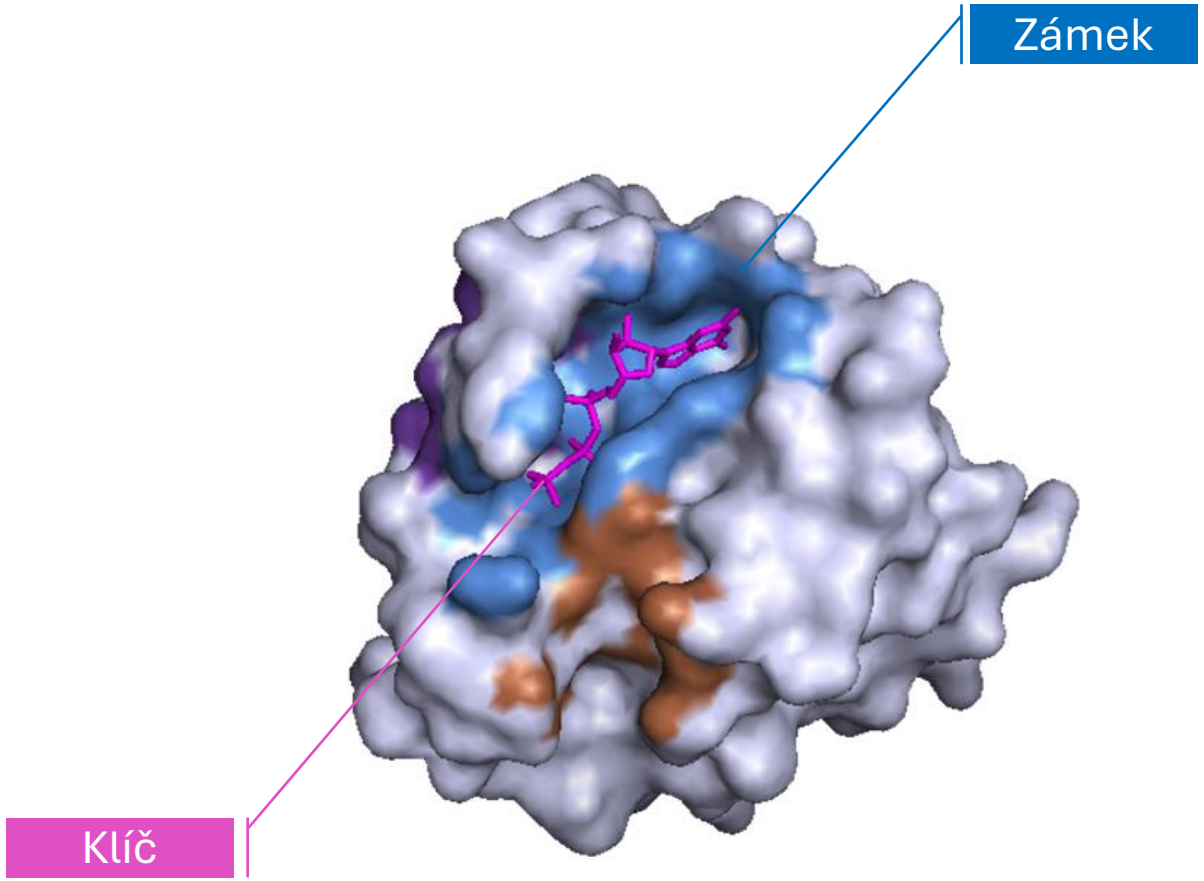
BIO INFORMTIKA

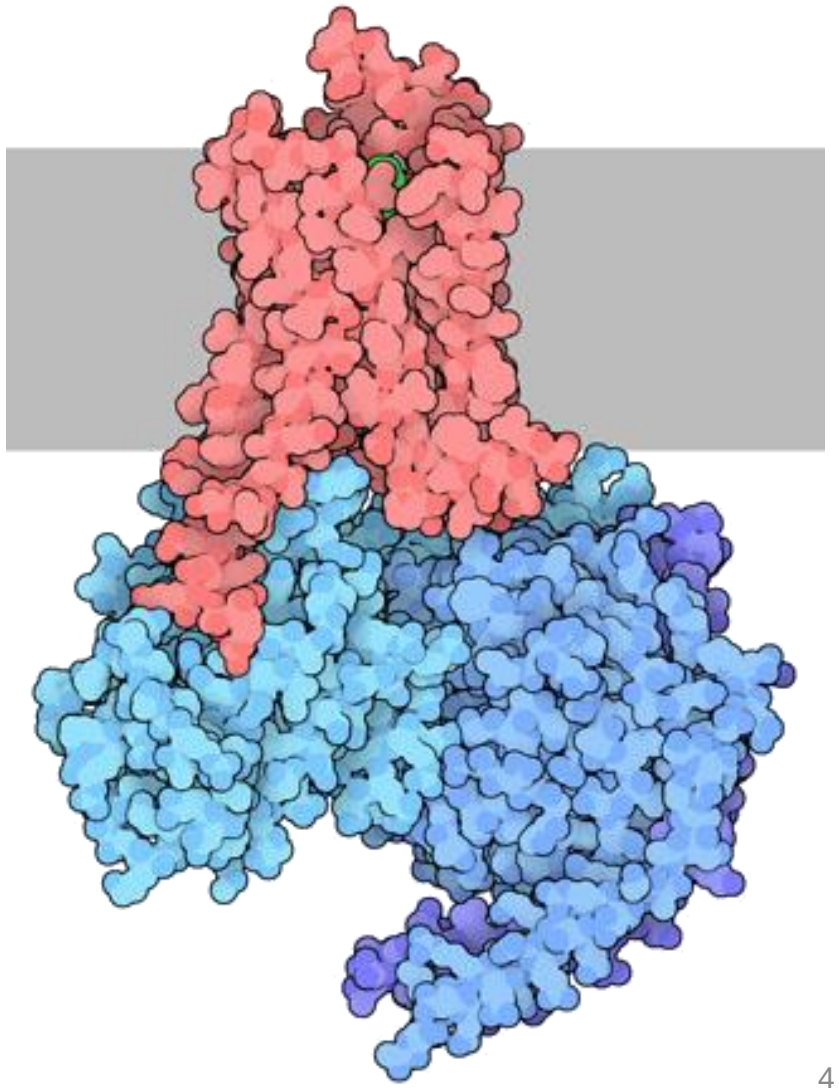
CHEMO INFORMATIKA

- **DNA (deoxyribonukleová kyselina)**
  - Posloupnost nukleotidů (A, C, G, T)
  - Dvoušroubovice
- **RNA (ribonukleová kyselina)**
  - Posloupnost nukleotidů (A, C, G, U)
  - Jednošroubovice
- **Proteiny**
  - Překlad z mRNA v ribozomech
  - Sekvence aminokyselin (20 AAs)
- **Metabolity**
  - Malé molekuly
  - Reaktanty a produkty biochemických reakcí zajišťujících růst, vývoj, přenos informace…



Co jsou léčiva

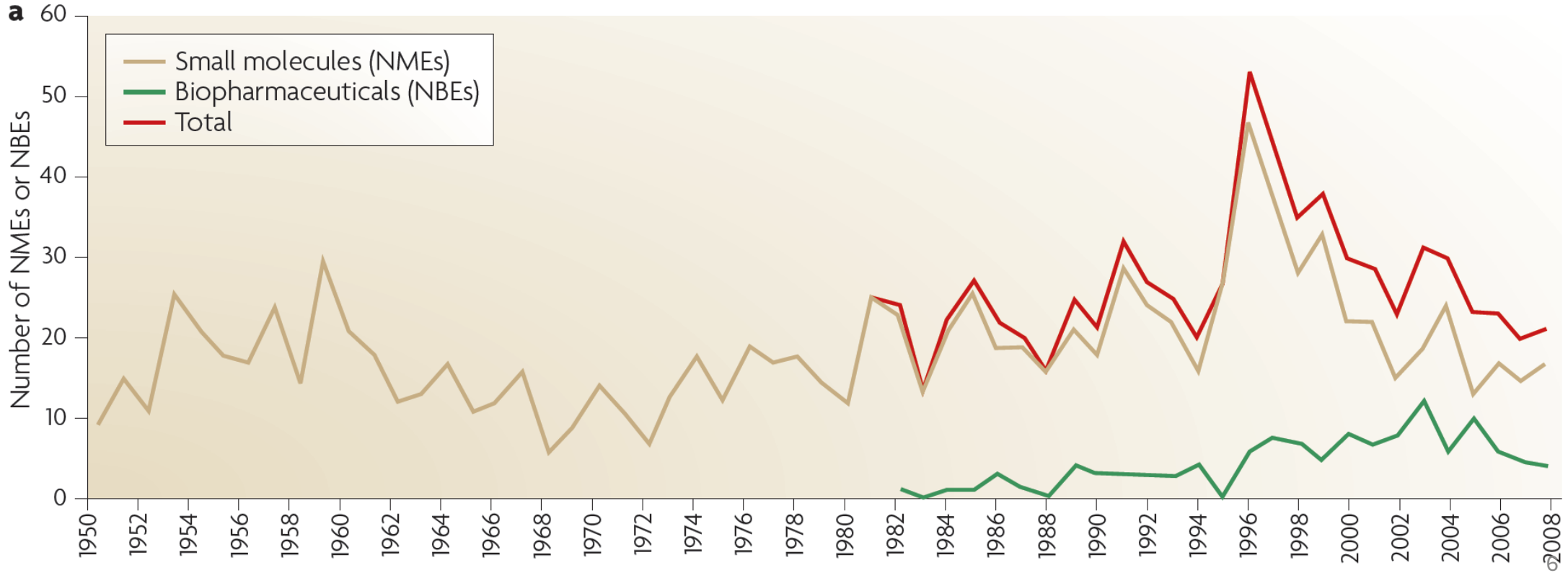




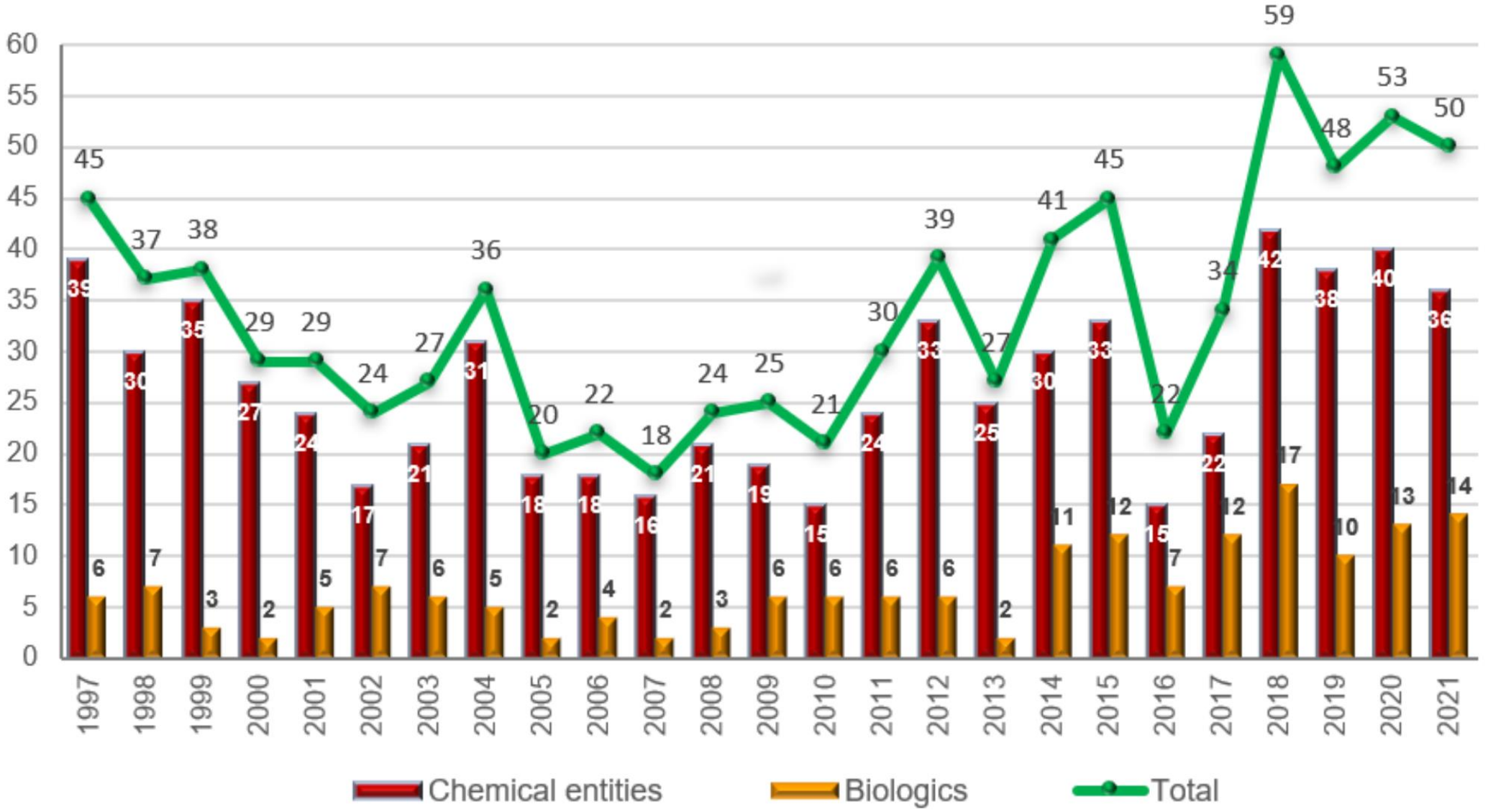
Proces a cena vývoje léčiv

Počet nových léků

- Od roku 1950 do 2008 schválila FDA 1222 léků (1103 nových malých molekulárních entit, 119 biologics)

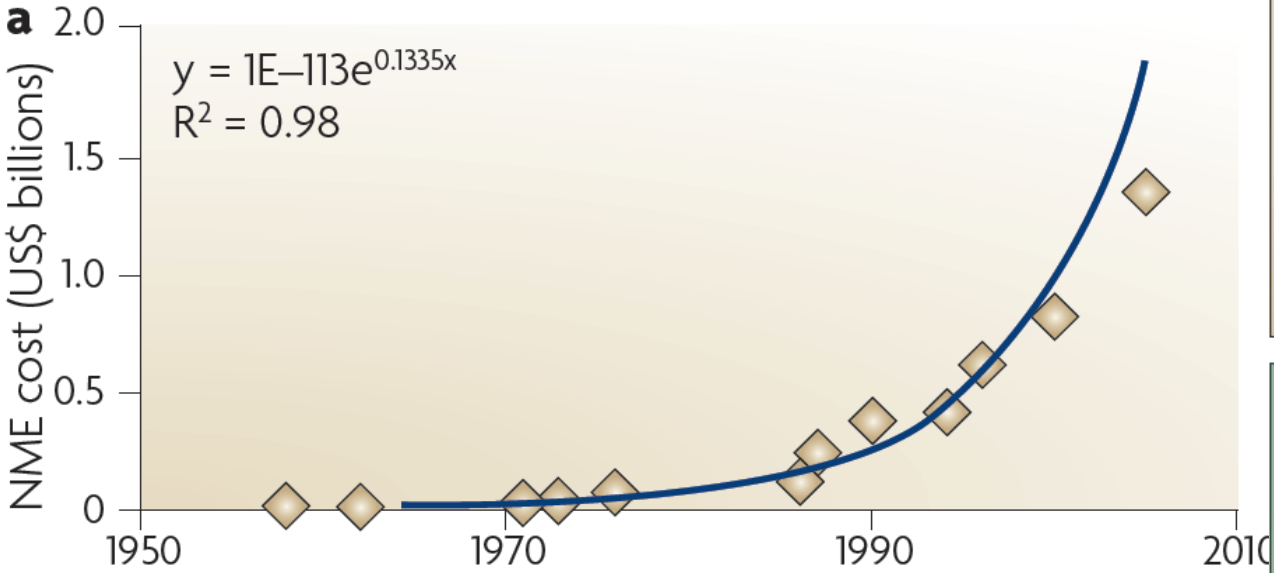


source: Munos, B. (2009). Lessons from 60 years of pharmaceutical innovation. *Nature reviews Drug discovery*, *8*(12), 959-968.



source: de la Torre, B. G., & Albericio, F. (2022). The pharmaceutical industry in 2021. An analysis of FDA drug approvals from the perspective of molecules. *Molecules*, *27*(3), 1075.

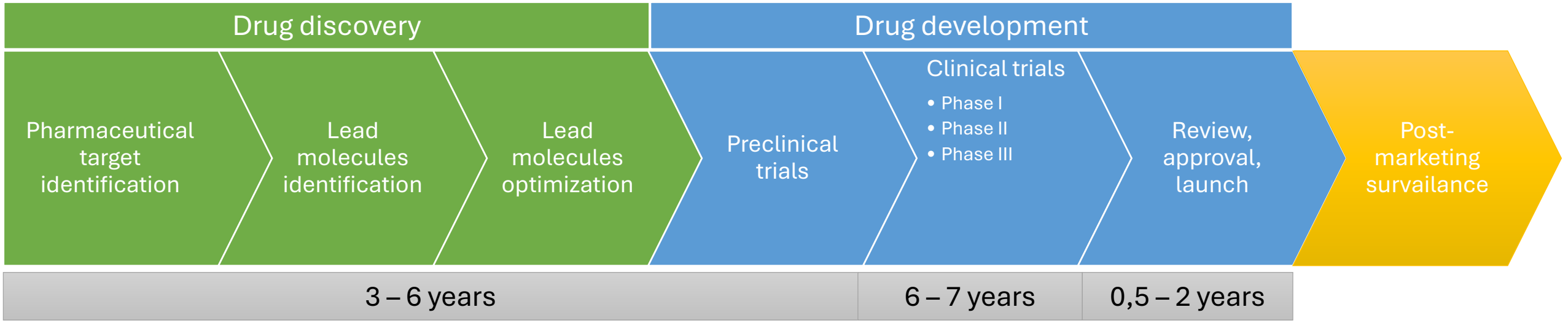
Cena nového léku



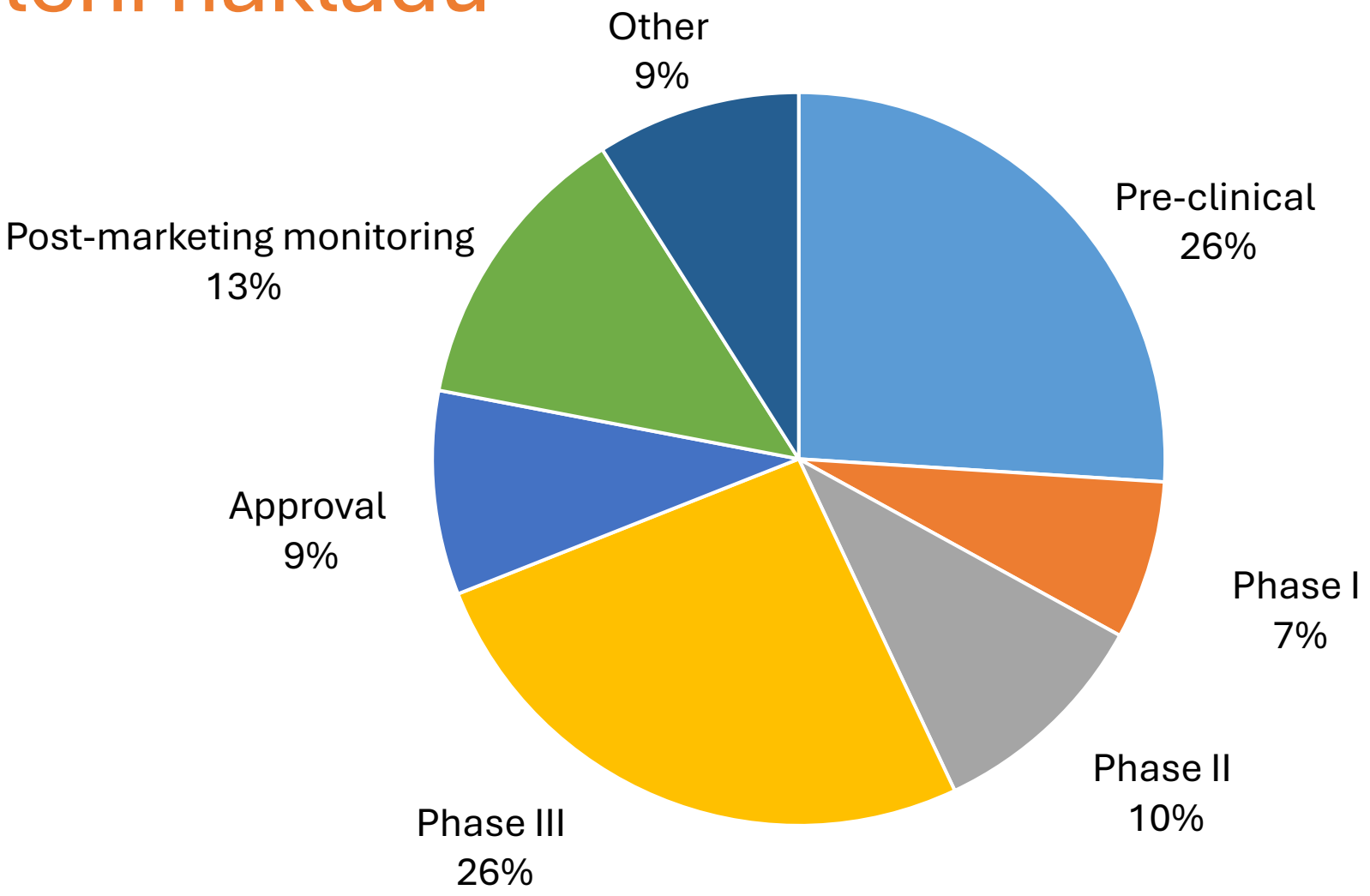
c Estimate	Value of item (US$ millions)
DiMasi estimate in 2000 dollars	802
• Adjustment for post-approval R&D	95
• Adjustment for new indications and non-US approvals (20%)	160
• Adjustment for success rate (11.5% versus 21.5%)	697
Adjusted DiMasi estimate in 2000 dollars	1,754
• Adjustment for inflation (3.7% per year)	592
• Adjustment for other cost increases, such as regulation (8.3% per year)	1,565
Adjusted DiMasi estimate in 2008 dollars	3,911

source: Munos, B. (2009). Lessons from 60 years of pharmaceutical innovation. *Nature reviews Drug discovery*, *8*(12), 959-968.

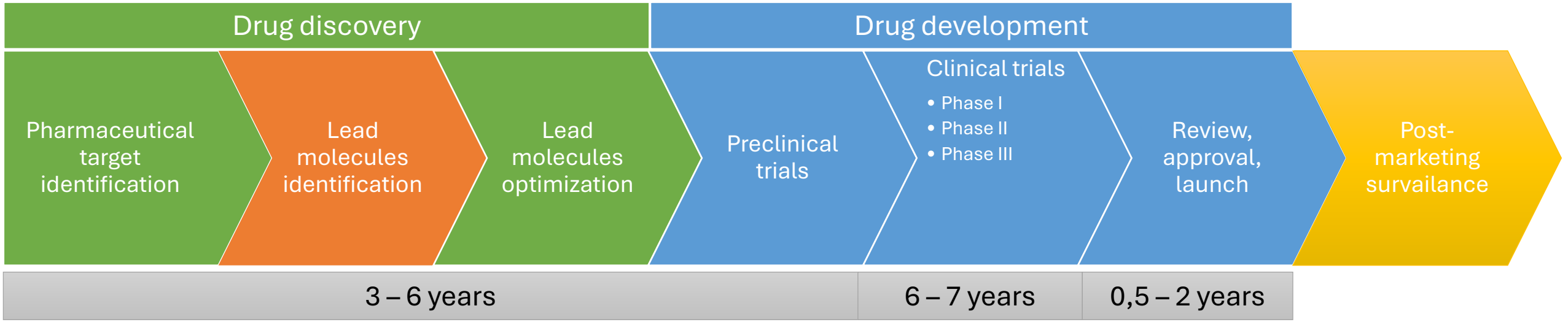
Proces vývoje léku



Dělení nákladů



Hlavní využití informatiky



Jak nalézt aktivní sloučeniny

Přístupy k identifikaci léčiv

Racionální (target-based) design

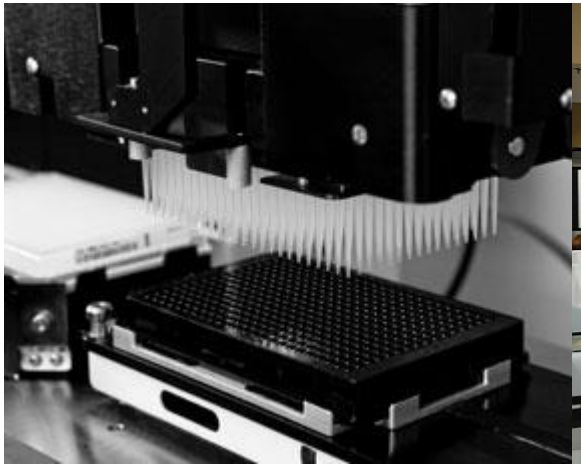
- Pochopení nemoci/systému
- Identifikace biologický procesů (biological pathways)
- Identifikace klíčového proteinu (cíl léčiva)
- Navrhnutí biochemického testu (in vitro test) a knihovy látek
- Optimalizace

Fenotypický přístup

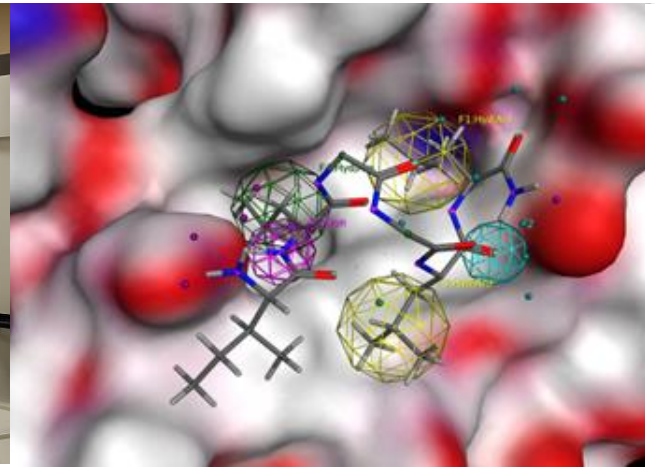
- Pozorování
  - rostlinu, látka, proces, … mající podobné efekty
- Isolace aktivní sloučeniny
- Reprodukce efektu v cílovém organismu (in vivo test)
- Optimalizace

Identifikace bioaktivních molekul

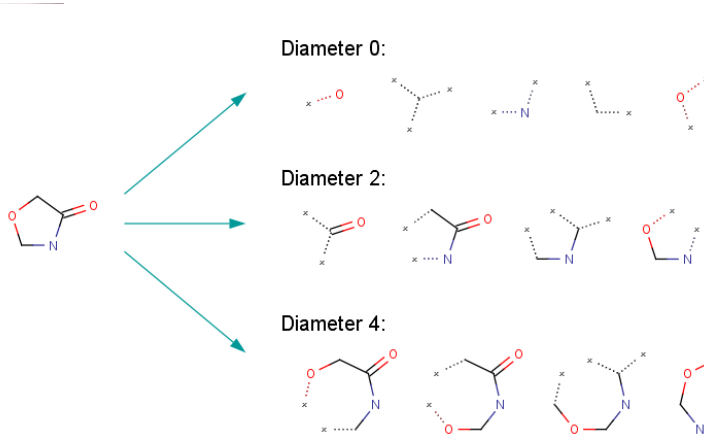
Vysokopropustný (high-throughput) **screening** (HTS)



source: Hybrigenics Services, Wikipedia



Virtuální HTS



Virtuální screening

Založené na struktuře cíle

Strukturní modelování

Založené na znalosti ligandů

Podobnostní hledání

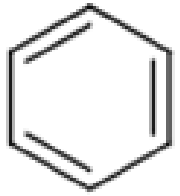
Farmakoforové modelování

BIOIFORMATIKA

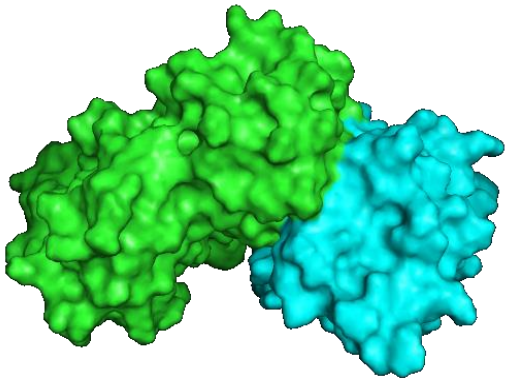
CHEMINFORMATIKA

Počítačová reprezentace sloučenin

Malé molekuly



Makromolekuly



```
benzene
ACD/Labs0812062058

  6  6  0  0  0  0  0  0  0  0  1 V2000
    1.9050   -0.7932    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.9050   -2.1232    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7531   -0.1282    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7531   -2.7882    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3987   -0.7932    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3987   -2.1232    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  2  1  1  0  0  0  0
  3  1  2  0  0  0  0
  4  2  2  0  0  0  0
  5  3  1  0  0  0  0
  6  4  1  0  0  0  0
  6  5  2  0  0  0  0
M  END
$$$$
```

```
          1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
ATOM      1  N   MET A   1      19.555  29.903  42.490  1.00 50.76           N
ATOM      2  CA  MET A   1      20.614  30.851  42.171  1.00 54.92           C
ATOM      3  C   MET A   1      21.995  30.221  42.321  1.00 56.63           C
ATOM      4  O   MET A   1      22.237  29.387  43.195  1.00 55.52           O
ATOM      5  CB  MET A   1      20.491  32.089  43.048  1.00 60.12           C
ATOM      6  N   LEU A   2      22.910  30.635  41.449  1.00 47.67           N
ATOM      7  CA  LEU A   2      24.304  30.230  41.561  1.00 39.94           C
ATOM      8  C   LEU A   2      25.108  31.273  42.334  1.00 37.57           C
ATOM      9  O   LEU A   2      25.342  32.368  41.821  1.00 58.96           O
ATOM     10  CB  LEU A   2      24.921  30.038  40.176  1.00 40.96           C
ATOM     11  CG  LEU A   2      24.451  28.823  39.376  1.00 40.06           C
ATOM     12  CD1 LEU A   2      24.794  28.991  37.903  1.00 38.65           C
ATOM     13  CD2 LEU A   2      25.054  27.548  39.941  1.00 31.43           C
ATOM     14  N   SER A   3      25.535  30.937  43.543  1.00 31.50           N
ATOM     15  CA  SER A   3      26.465  31.776  44.291  1.00 30.86           C
ATOM     16  C   SER A   3      27.633  32.236  43.430  1.00 37.53           C
ATOM     17  O   SER A   3      27.927  31.657  42.381  1.00 34.29           O
ATOM     18  CB  SER A   3      27.032  30.985  45.481  1.00 33.36           C
ATOM     19  OG  SER A   3      28.180  30.254  45.070  1.00 29.16           O
```

Virtuální screening

Založené na struktuře cíle

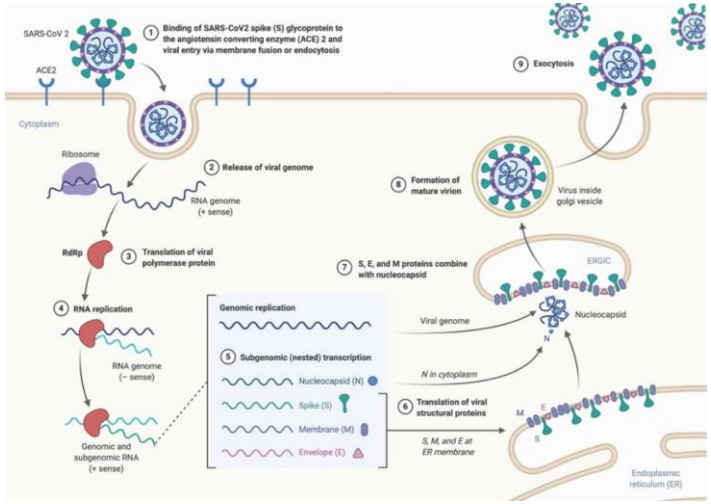
Založené na znalosti ligandů

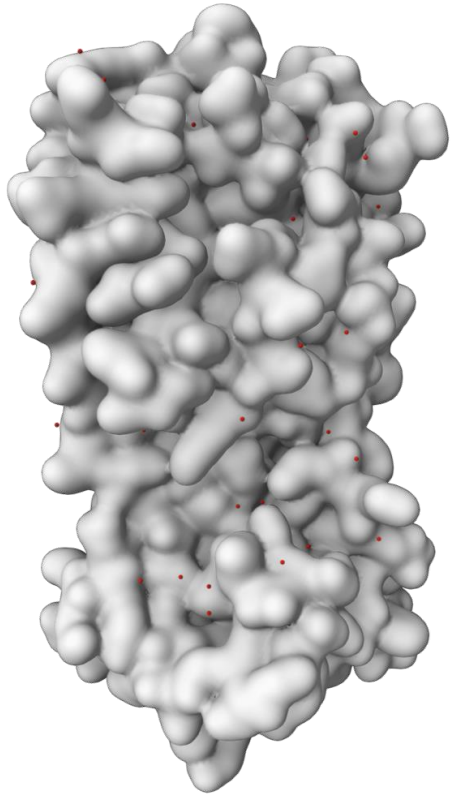
Strukturní modelování

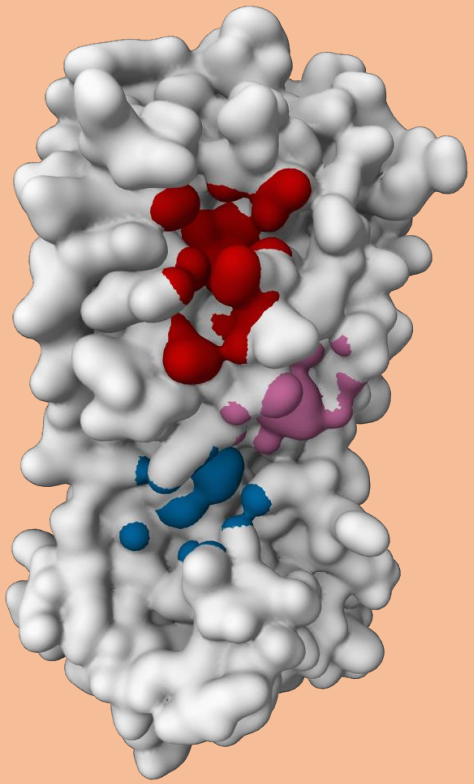
Podobnostní princip

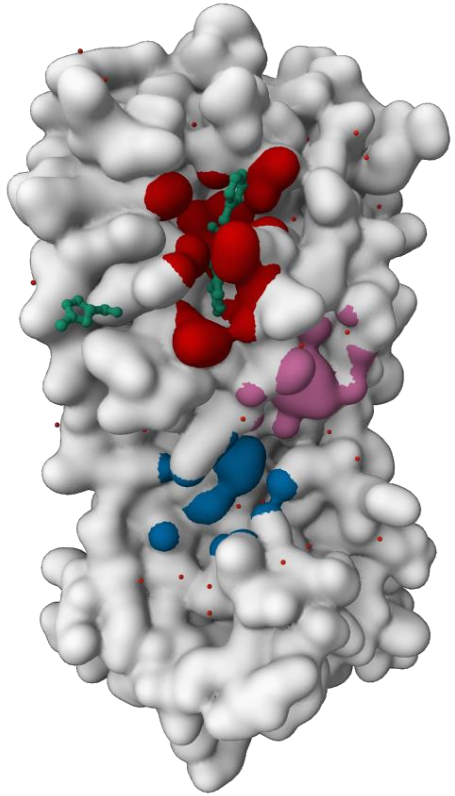
Farmakoforové modelování

Racionální design









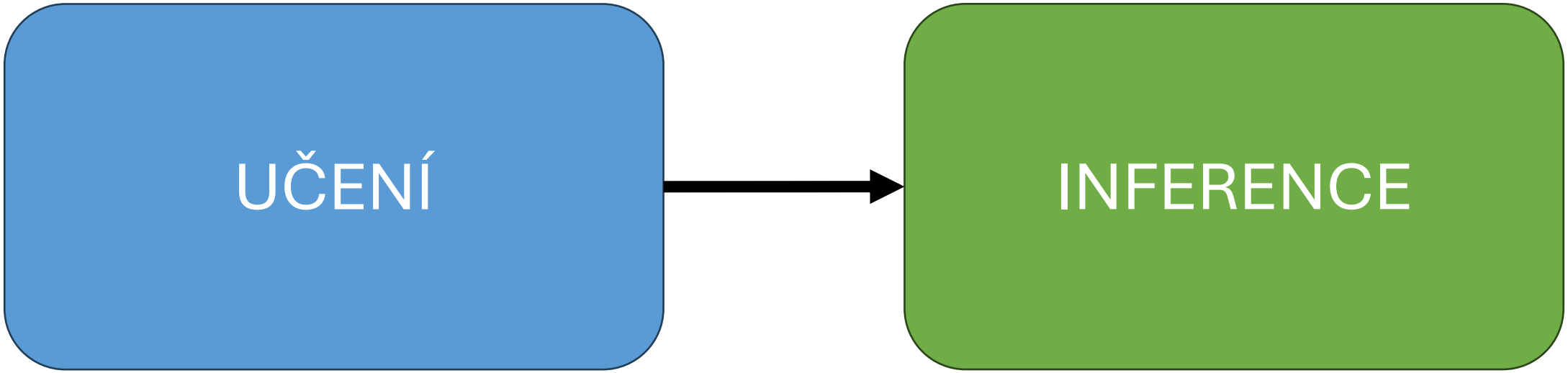
POCHOPENÍ BIOLOGICKÉHO PROCESU

VÝBĚR CÍLE

IDENTIFIKACE POTENCIÁLNÍCH VAZEBNÝCH MÍST

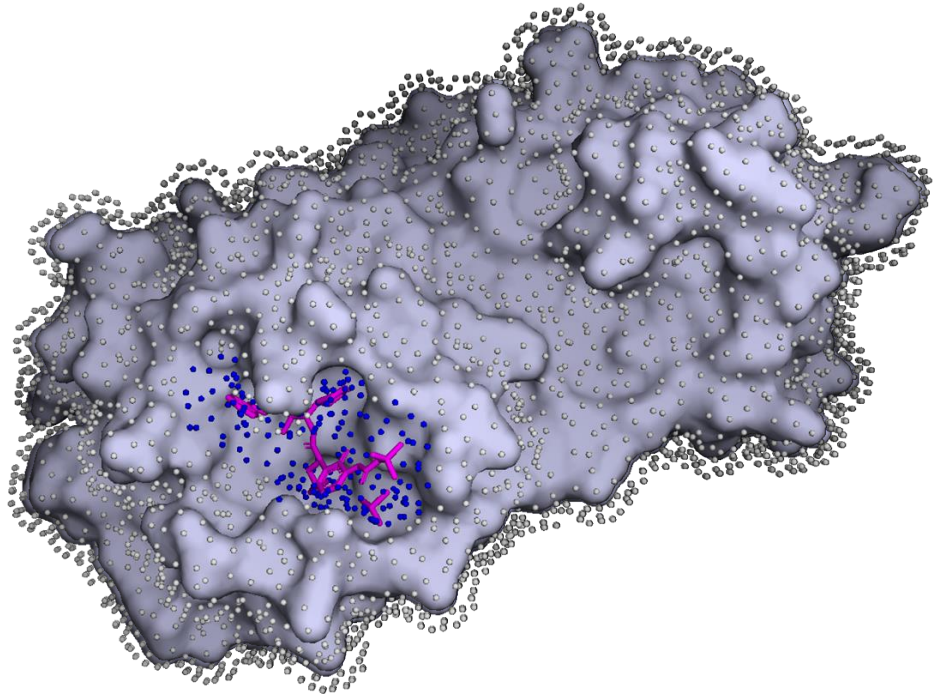
IDENTIFIKACE KANDIDÁTNÍCH MOLEKUL

Detekce vazebných míst pomocí strojového učení



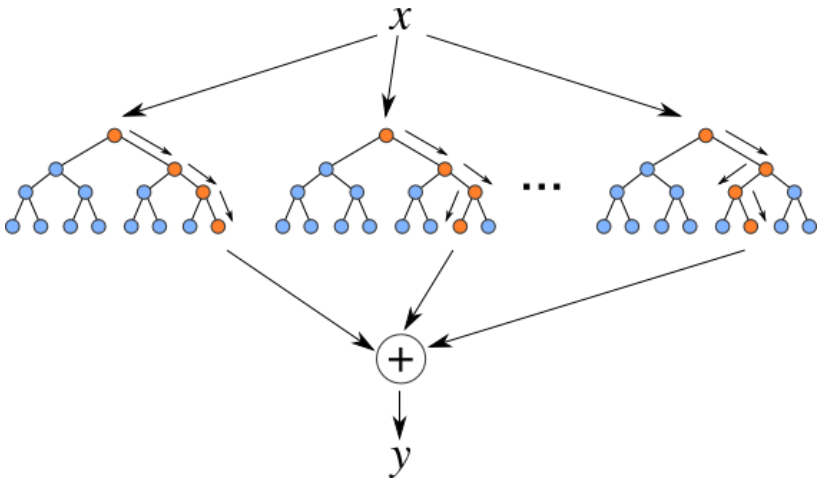
Učení (vlastností vazebných míst)

1. Získání struktur známých protein-ligand komplexů
2. Pokryj protein sítí bodů
3. **Označ body** jako vazebné/nevzebné
4. **Vyextrahuj fyzikálně-chemické vlastnosti** bodů
5. **Vybuduj/natrénuj model** pro rozhodnutí, zda je bod součástí vazebného místa



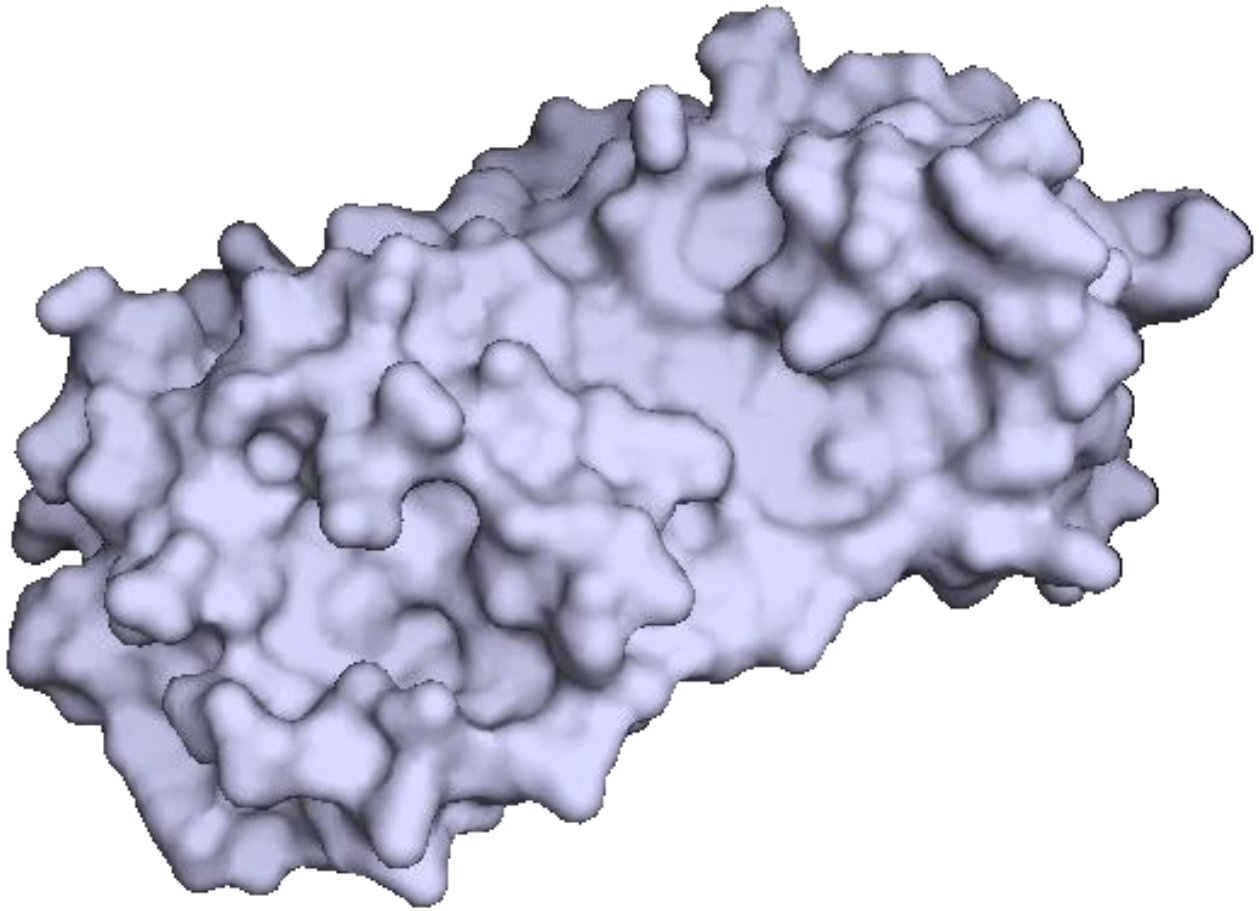
Random Forests classification

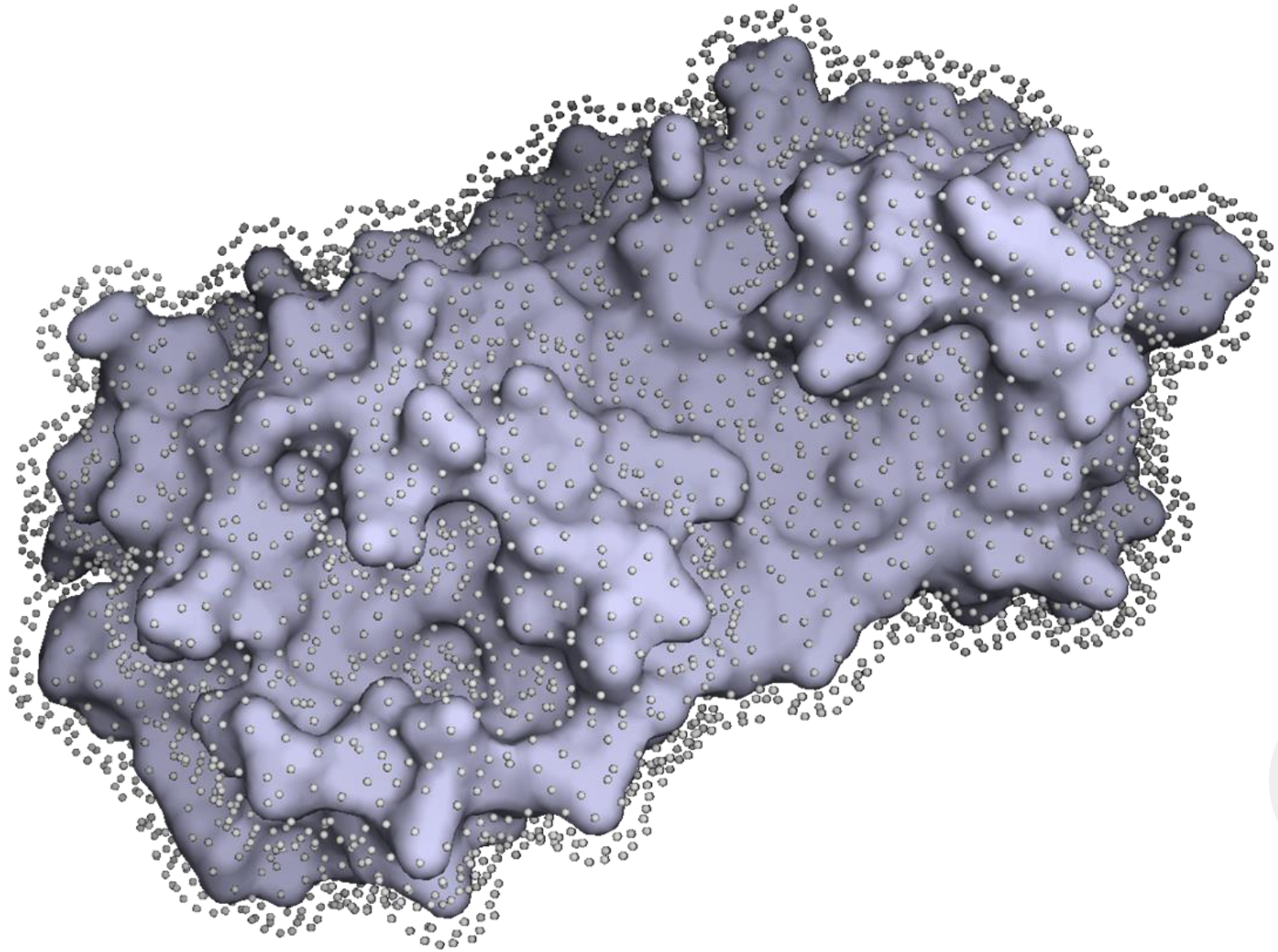
- **Ensemble** of decision trees
- **Single** decision **tree**
  - Unstable
  - Overfits

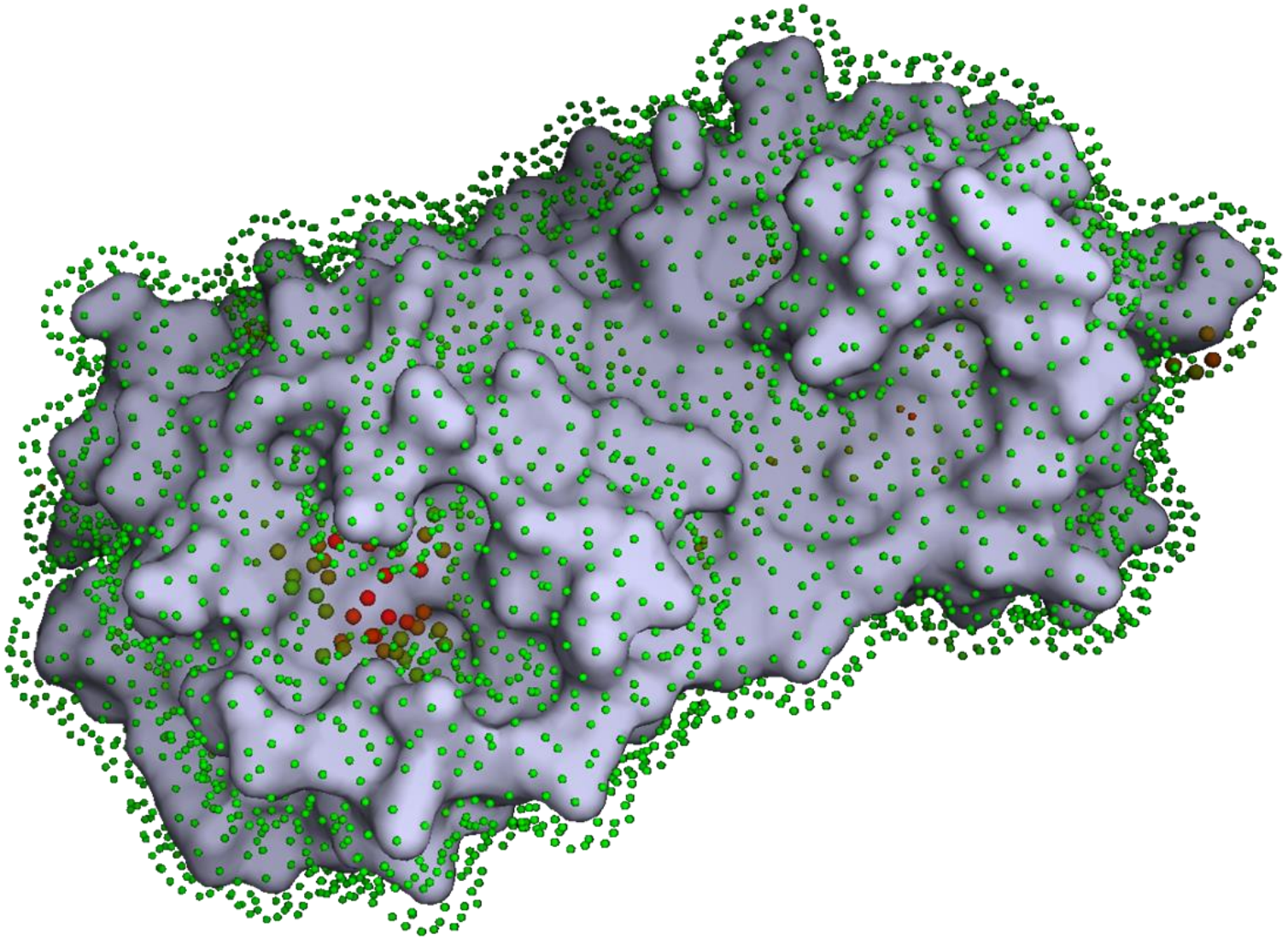


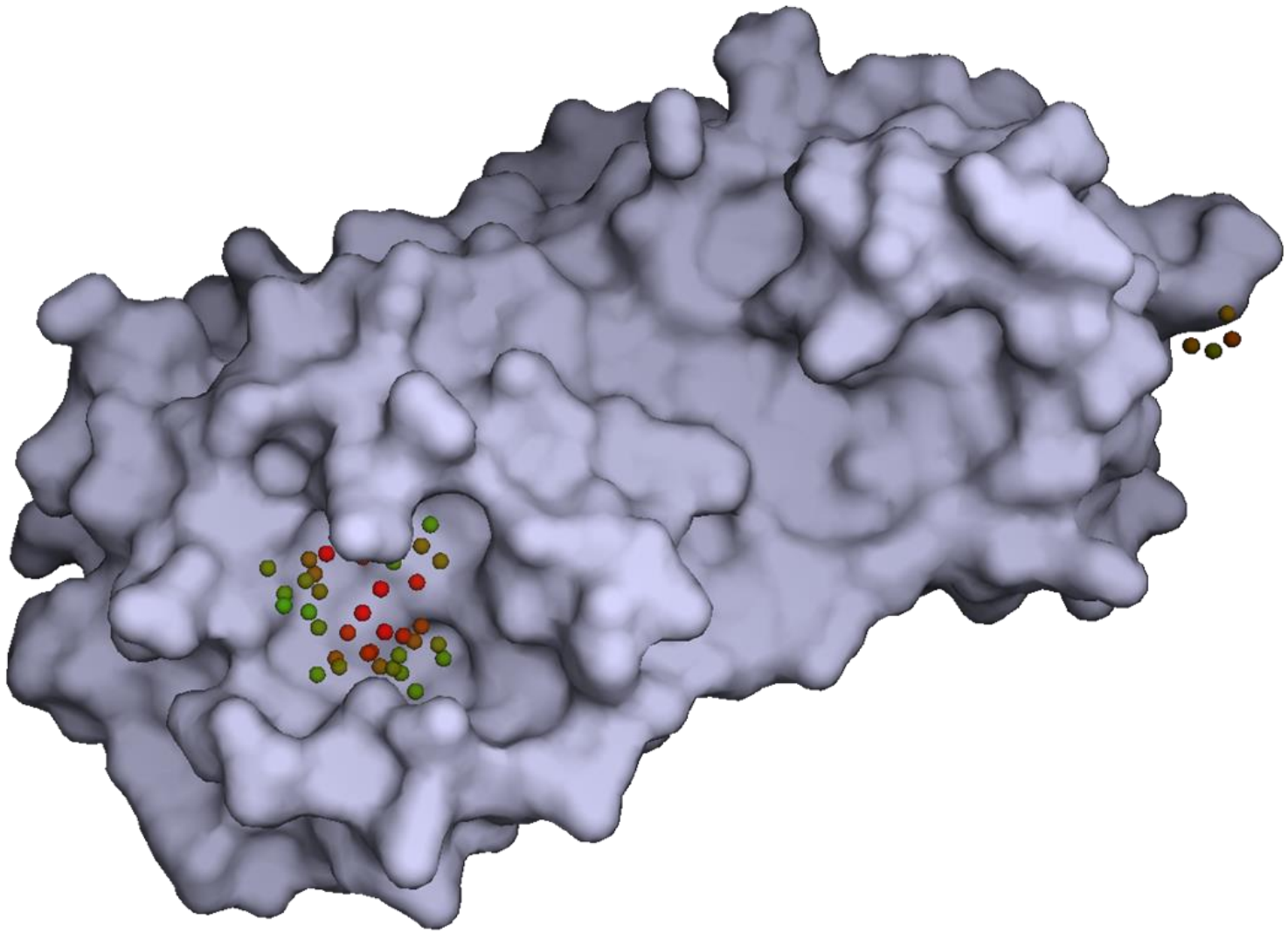
- **Solution**
  - Aggregate **multiple decision trees** on **bootstrapped data**
  - **Random choice of descriptors**
- Advantages
  - Suitable for **imbalanced data sets**
  - Estimates of what **variables** are **important** in the classification

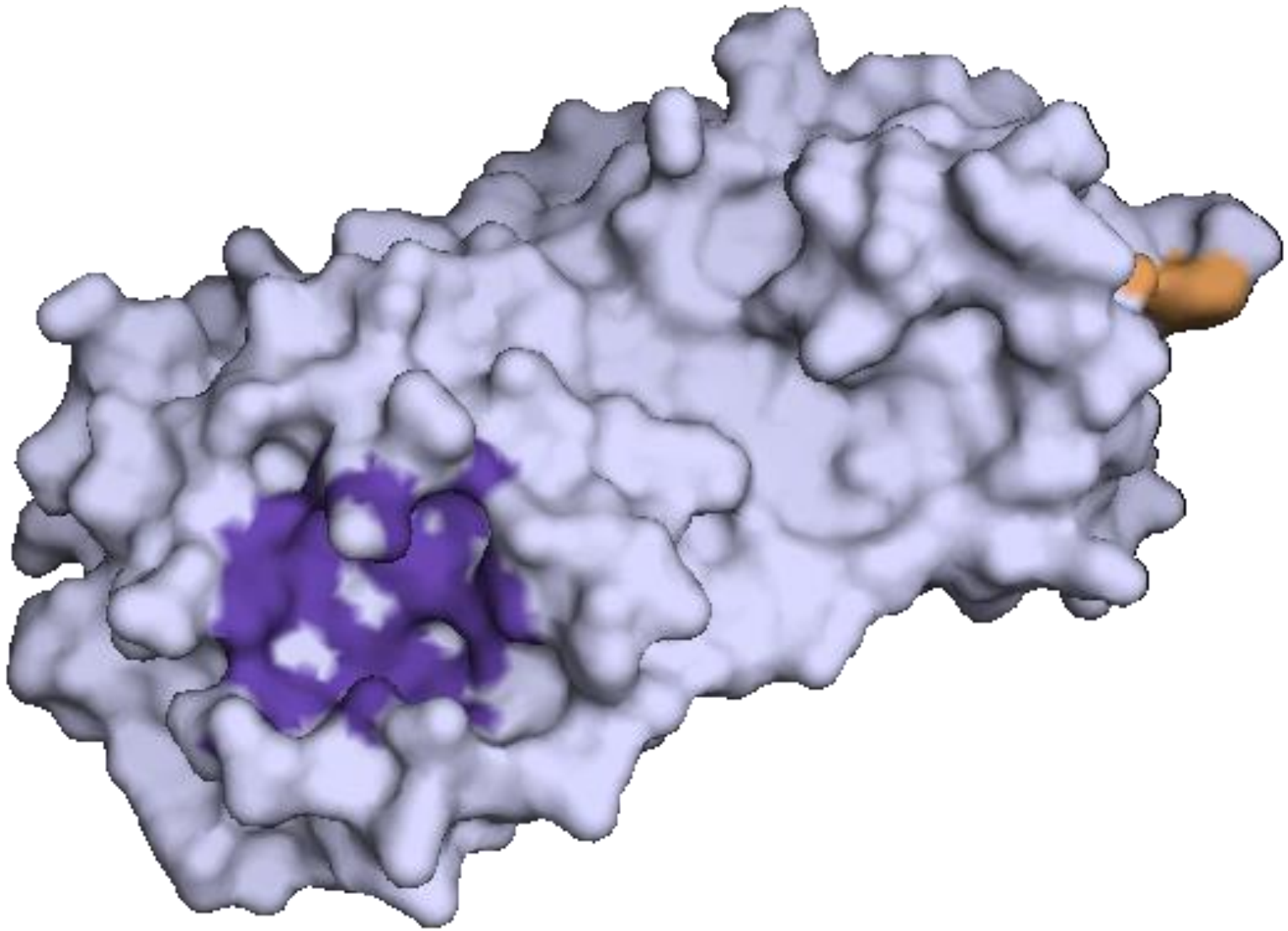
Inference



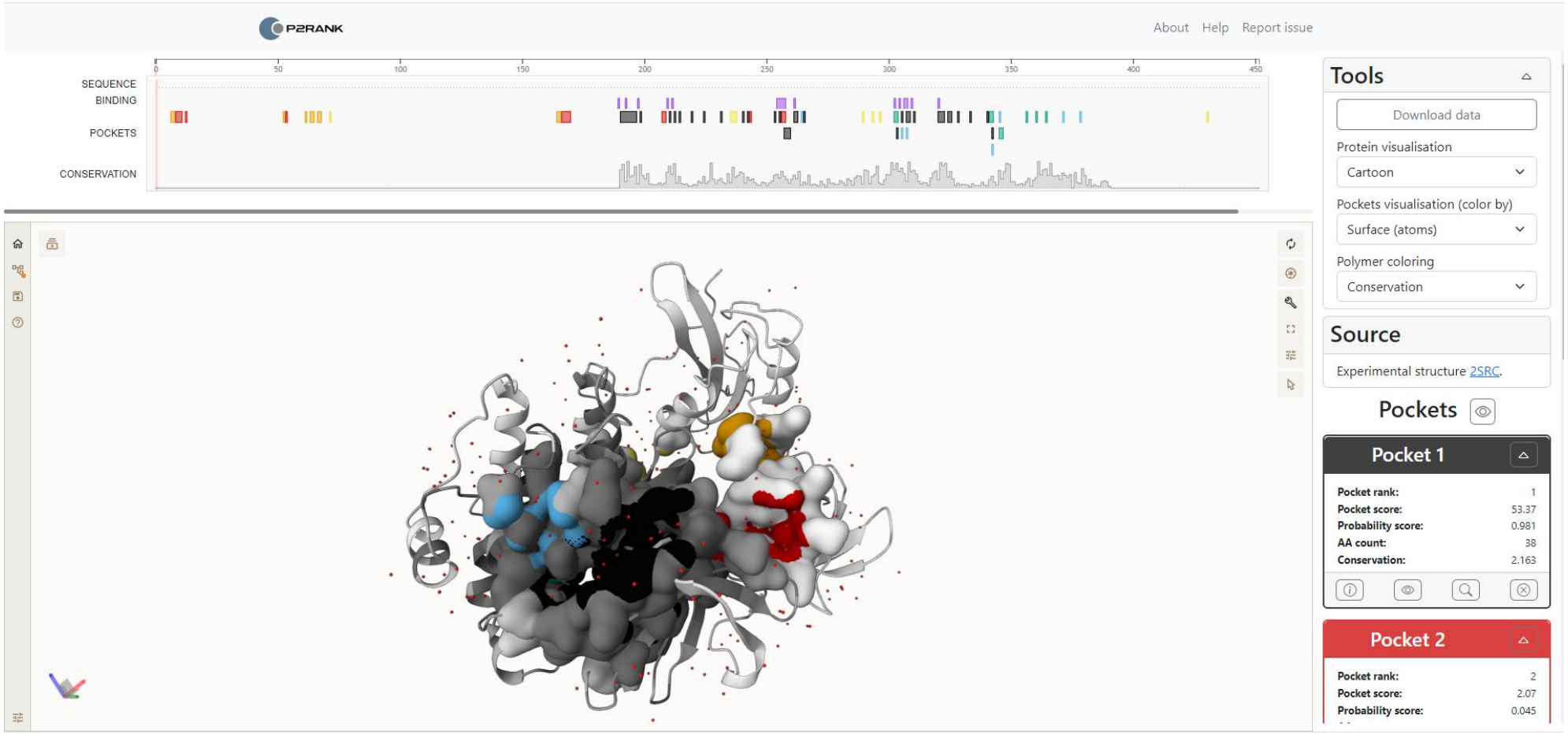






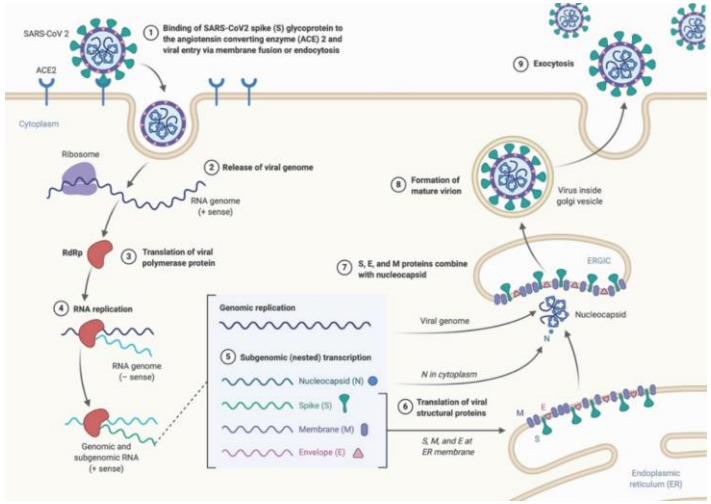


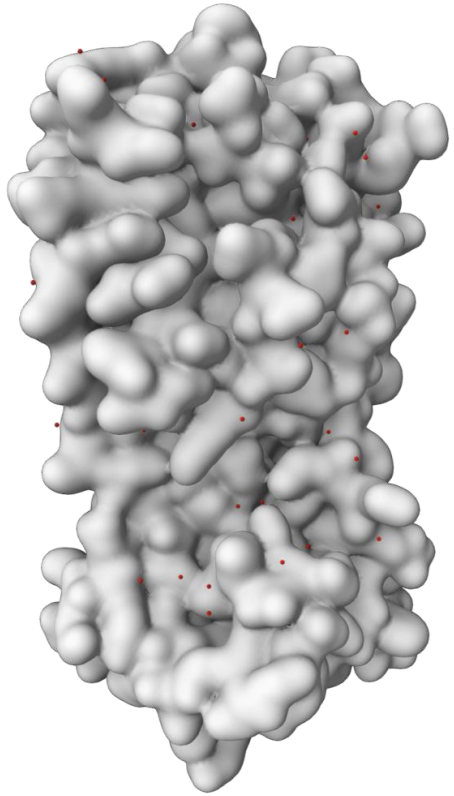
PrankWeb

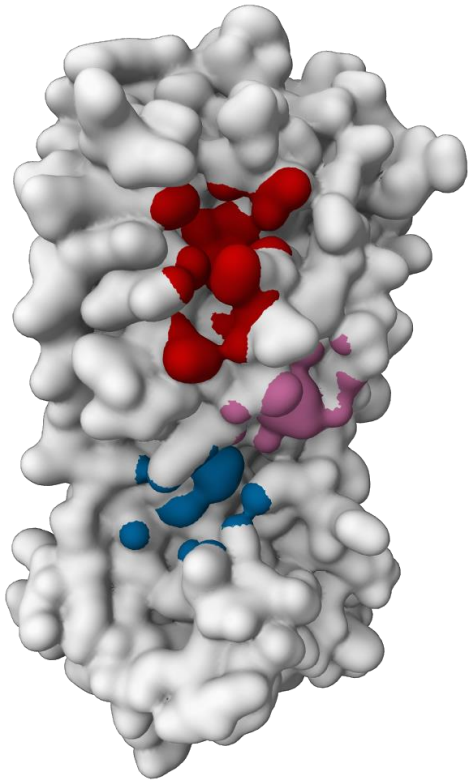


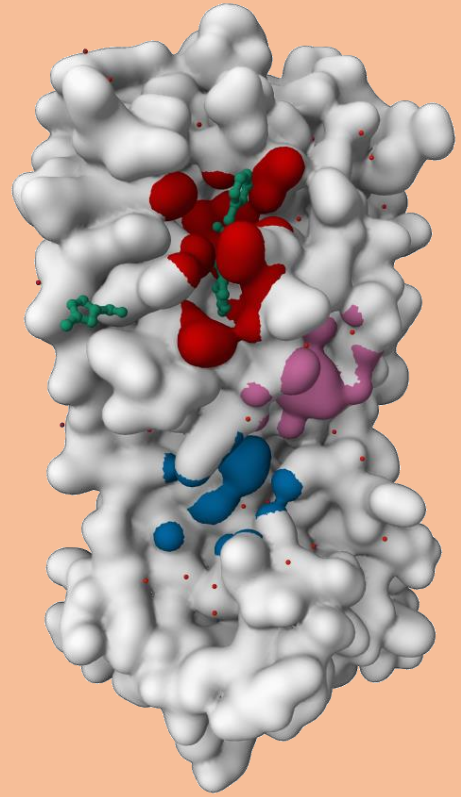
https://prankweb.cz

Racionální design









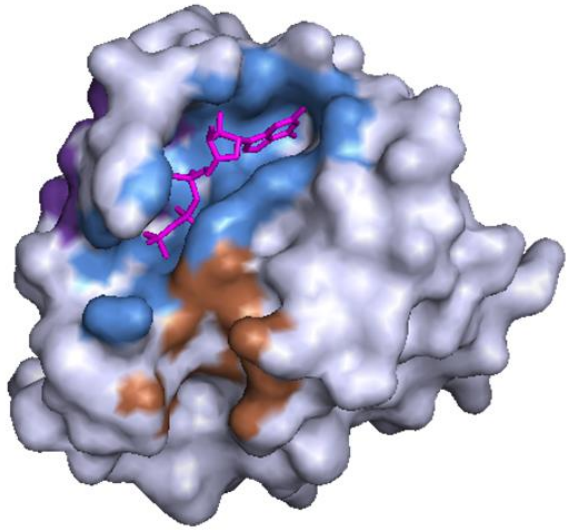
POCHOPENÍ BIOLOGICKÉHO PROCESU

VÝBĚR CÍLE

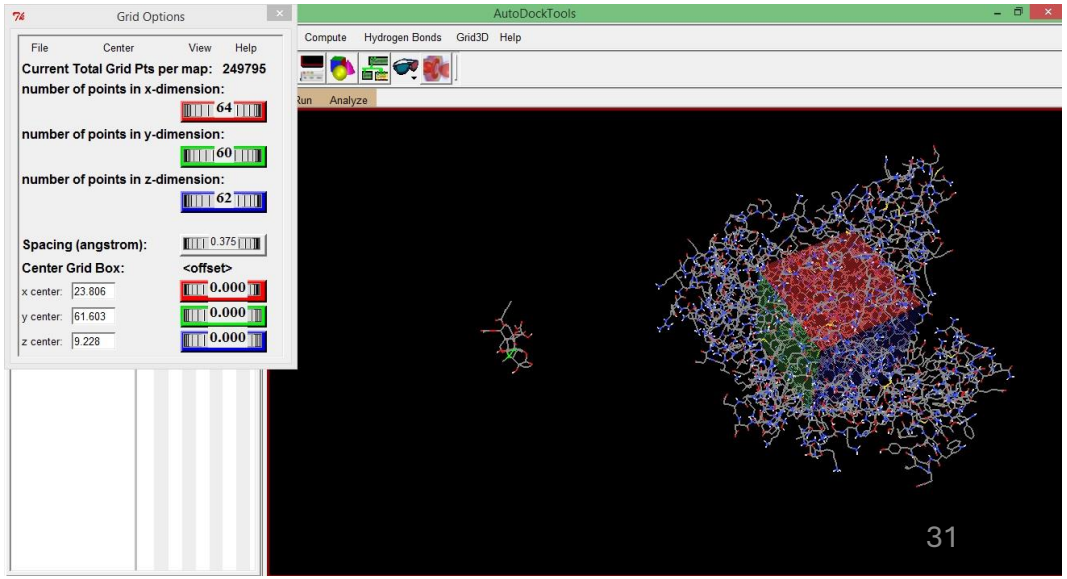
IDENTIFIKACE POTENCIÁLNÍCH VAZEBNÝCH MÍST

IDENTIFIKACE KANDIDÁTNÍCH MOLEKUL

Protein-ligand dockování

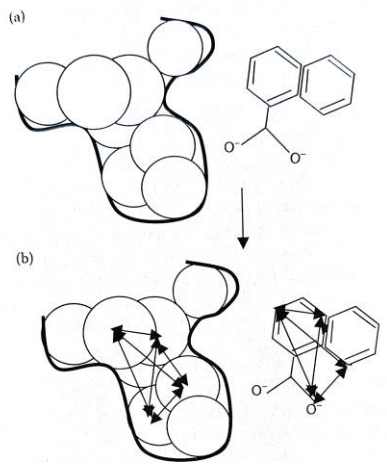


- **Generování póz**
  - Zacílení na potenciální aktivní místo
  - Možné orientace, lokace a konformace ligandu v aktivním místě
  - Rigidní vs. flexibilní
- **Skórovací funkce**
  - Silové pole
  - Empirické
  - Znalostní

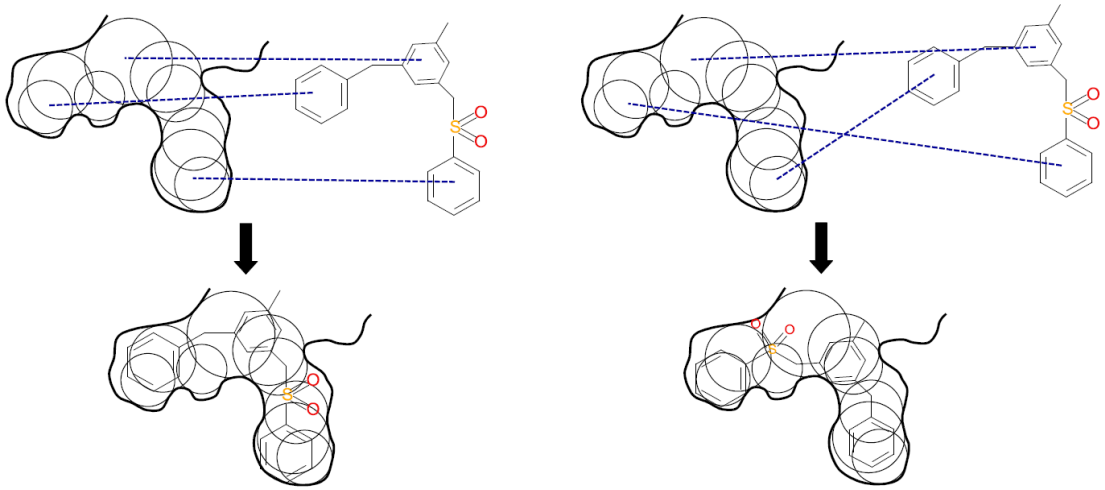


Vyhledávání póz

- Tvarová komplementarita
- Systematické prohledávání prostoru
- Stochastické algoritmy
- MD simulace



source: Faulon, J. L., & Bender, A. (2010). Handbook of chemoinformatics algorithms. CRC press.

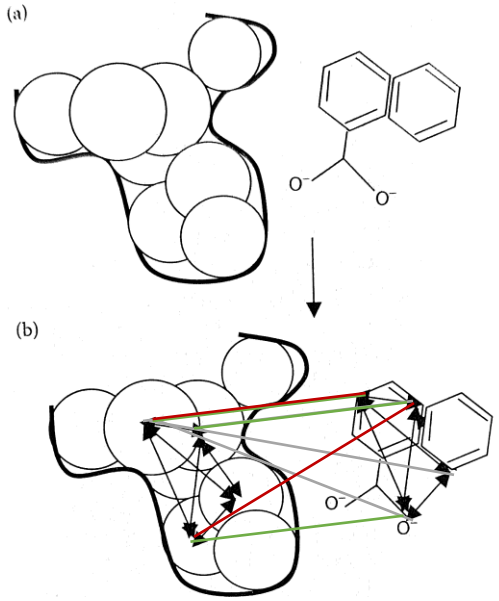


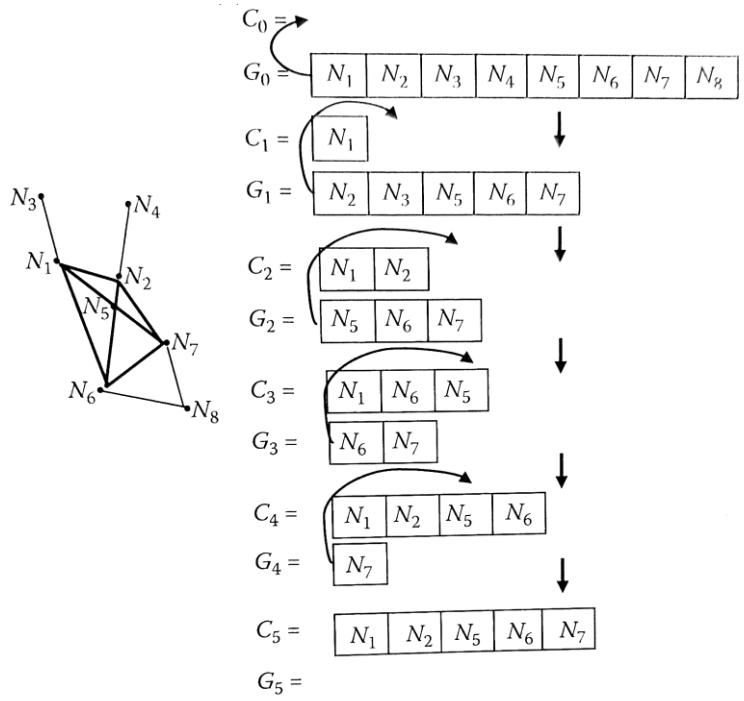
source: Leach, A. R., & Gillet, V. J. (2007). *An introduction to chemoinformatics*. *An Introduction To Chemoinformatics*. http://doi.org/10.1007/978-1-4020-6291-9

Vyhledávání póz – tvarová komplementarita

- Reprezentace prostoru **grafem**
  - vrcholy = koule vazebného místa × koule ligandu
  - hrany = podobné vzdálenosti mezi středy koulí

- Hledání **maximální kliky** v grafu





source: Faulon, J. L., & Bender, A. (2010). Handbook of chemoinformatics algorithms. CRC press.

Vyhledávání – explorace konformačního prostoru ligandů/fragmentový přístup

Identify flexible units in ligand

Start with the largest rigid unit/set of base units

Order flexible groups in layers starting from rigid unit

Orient rigid unit using orientational search/shape complementarity

Loop over flexible layers

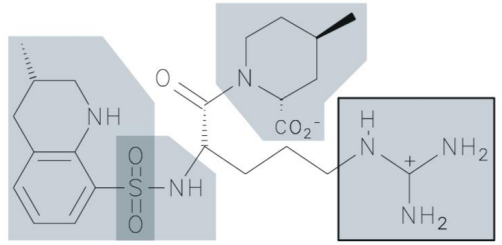
    Add next layer

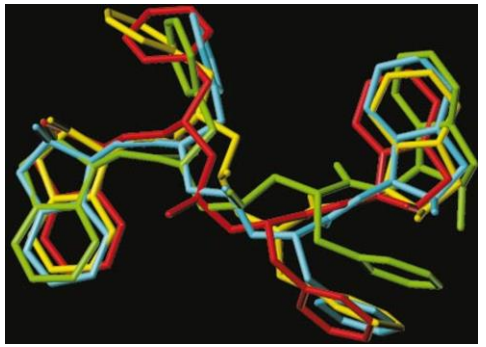
    Search torsions for each flexible group in layer

    Prune by score

End loop

Minimize and score the final structure

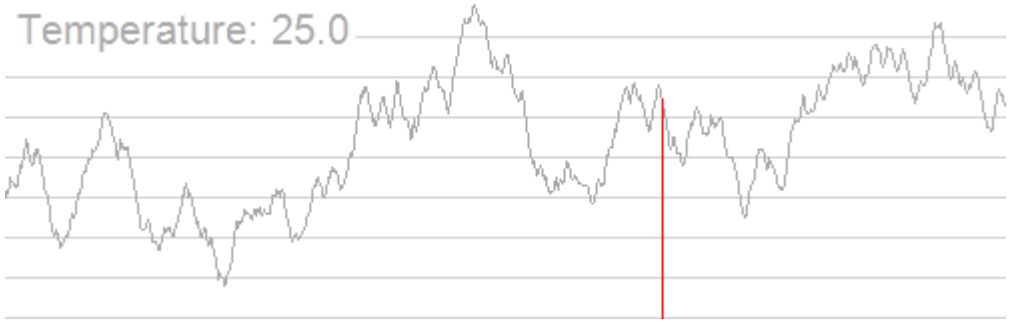


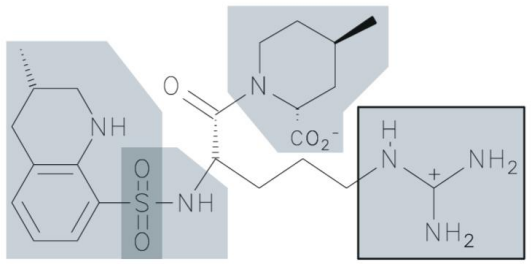


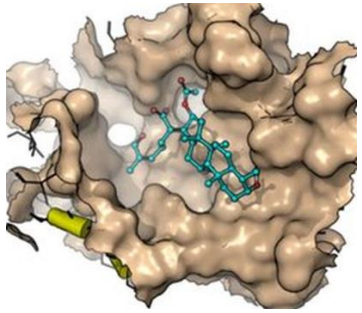
source: Rarey, M., Kramer, B., & Lengauer, T. (1997). Multiple automatic base selection: Protein–ligand docking based on incremental construction without manual intervention. Journal of computer-aided molecular design, 11, 369-384.

Vyhledávání póz – simulované žíhání

1. Nastav iniciální konfiguraci
2. Nastav iniciální templotu T
3. Vyber novou konfiguraci $\boldsymbol{\sigma'}$ **z blízkého okolí** $\boldsymbol{\sigma}$
4. **Vyhodnoť energii** E a E' konfigurací σ a σ'
5. **Pokud** $(\boldsymbol{E' < E}$ **nebo** $\boldsymbol{rand(0,1) < e^{-(E'-E)/T}})$ **pak** $\boldsymbol{\sigma \leftarrow \sigma'}$
6. **Sniž teplotu** a jdi na bod 3

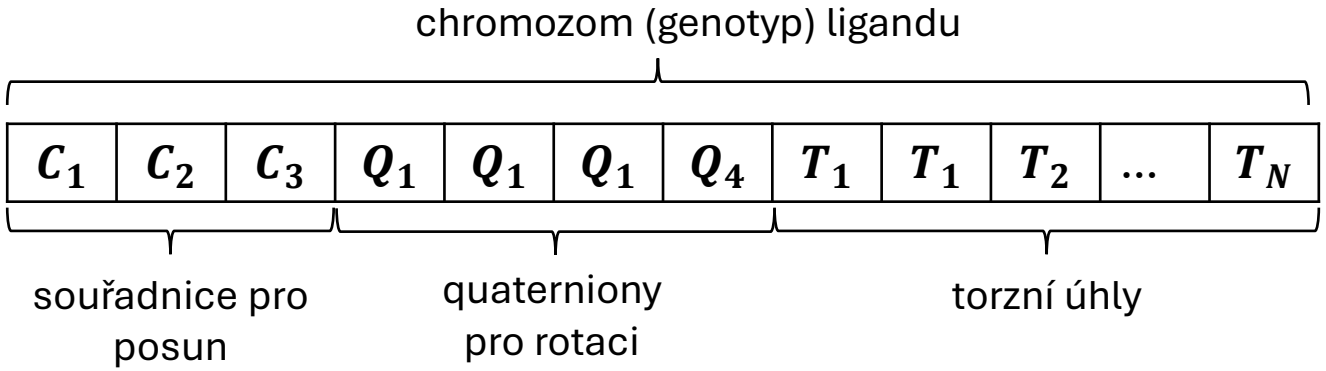


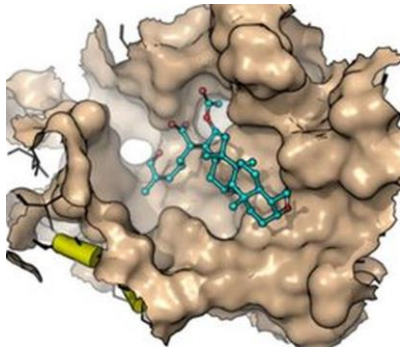




Vyhledávání póz – genetické algoritmy

- Populace velikosti N
- V každém kroku
  1. Mapuj pro každého jedince populace genotyp na fenotyp
  2. Vyhodnoť fitness každého jedince
  3. Vyber jedince pro křížení do další populace
  4. Mutace a crossover
- Lokální optimalizace





fenotyp

Skórování

- **Silové pole**
  - Interakční energie pro všechny typy interakcí
- **Empirické**
  - Volná energie vazby
  - Kalibrováno experimentálními daty

$$\Delta G_{bind} = \Delta G_0 + \Delta G_{hb} \sum_{h-bonds}' f(\Delta R, \Delta\alpha) + \Delta G_{ionic} \sum_{\substack{\text{ionic} \\ \text{interactions}}} f(\Delta R, \Delta\alpha) + \Delta G_{lipo} \left| A_{lipo} \right| + \Delta G_{rot} NROT$$

- **Znalostní**
  - Statistiky vzdáleností, frekvencí určitých typů párů atomů spočítané ze známých komplexů

Virtuální screening

Založené na struktuře cíle

Založené na znalosti ligandů

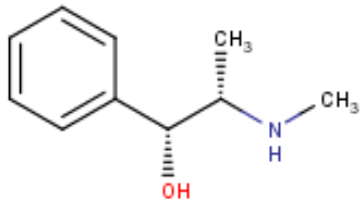
Strukturní modelování

Podobnostní hledání

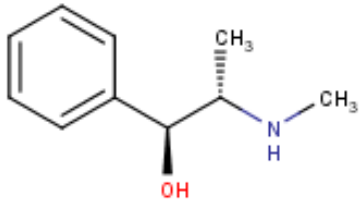
Farmakoforové modelování

Podobnostní princip

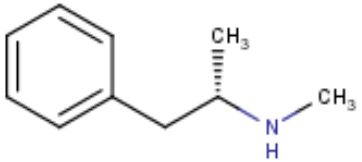
- Podobné molekuly mají tendenci se vázat na daný cíl/mít podobnou funkci častěji než molekuly nepodobné



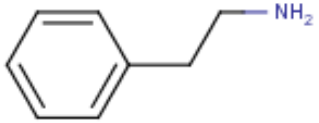
ephedrine
(stimulant)



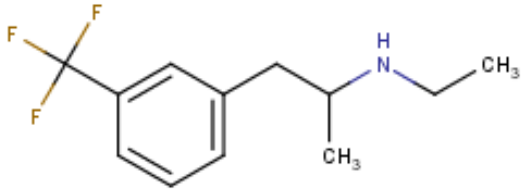
pseudoephedrine
(dekongestivum)



metamphetamin
(stimulant)

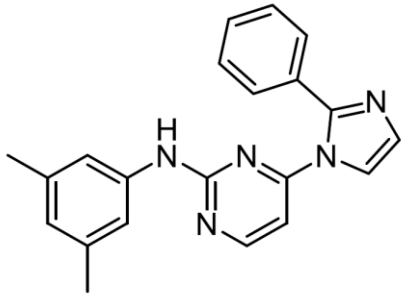


phenethylamine

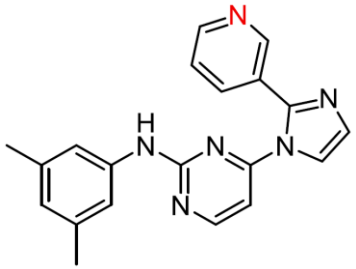


fenfluramine

- Aktivitní útesy

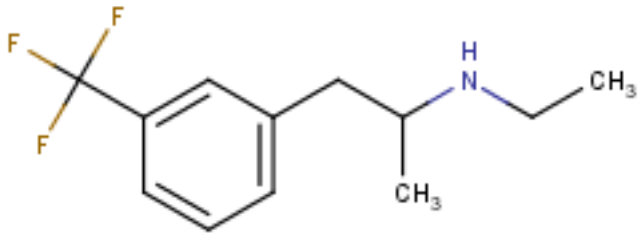


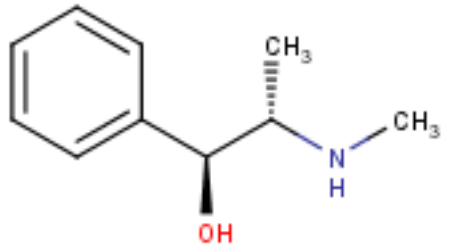
6 nM



2390 nM

Podobnost molekul





Charakteristiky

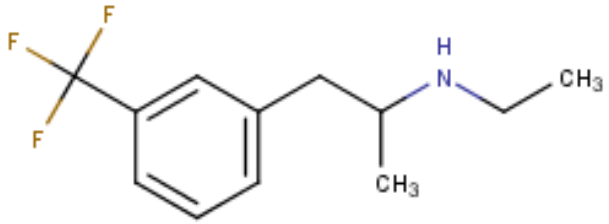
- Fyzikálně-chemické charakteristiky
  - váha, počet elektronů, rozpustnost, polární plocha, rozdělovací koeficienty (logP), statistiky
- 2D struktura
- 3D struktura

Charakteristiky

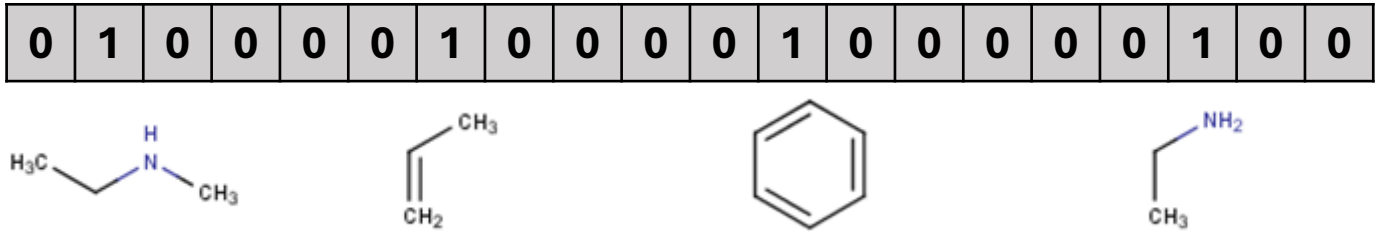
Podobnostní míra

Podobnost

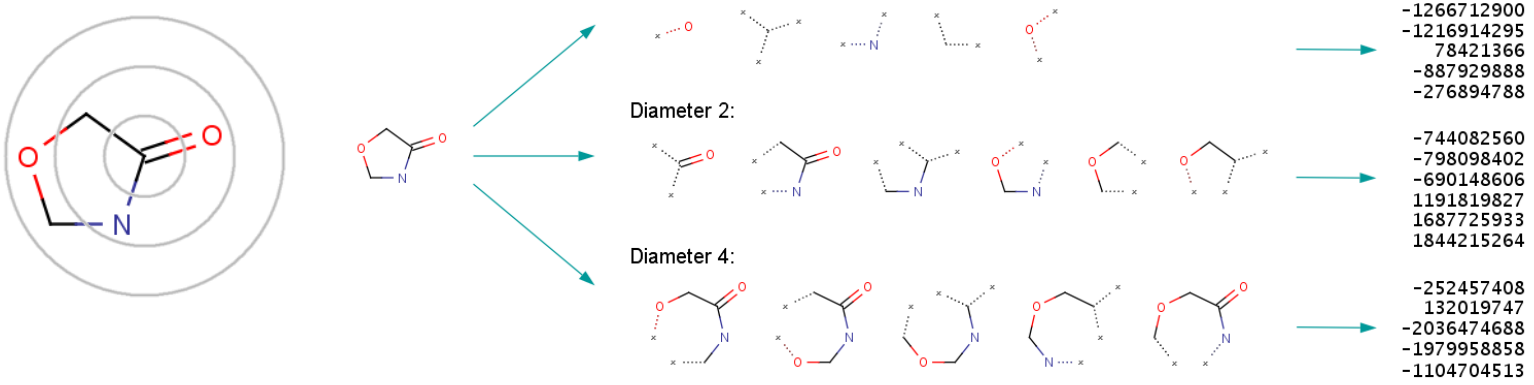
Molekulární otisky molekul

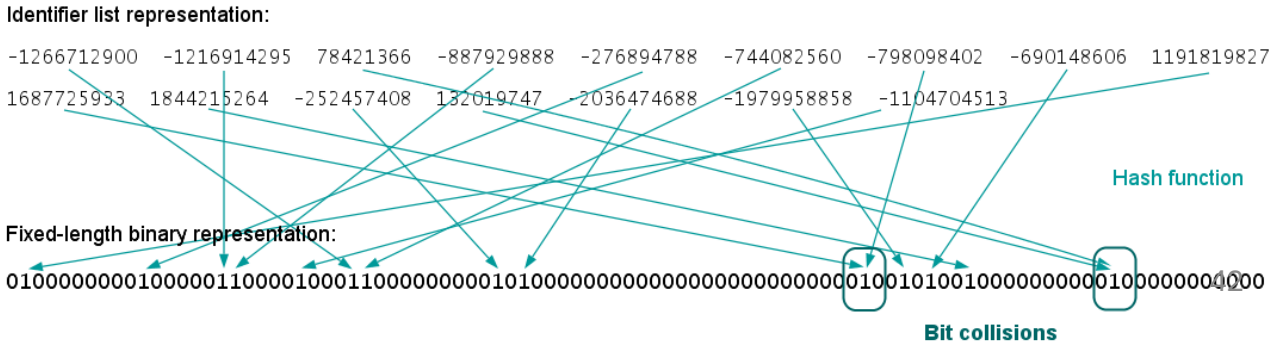


- **Slovníkové** otisky

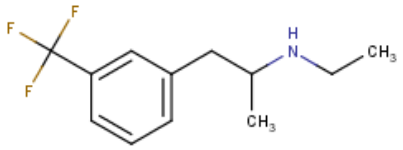


- **Hashované** otisky
  - Cirkulární otisky
  - Cesty dané délky
  - Atomové páry





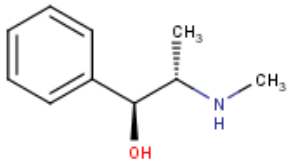
Podobnostní míra



1	0	1	1	0	1	0	1	1	0
1	0	1	1	0	1	1	0	1	0

$$\frac{A \,\&\, B}{A \mid B}$$

$$\text{podobnost} = \frac{5}{6+6-5} = \frac{5}{7} = 0{,}72$$



Virtuální screening

Založené na struktuře cíle

Založené na znalosti ligandů

Strukurní modelování

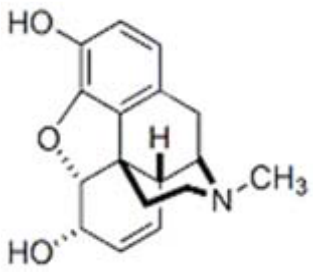
Podobnostní princip

Farmakoforové modelování

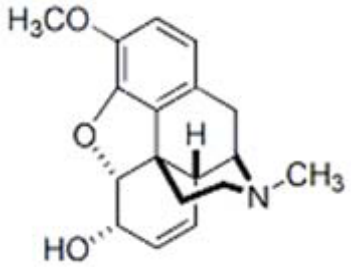
Farmakofor

- Podle definice IUPAC je **farmakofor** definován jako

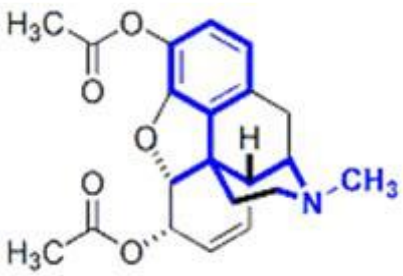
  “the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response”



morphine
1



codeine
2

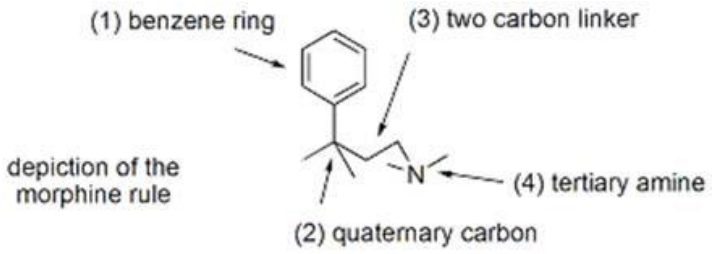


heroin
3

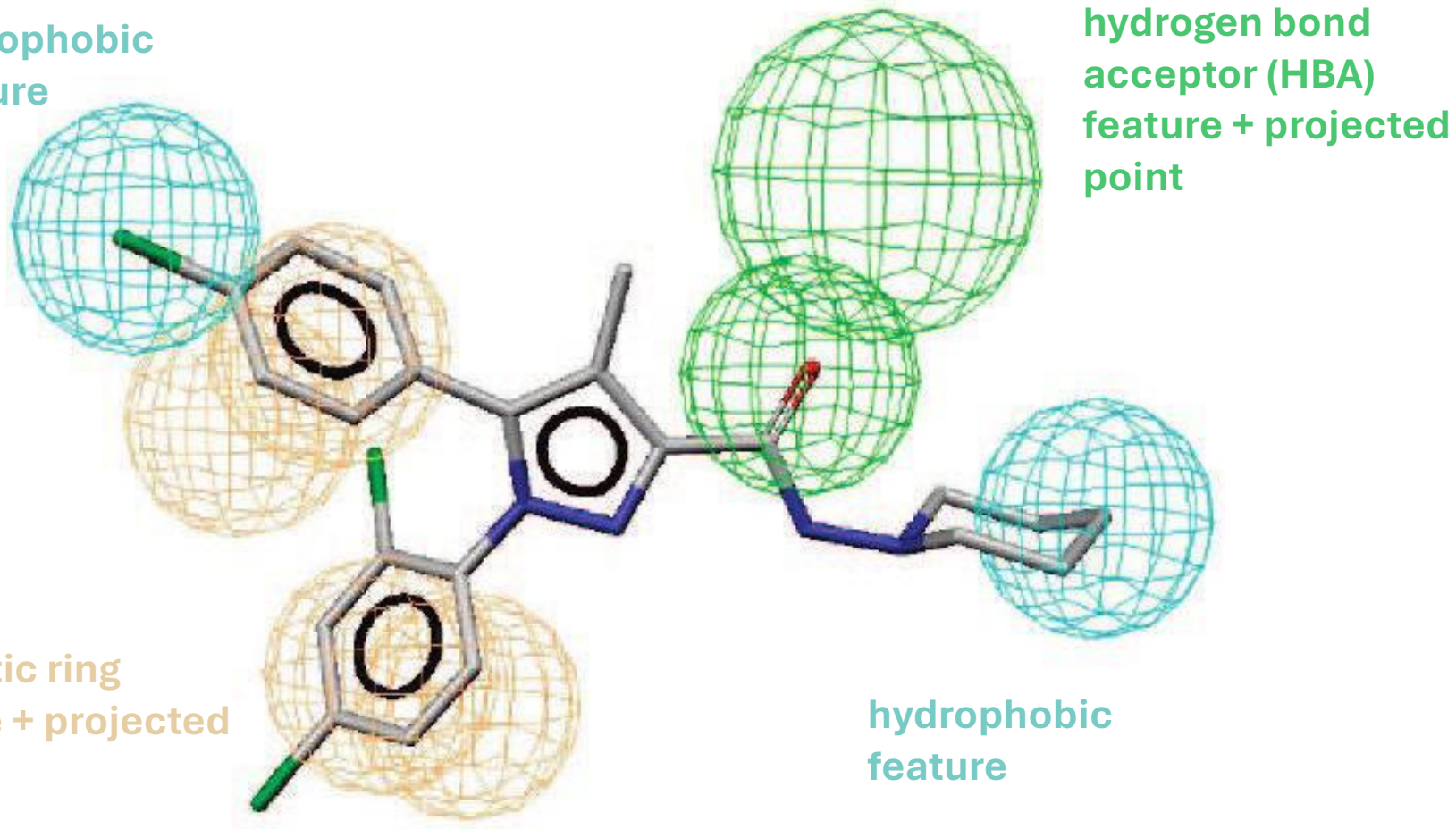
pharmacophore highlighted

methadone
4

pharmacophore highlighted



Farmakoforové vlastnosti

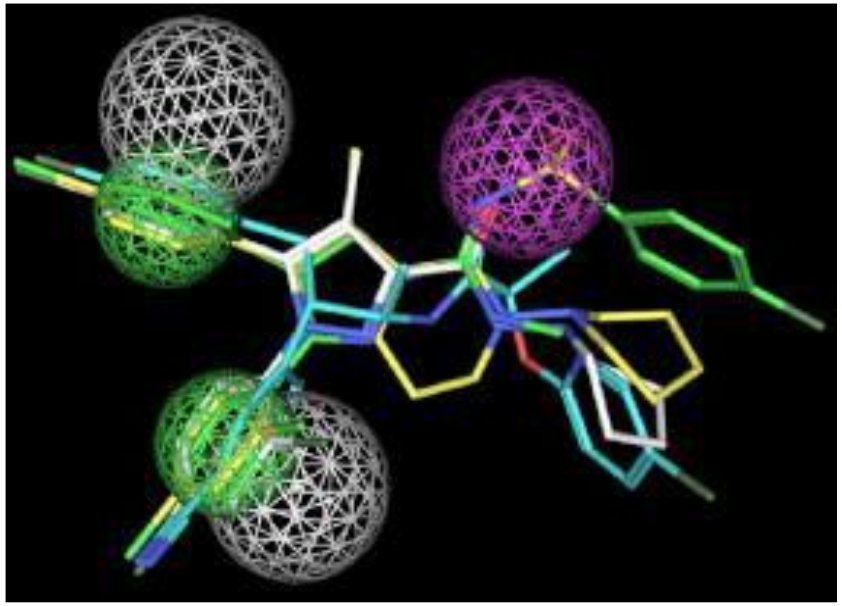


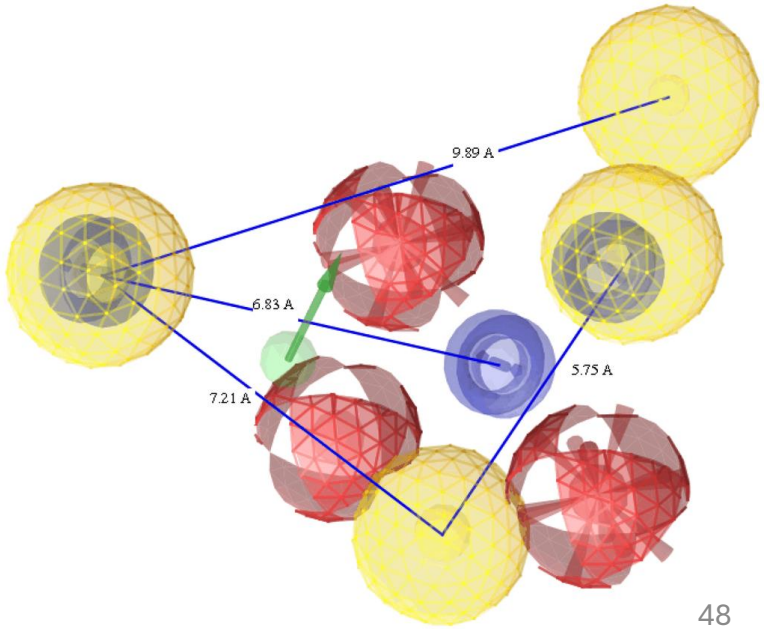
3D farmakoforový model

- **Identifikace farmakoforových vlastností** v každém z ligandů

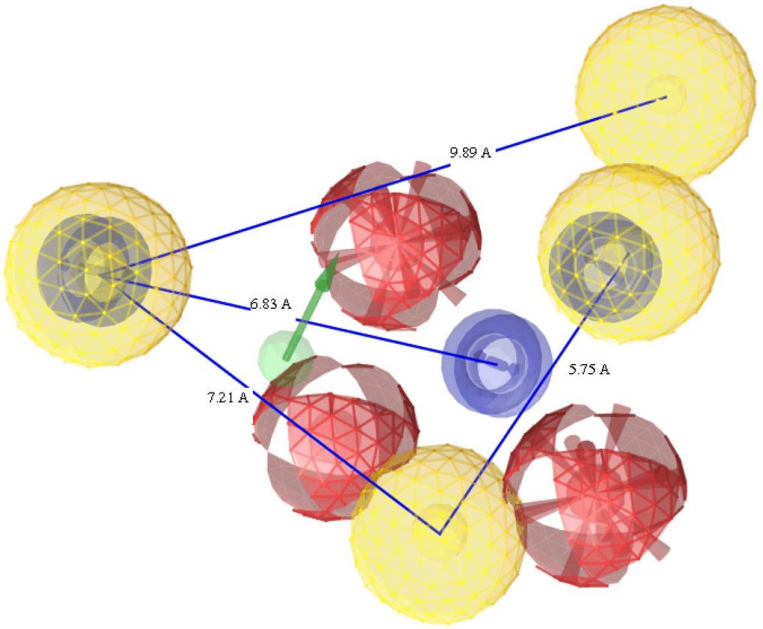
- Prostorová **superpozice** ligandů (může zahrnovat predikci prostorové struktury)

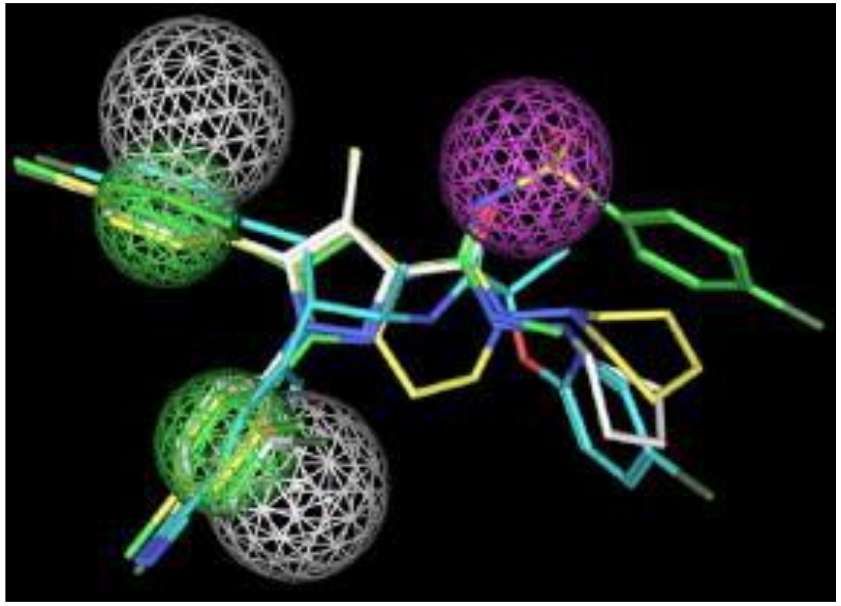
- Identifikace **společných farmakoforových vlastností** (automaticky nebo manuálně)





Aplikace farmakofórového modelu





Virtuální screening

Založené na struktuře cíle

Strukturní modelování

Založené na znalosti ligandů

Podobnostní hledání

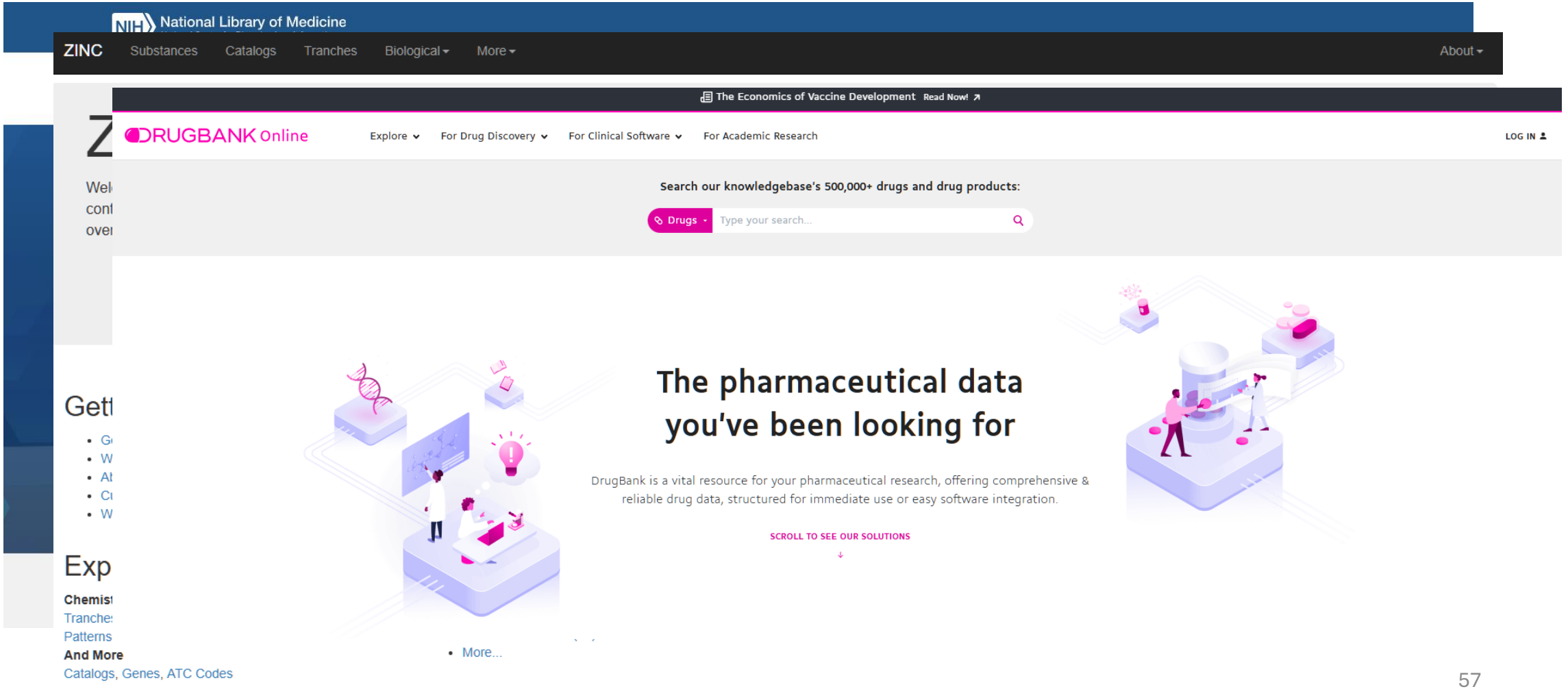
Farmakoforové modelování

BIOIFORMATIKA

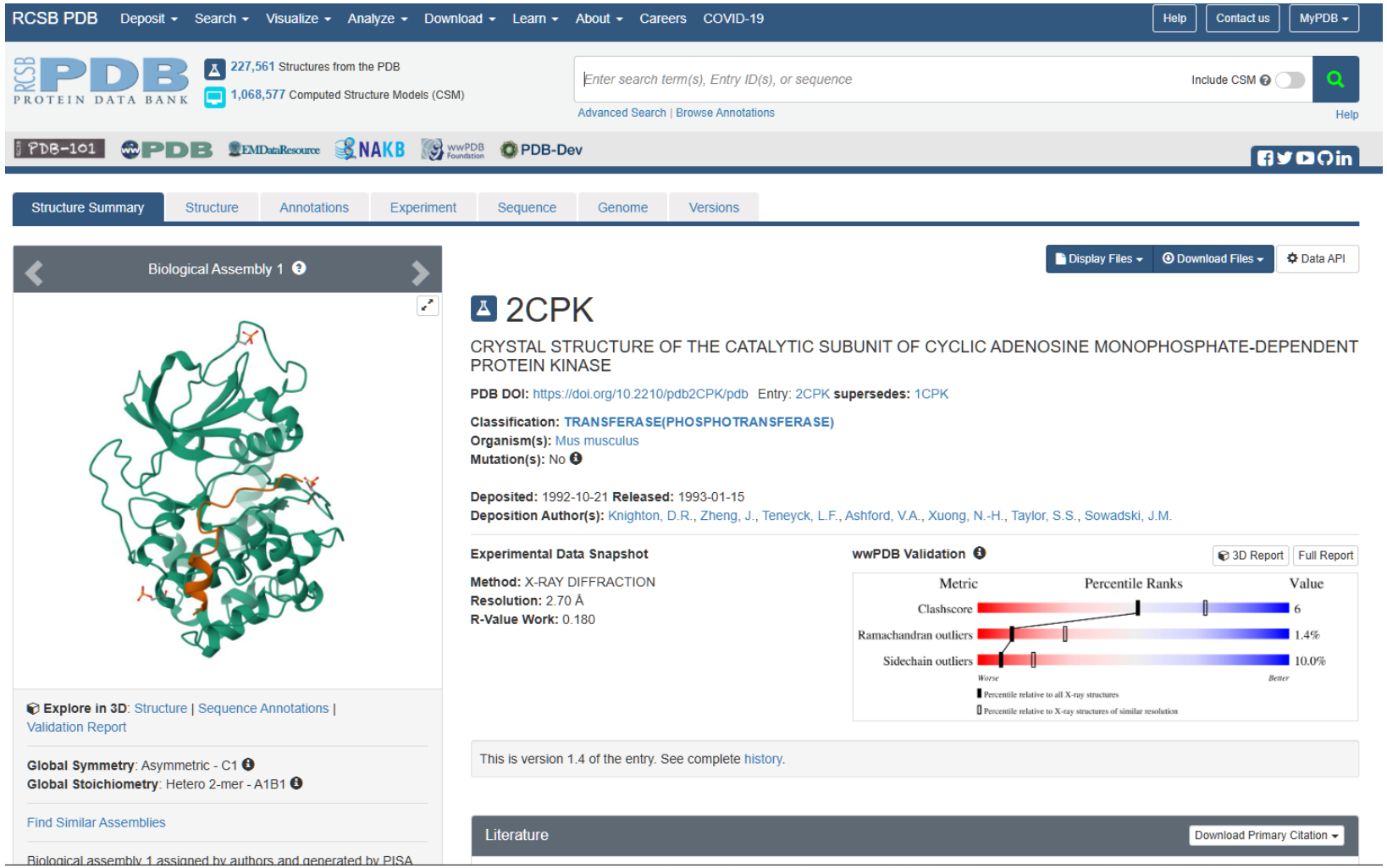
CHEMINFORMATIKA

Databáze sloučenin

Databáze malých molekul



Databáze proteinových struktur



Za hranice známého prostoru

Malé molekuly

Chemický prostor

- **Všechny organické sloučeniny** tvoří "**chemický prostor**"
  - Odhadovaná velikost chemického prostoru: 10^{100} - 10^{200}

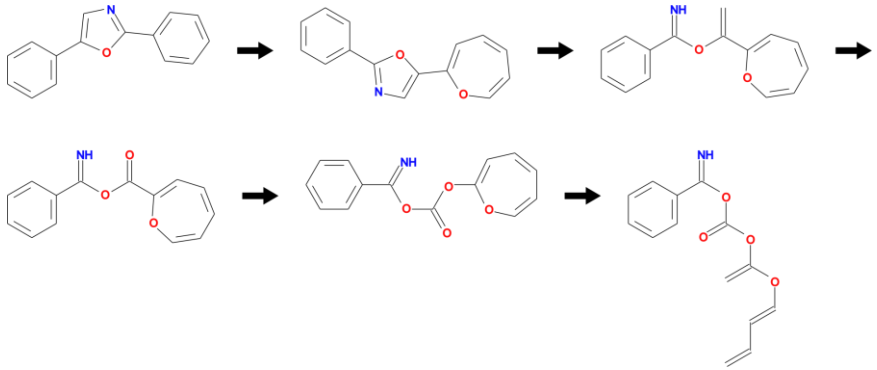
- Ne všechny teoretický možné sloučeniny jsou **syntetizovatelné**

  - Automatické generování prostoru

    - GDB11 databáze – 11 atomů (C, N, O, F), 26.4 milionů molekul
    - GDB13 databáze – 13 atomů (C, N, O, S, Cl), 970 milionů molekul
    - GDB17 databáze – 13 atomů (C, N, O, S, and halogens), 166.4 miliard molekul

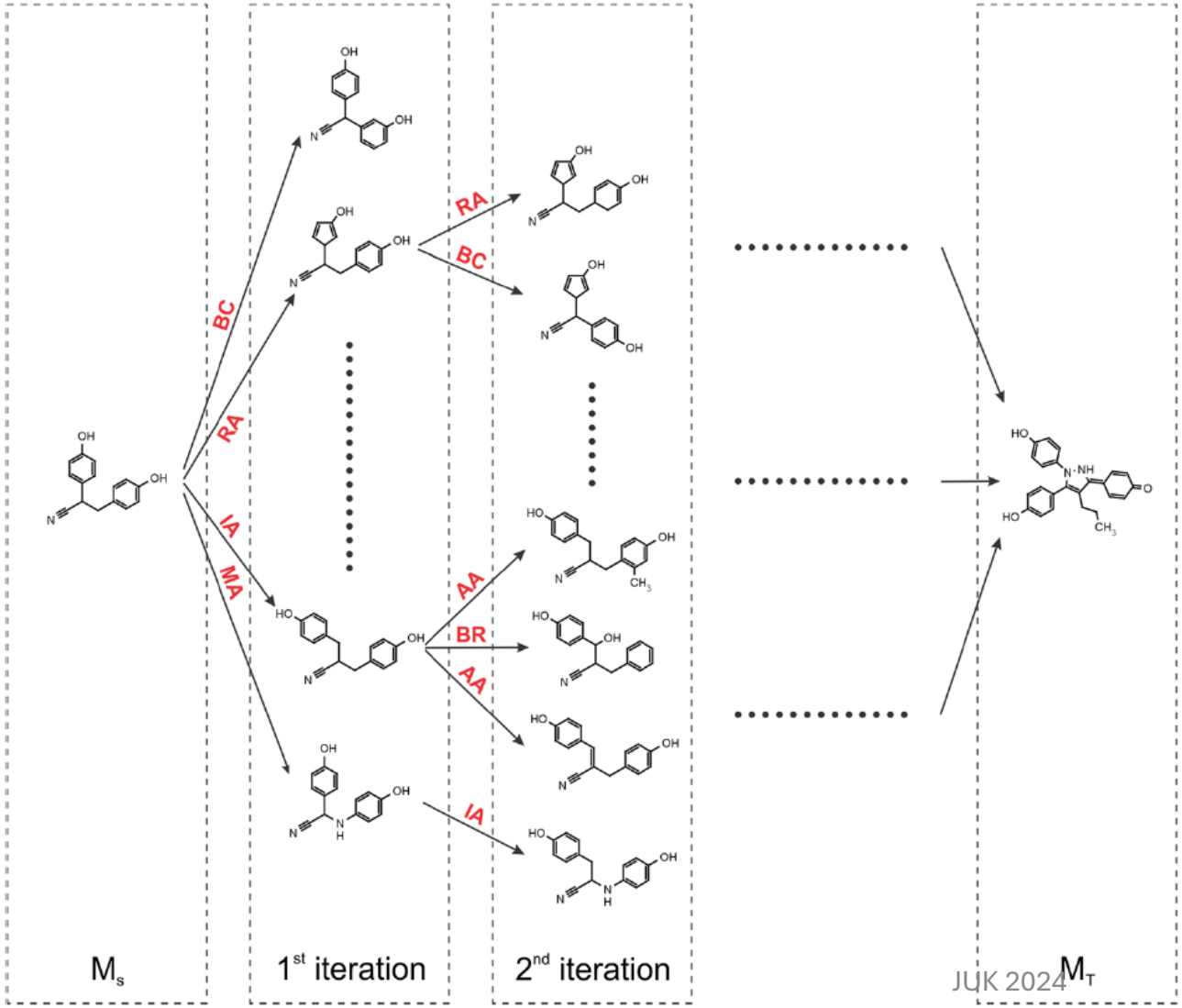
Explorace chemického prostoru

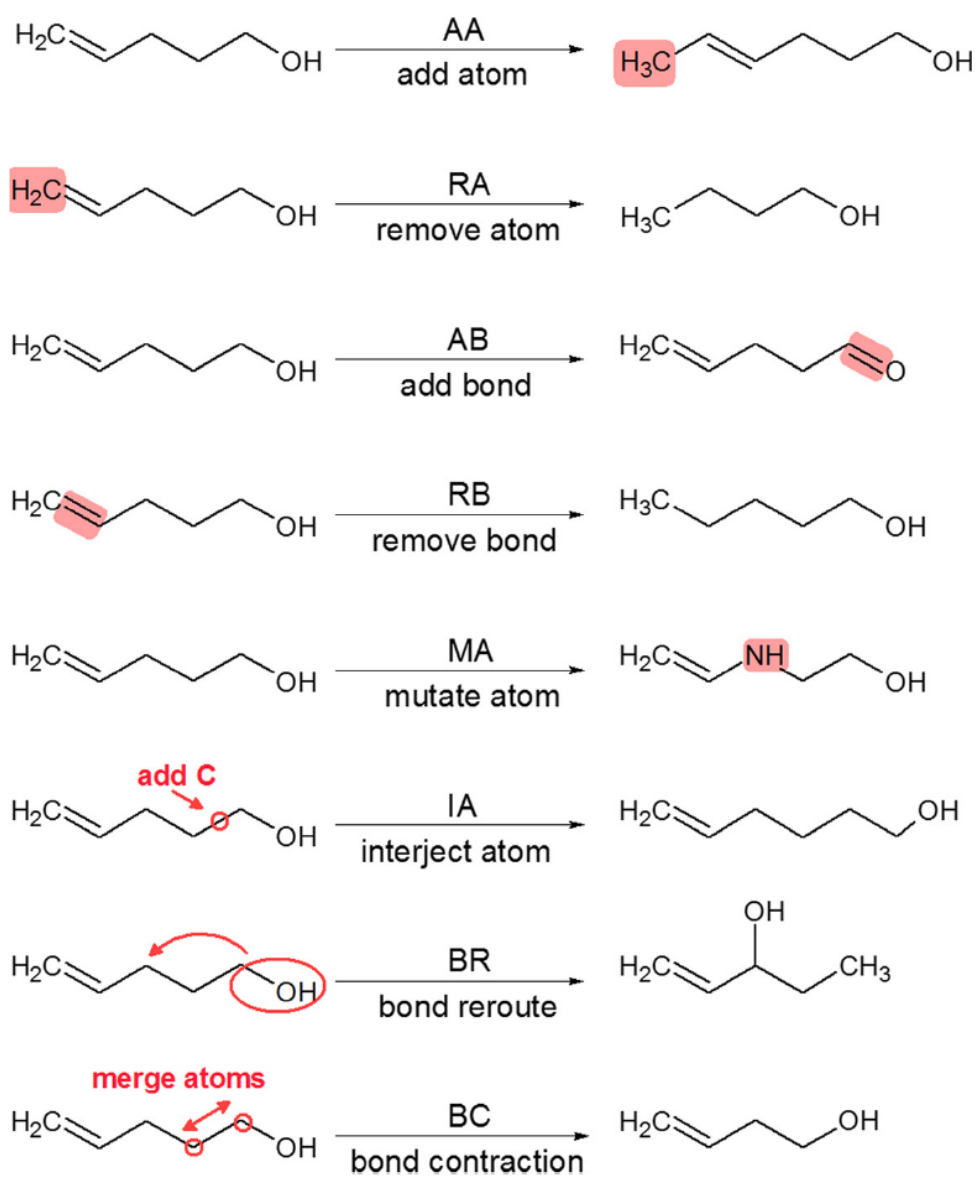
- Aplikace “molekulárního **morfingu**”

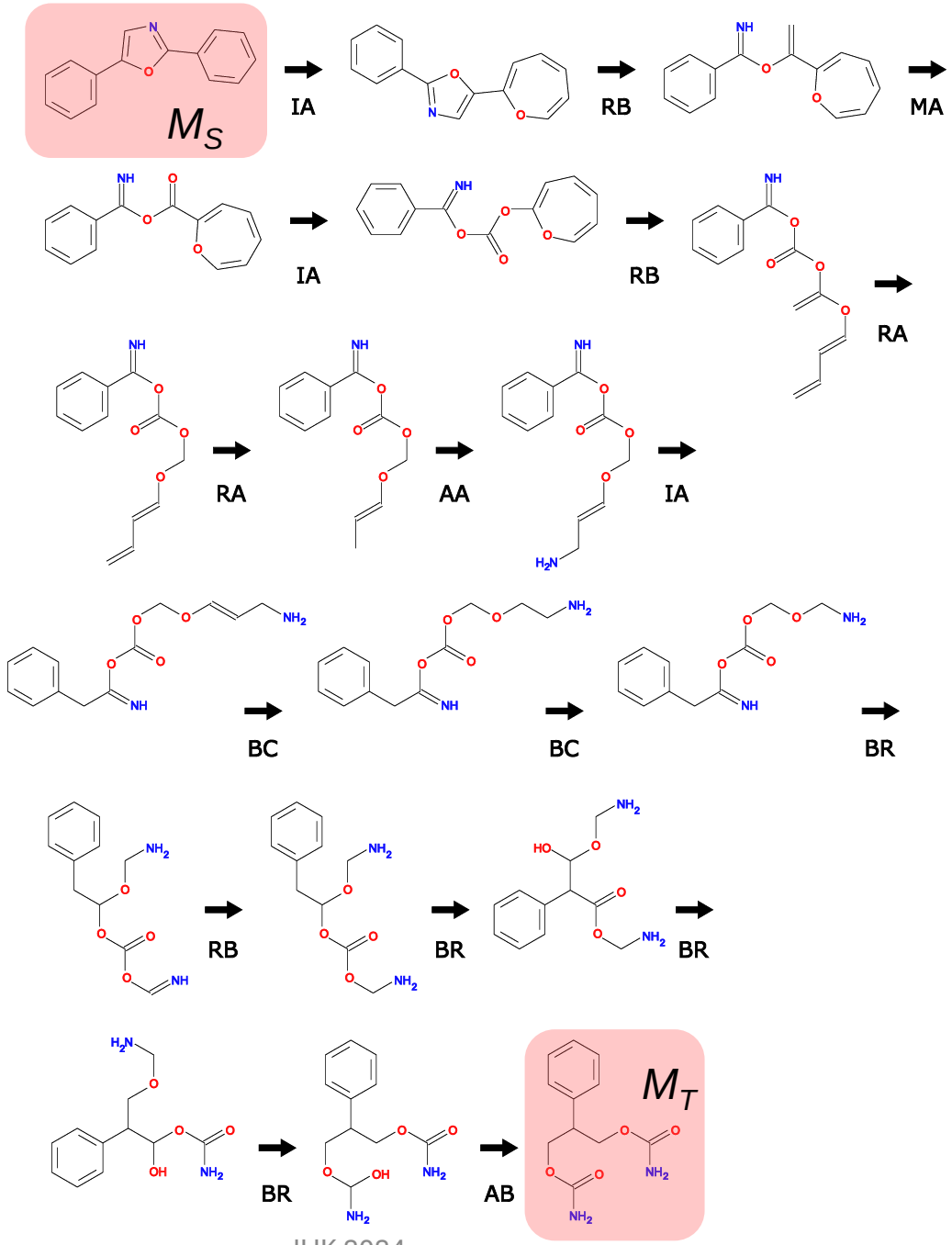


- Startovní a cílová molekula pochází ze stejné mechanistické třídy
  → **okolí cesty** reprezentuje **zacílenou virtuální knihovu**

Algoritmus Molpher



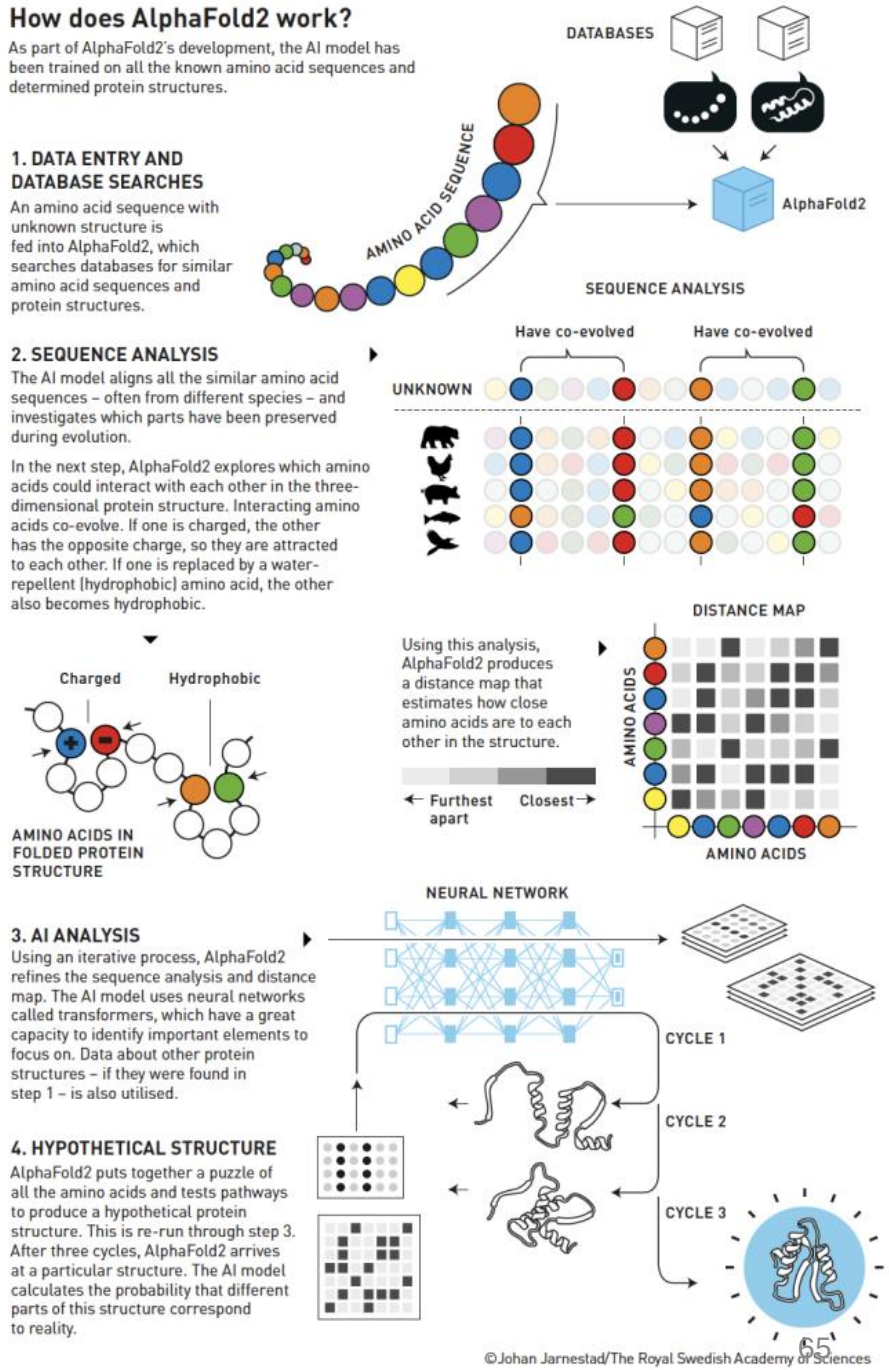



M_S
IA
RB
MA
IA
RB
RA
RA
AA
IA
BC
BC
BR
RB
BR
BR
BR
AB
M_T

Za hranice známého prostoru

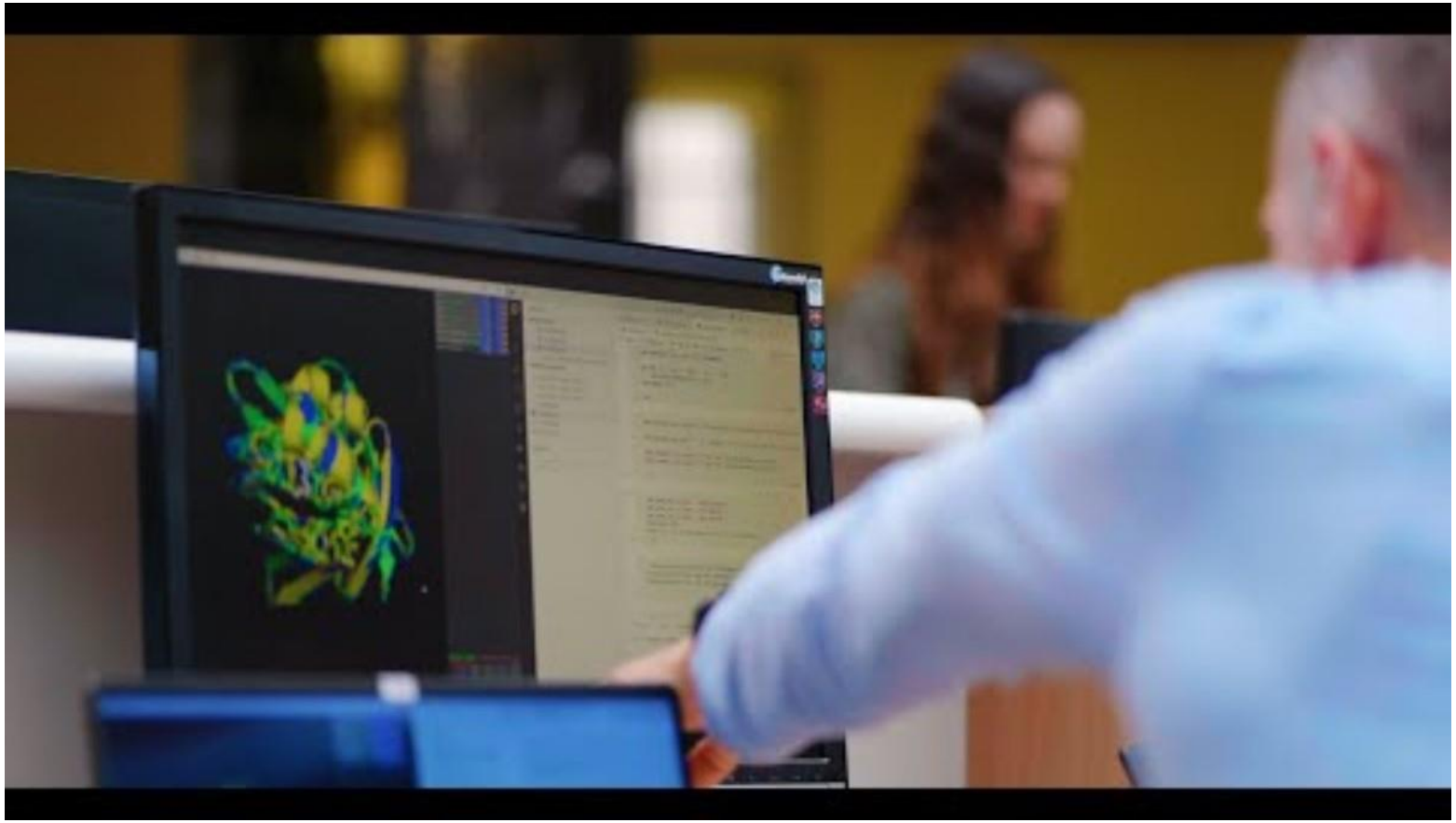
Makromolekuly

Proteinové struktury



source: https://www.nobelprize.org/uploads/2024/10/fig2_ke_en_24.pdf

Bioinformatika není jenom vývoj léčiv!!!!



https://www.youtube.com/@CharlesUniversityBioinfo

Dotazy