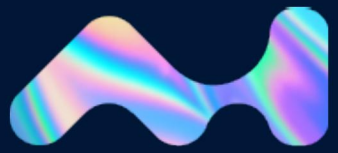


AI CENTER
FEE CTU

Využití AI v oblasti predikce vlastností chemických látek

Milan Papež

Centrum Umělé Inteligence

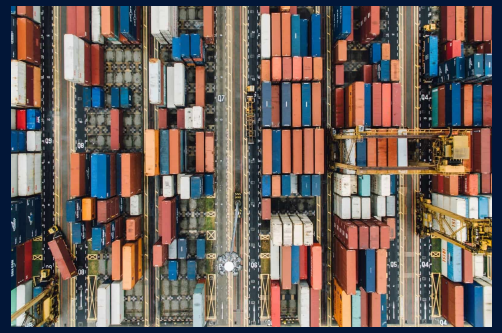
Fakulta Elektrotechniky

České Vysoké Učení Technické v Praze

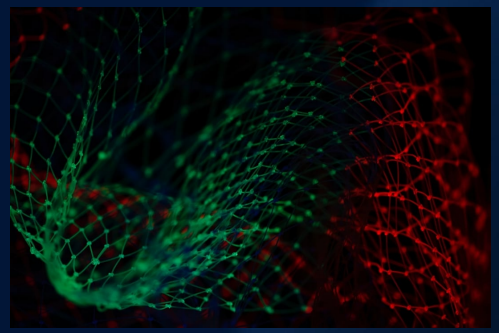
18. 9. 2025

Centrum Umělé Inteligence

Automatické Plánování

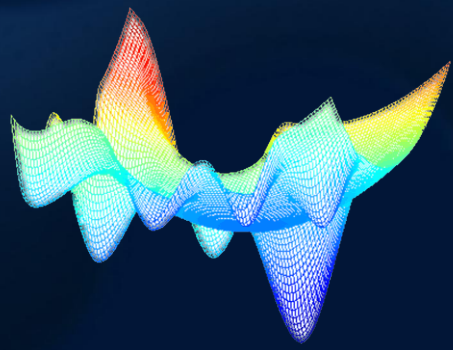


Strojové Učení

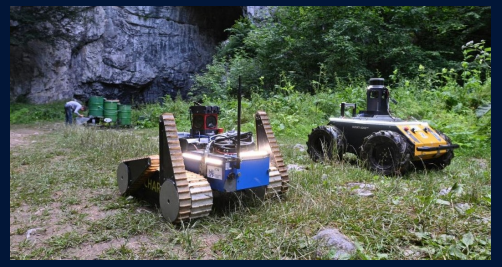


Teorie Her

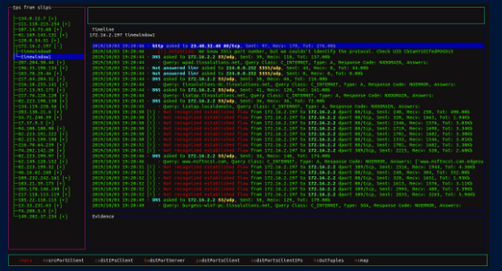
Optimalizace



Robotika



Kybernetická Bezpečnost



Chytrá Doprava

AI Center FEE CTU

Skupina Strojového Učení

AI Center FEE CTU

Členové Skupiny

Vedení skupiny

Tomáš Pevný

Vašek Šmídl

Postdoktorandi

Milan Papež

Ph.D. studenti

Matěj Zorek

Michaela Mašková

Šimon Mandlík

Martin Rektoris

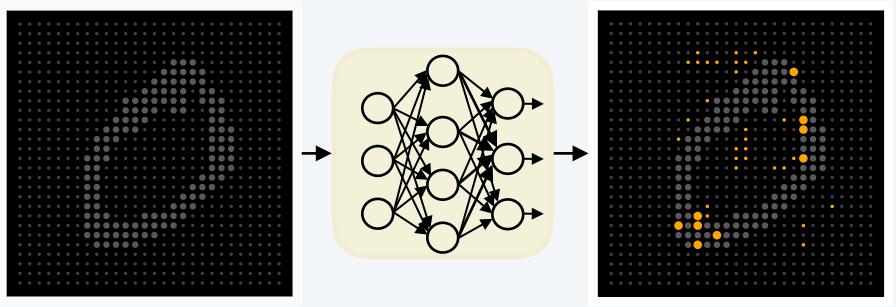
Armin Hadžić

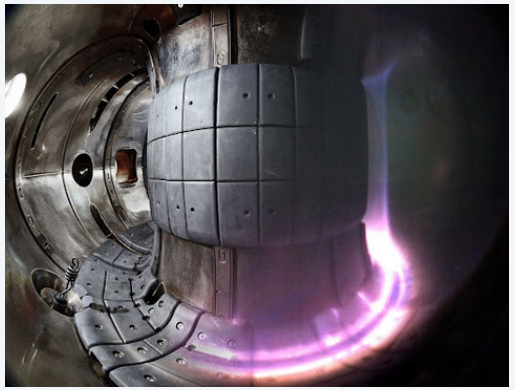
Oleksii Shuhailo

Šimon Soldát

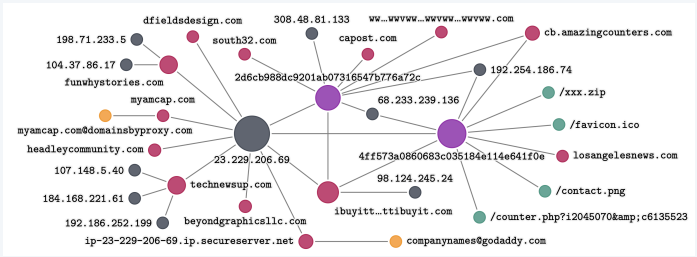
AI Center FEE CTU

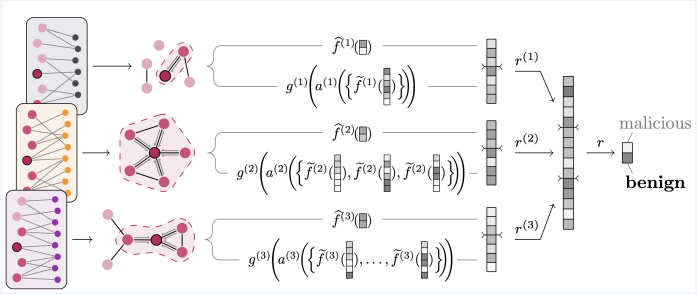
Zaměření Skupiny

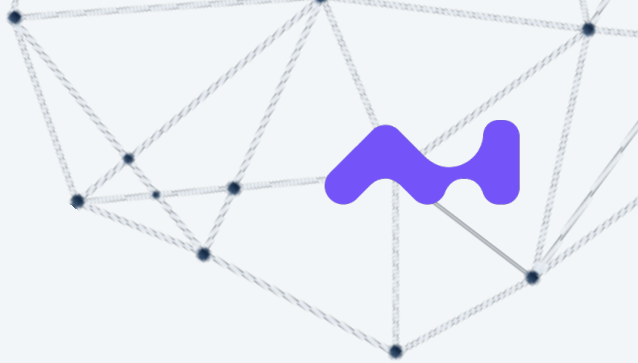


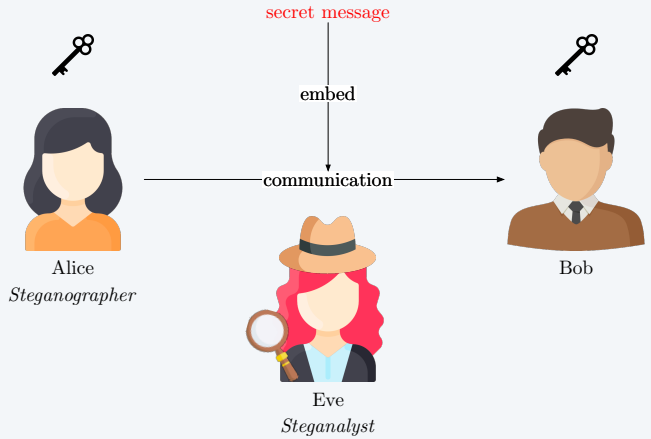


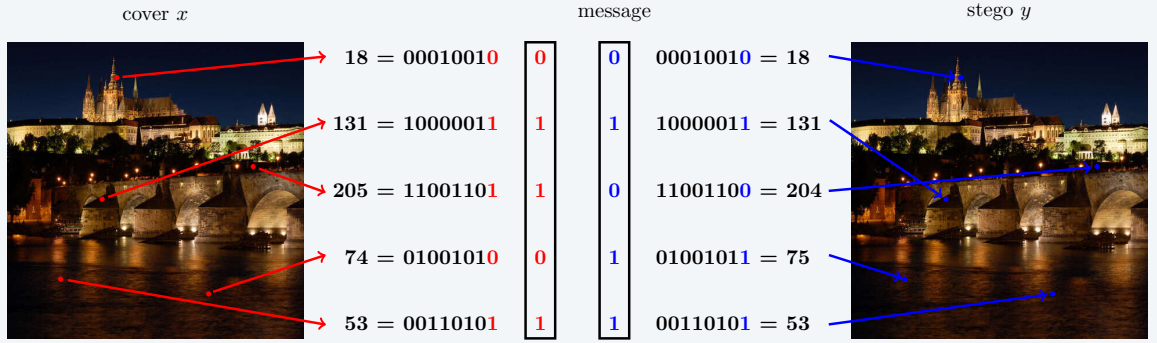
AI Center FEE CTU











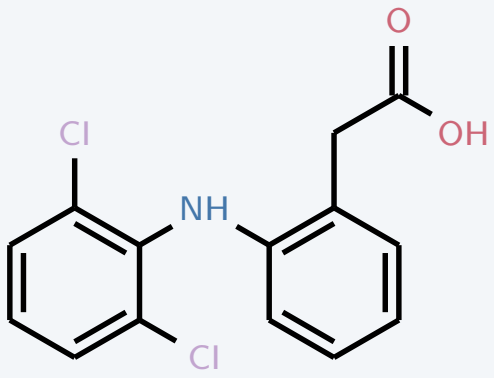
Strojové Učení v Chemii

AI Center FEE CTU

Datové Reprezentace Molekul

Diklofenac ($C_{14}H_{11}Cl_2NO_2$)

Grafická reprezentace



Textová reprezentace

SMILES	C1=CC=C(C(=C1)CC(=O)O)NC2=C(C=CC=C2Cl)Cl
SELFIES	[C][=C][C][=C][Branch1][=N][C][=Branch1][Ring2][=C][Ring1][=Branch1][C][C][=Branch1][C][=O][O][N][C][=C][Branch1][Branch2][C][=C][C][=C][Ring1][=Branch1][Cl][Cl]
InChI	InChI=1S/C14H11Cl2NO2/c15-10-5-3-6-11(16)14(10)17-12-7-2-1-4-9(12)8-13(18)19/h1-7,17H,8H2,(H,18,19)

Maticová reprezentace

X

1
1
3
⋮
7

A

0	1	2	⋯	0
1	0	1	⋯	0
2	1	0	⋯	1
⋮	⋮	⋮		⋮
0	0	1	⋯	0

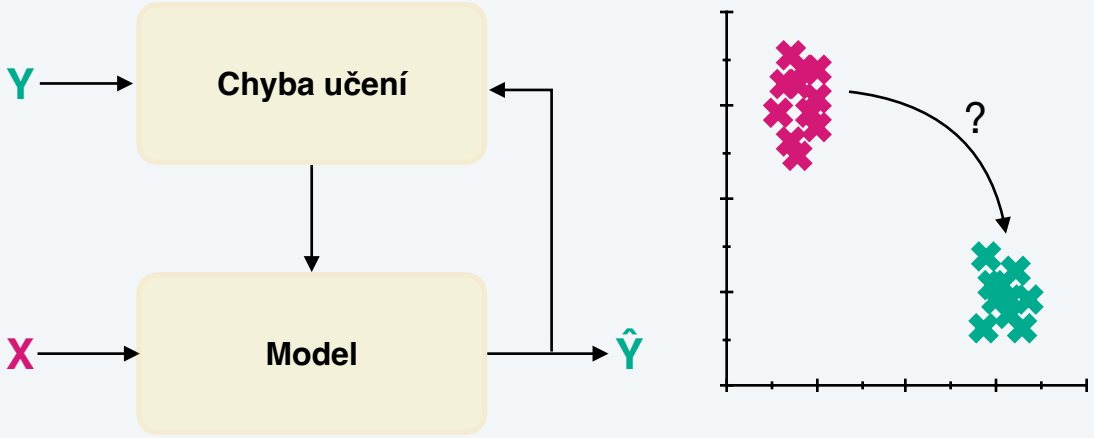
Molekulární Databáze

Dataset	Počet vzorků (miliony)	Maximální počet atomů (přibližně)	Poznámka
QM9	0,134	9	Simulované molekuly
Zinc250k	0,250	38	Filtrováno pro GenAI
Guacamol	1,6	88	Anotace bioaktivity
Moses	1,9	27	Převážně syntetizovatelné
GEOM	38	50	3D konformace
PubChem	86	91	Široké pokrytí molekulárního prostoru

SMILES	μ	α	homo	lumo	⋯	U	Cv
OC1CCCC(=O)C1=O	3.56	69.15	-0.2477	-0.0795	⋯	-458.99	30.94
CN(CC#N)C(=N)C#N	3.34	73.79	-0.2658	-0.0604	⋯	-412.99	32.01
CC1C2CCC=CC1O2	1.59	82.26	-0.2359	0.0099	⋯	-387.07	31.79
CN1N=NC(F)=CC1=N	1.69	69.88	-0.2373	-0.0761	⋯	-474.12	27.83
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
O=C1C=C(C#C)C2NC12	3.97	80.61	-0.2410	-0.0803	⋯	-399.50	28.33

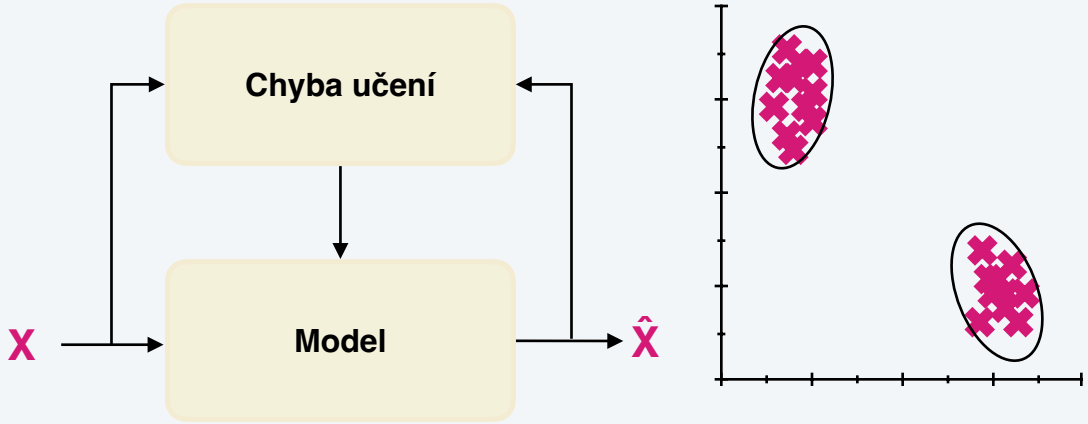
Základní Úlohy Strojového Učení

Učení s učitelem



X	Y			
SMILES	μ	α	⋯	Cv
OC1CCCC(=O)C1=O	3.56	69.15	⋯	30.94
CN1N=NC(F)=CC1=N	1.69	69.88	⋯	27.83
⋮	⋮	⋮	⋮	⋮
O=C1C=C(C#C)C2NC12	3.97	80.61	⋯	28.33

Učení bez učitele



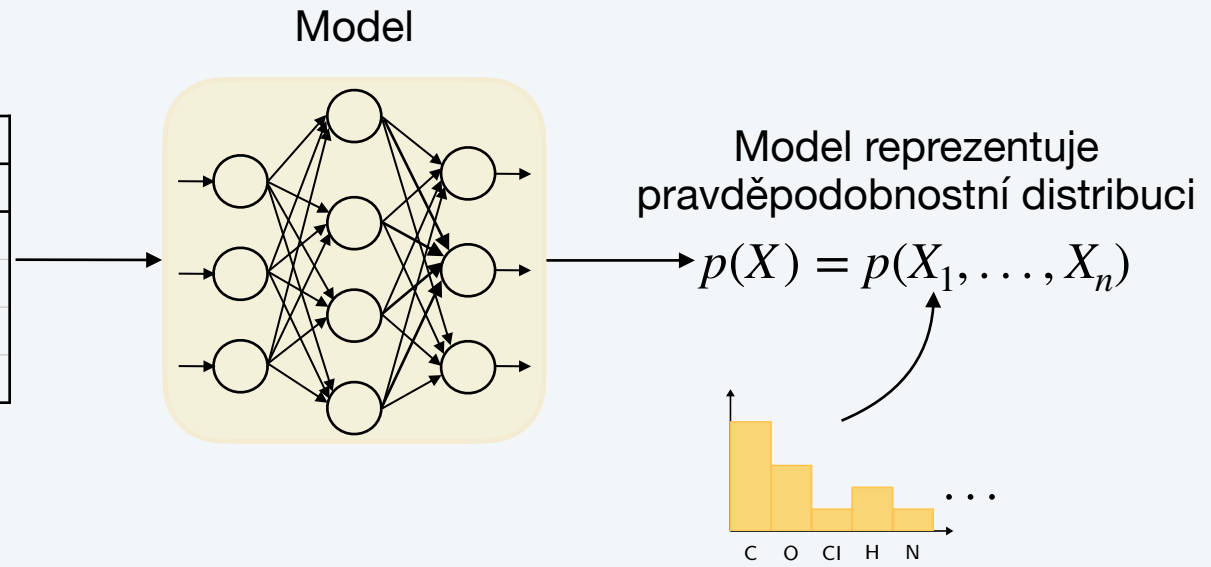
X
SMILES
OC1CCCC(=O)C1=O
CN1N=NC(F)=CC1=N
⋮
O=C1C=C(C#C)C2NC1

Hluboké Generativní Modely v Chemii

AI Center FEE CTU

Hluboké Generativní Modely

X				
SMILES	μ	α	⋯	Cv
OC1CCCC(=O)C1=O	3.56	69.15	⋯	30.94
CN1N=NC(F)=CC1=N	1.69	69.88	⋯	27.83
⋮	⋮	⋮	⋮	⋮
O=C1C=C(C#C)C2NC12	3.97	80.61	⋯	28.33

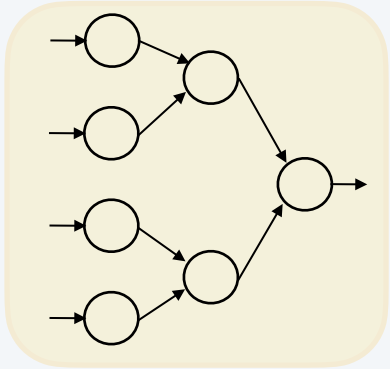


Generování nových dat je provedeno vzorkováním z pravděpodobnostní distribuce, $X \sim p(X)$.

Hluboké Generativní Modely

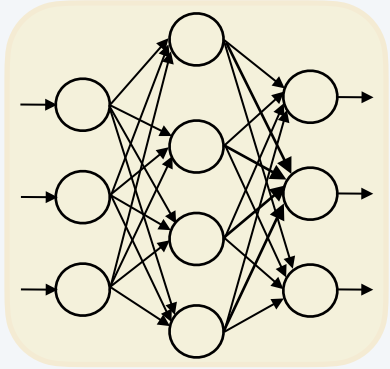
Traktabilní modely

- Výpočetně levné
- Řídká propojení
- Velký počet úloh (téměř všechny)

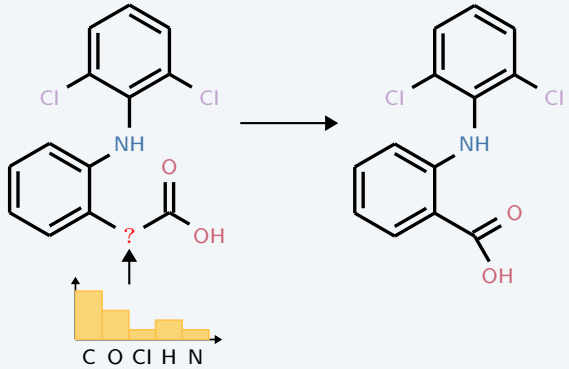


Netraktabilní modely

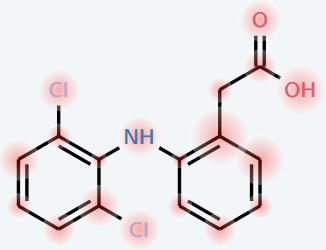
- Výpočetně drahé
- Hustá propojení
- Nízký počet úloh (jedna až dvě)



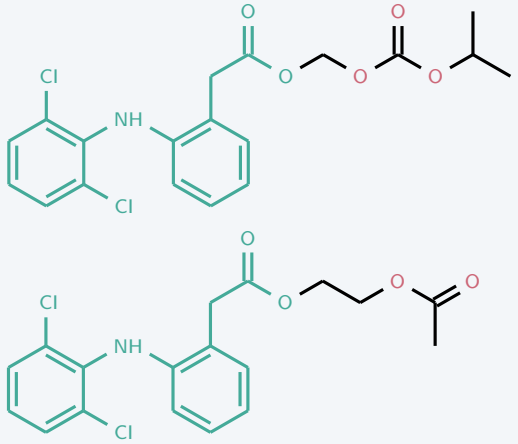
Typy úloh



Odstraň vybraný atom (marginalizace)

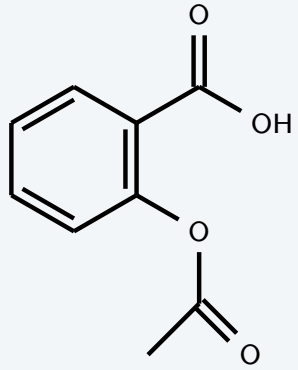


Vyjádření nejistoty (očekávaná hodnota)



Rozšiř existující molekulu (podmiňování)

Tokenizace SMILES Řetězců

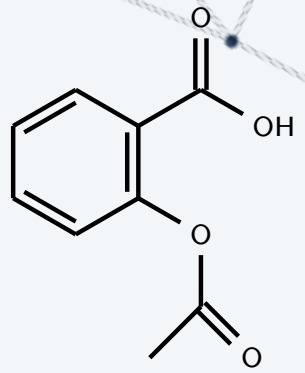


CC(=O)OC1=CC=CC=C1C(=O)O

Tokenizátor (encode)

9	5	5	0	…	8	10
BOS	C	C	(	…	O	EOS

Tokenizátor (decode)

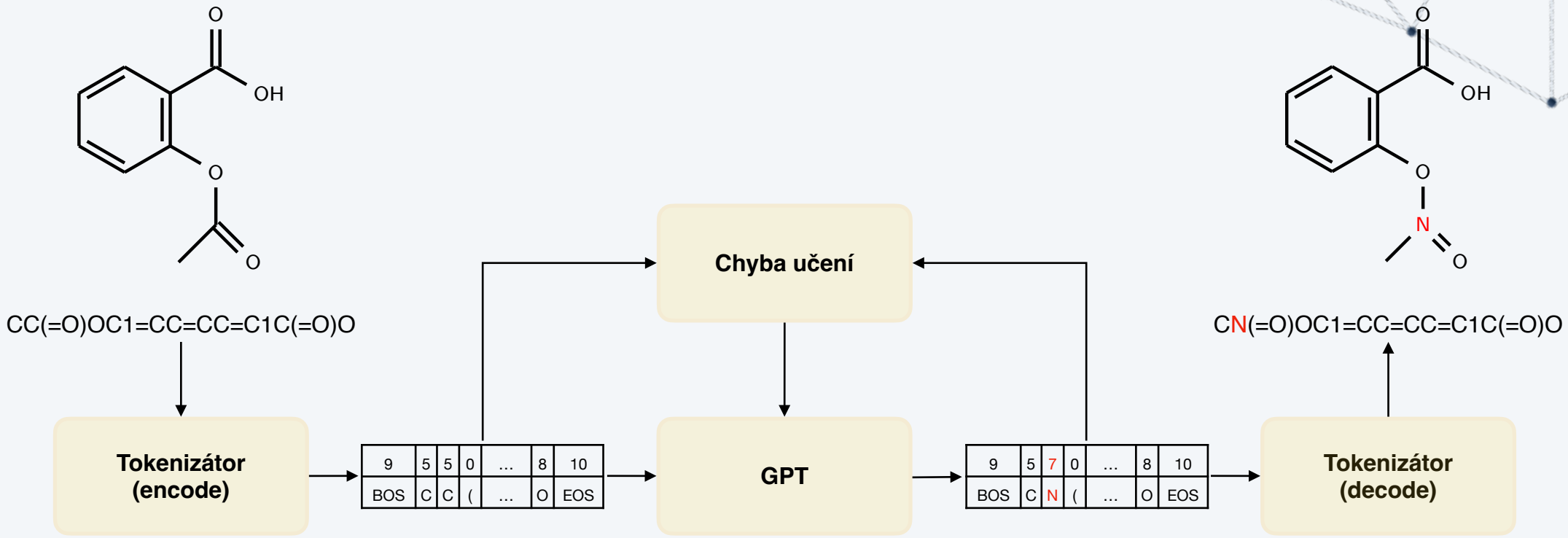


CC(=O)OC1=CC=CC=C1C(=O)O

Slovník

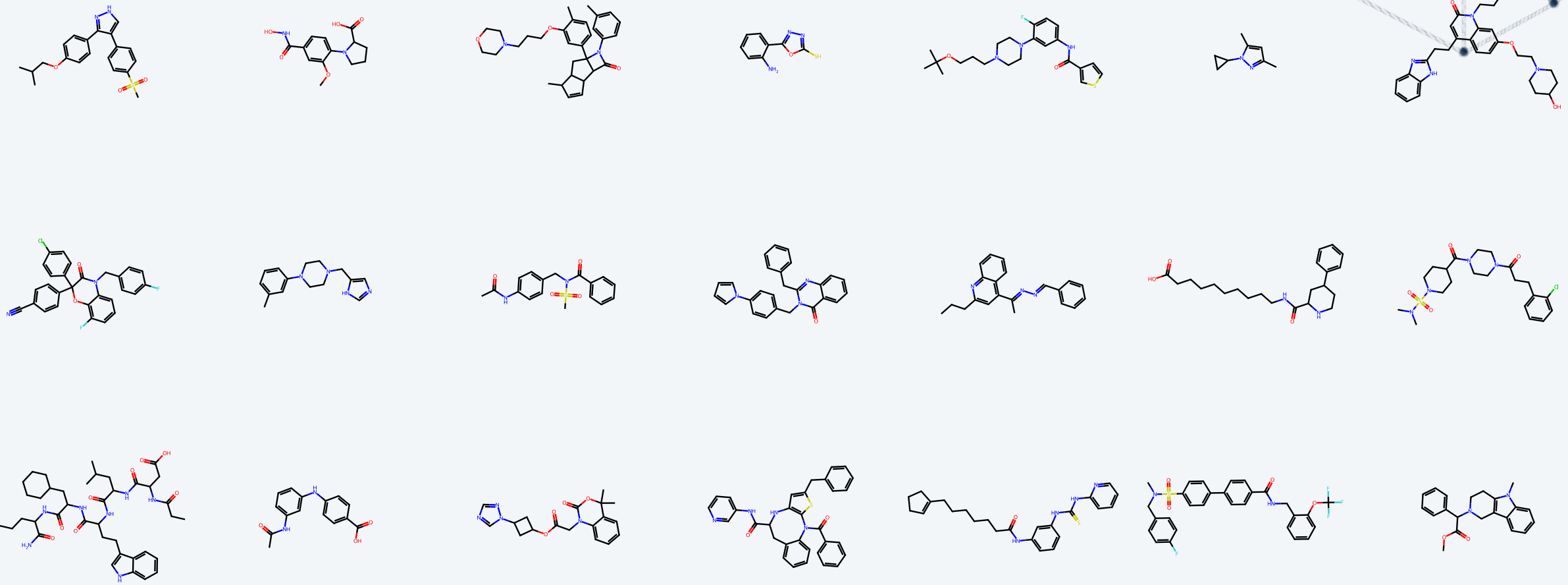
(	)	1	2	=	C	Cl	N	O	BOS	EOS	PAD	UNK
0	1	2	3	4	5	6	7	8	9	10	11	12

Chemické Jazykové Modely I Text



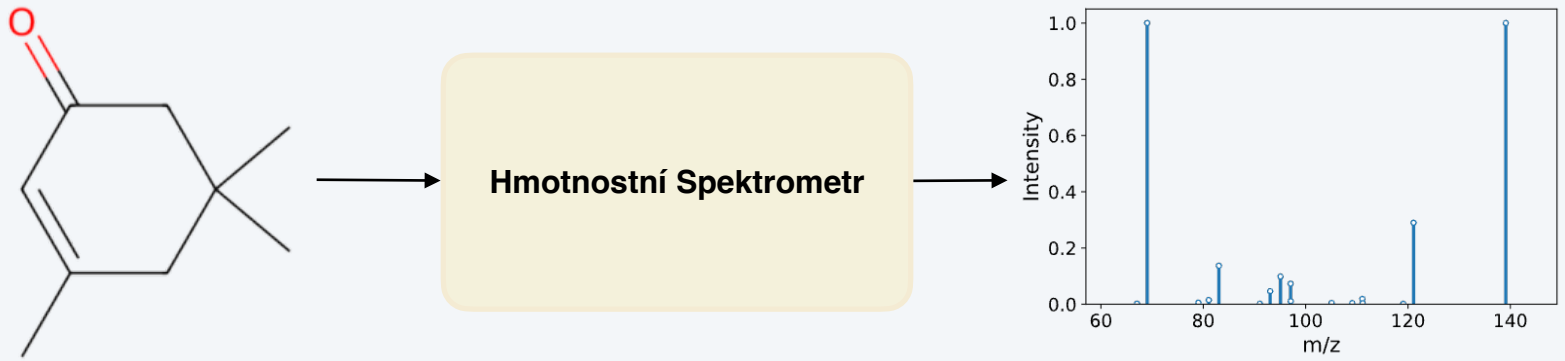
- Generativní předtrénovaný transformer (GPT)
- Strukturovaný stavový model (S4, Mamba)
- Difuzní model

Chemické Jazykové Modely I Text

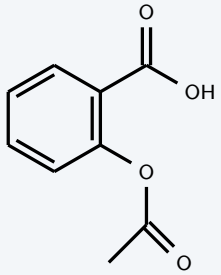


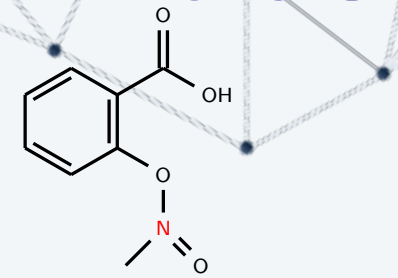
AI Center FEE CTU

Chemické Jazykové Modely I Text+Spektrum



Chemické Jazykové Modely I Text+Spektrum





CC(=O)OC1=CC=CC=C1C(=O)O

CN(=O)OC1=CC=CC=C1C(=O)O

Chyba učení (Text)

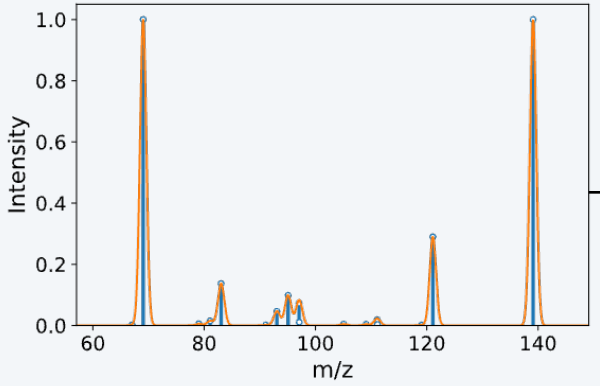
Tokenizátor (encode)

9	5	5	0	...	8	10
BOS	C	C	(	...	O	EOS

GPT

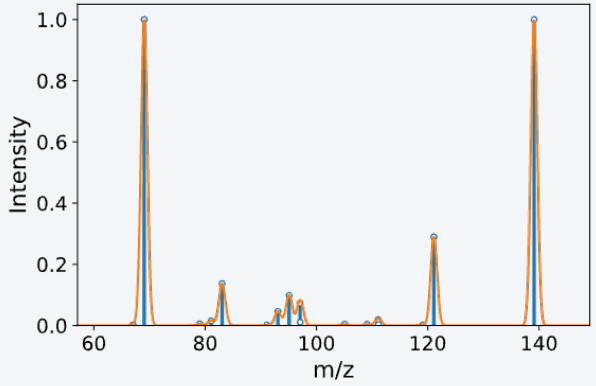
9	5	7	0	...	8	10
BOS	C	N	(	...	O	EOS

Tokenizátor (decode)



Neuronová síť

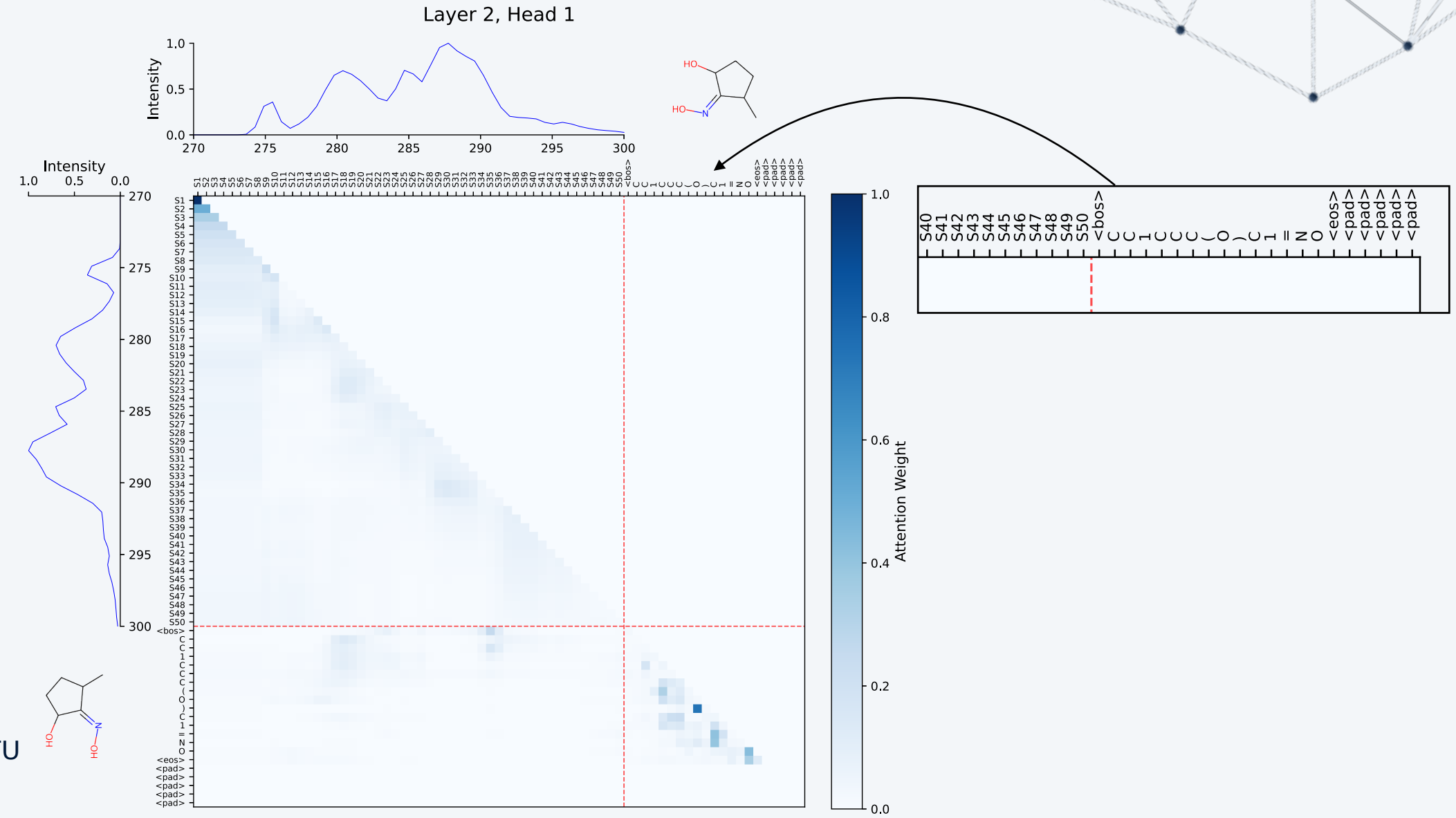
Neuronová síť



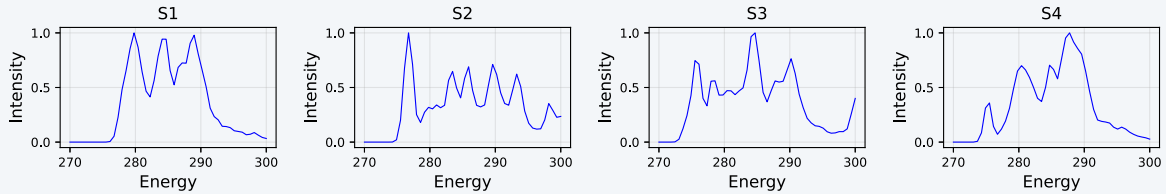
Chyba učení (Spektrum)

AI Center FEE CTU

Co se model učí?



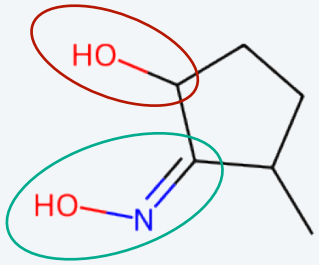
Generovaní Molekul Pomocí Spektra



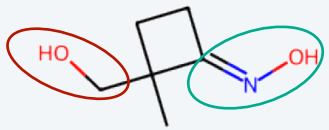
AI Center FEE CTU

Generovaní Molekul Pomocí Spektra

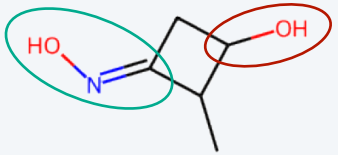
R4



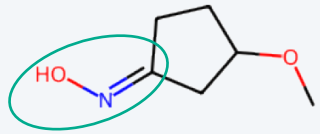
M1.4



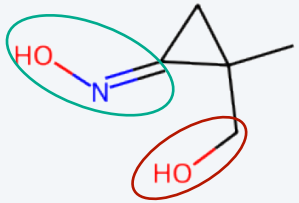
M2.4



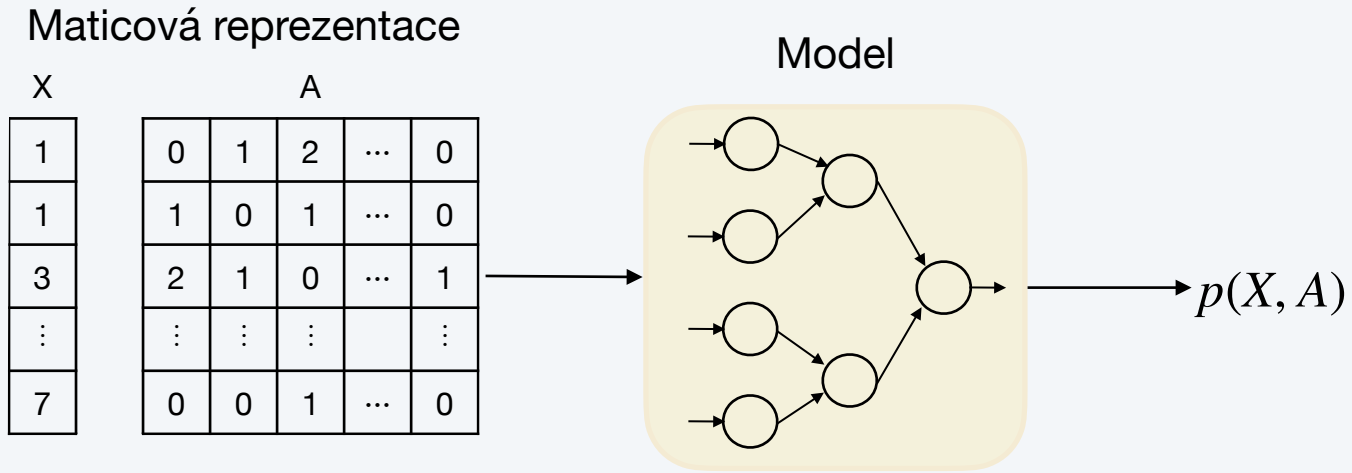
M3.4



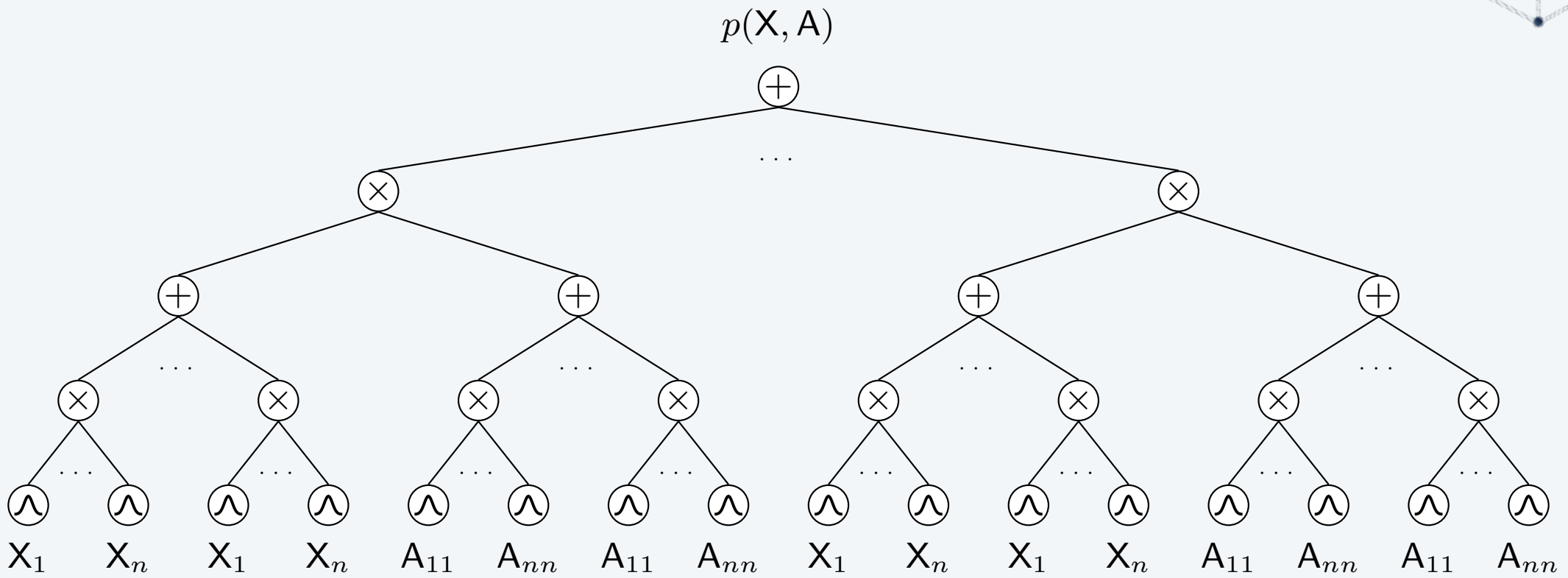
M4.4



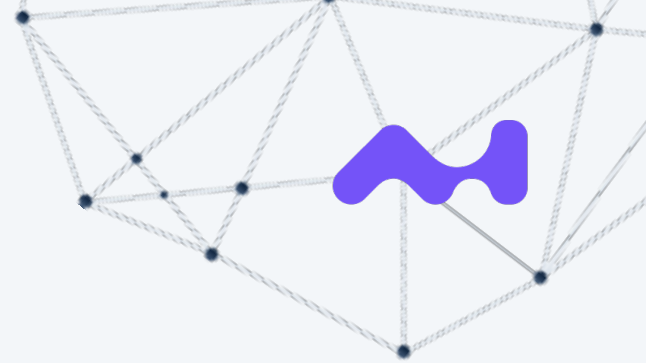
Pravděpodobnostní Grafové Obvody



Pravděpodobnostní Grafové Obvody



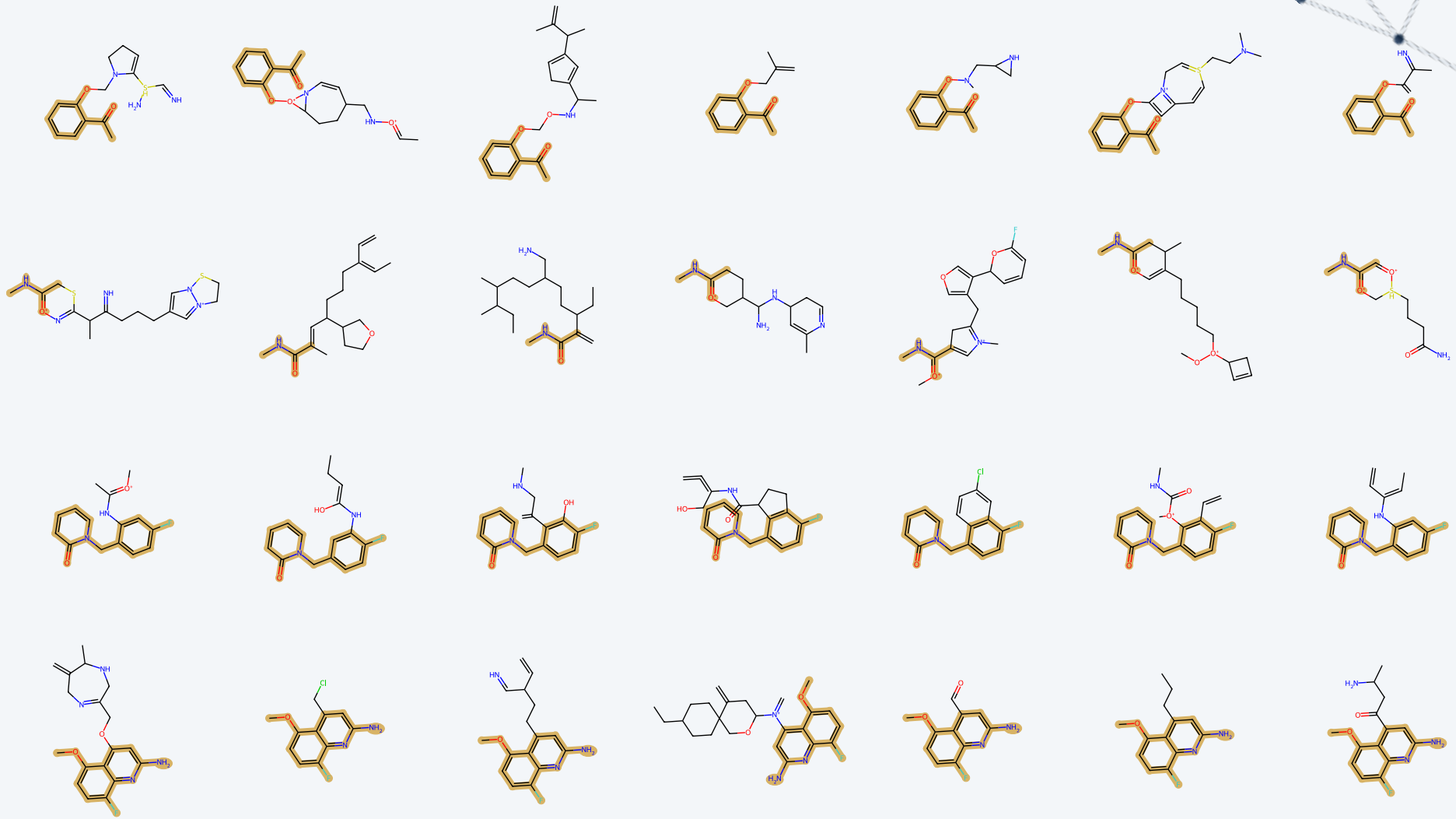
Pravděpodobnostní Grafové Obvody



	QM9					Zinc250k				
Model	Validita↑	NSPDK↓	FCD↓	Unikátnost↑	Novost↑	Validita↑	NSPDK↓	FCD↓	Unikátnost↑	Novost↑
GraphAF	74.43	0.021	5.27	88.64	86.59	68.47	0.044	16.02	98.64	100.00
GraphDF	**93.88**	0.064	10.93	98.58	**98.54**	**90.61**	0.177	33.55	99.63	99.99
MoFlow	91.36	0.017	4.47	98.65	94.72	63.11	0.046	20.93	99.99	100.00
EDP-GNN	47.52	0.005	2.68	99.25	86.58	82.97	0.049	16.74	99.79	100.00
GraphEBM	8.22	0.030	6.14	97.90	**97.01**	5.29	0.212	35.47	98.79	100.00
SPECTRE	87.30	0.163	47.96	35.70	**97.28**	90.20	0.109	18.44	67.05	100.00
GDSS	**95.72**	**0.003**	2.90	98.46	86.27	**97.01**	**0.019**	14.66	99.64	100.00
DiGress	**99.00**	0.005	**0.36**	96.20	33.40	**91.02**	0.082	23.06	81.23	100.00
GRAPHARM	90.25	**0.002**	**1.22**	95.62	70.39	88.23	0.055	16.26	99.46	100.00
BT	78.15	0.004	1.68	**99.66**	93.20	17.00	0.050	**9.42**	100.00	100.00
LT	62.81	0.007	2.57	**99.72**	95.95	4.11	0.056	11.53	100.00	100.00
RT	81.83	0.003	1.29	99.37	90.67	6.90	0.047	9.59	**100.00**	**100.00**
RT-S	88.83	**0.002**	**1.11**	99.38	88.49	14.66	**0.043**	**8.78**	**100.00**	**100.00**
HCLT	89.00	0.003	1.45	**99.45**	90.62	23.67	**0.035**	**8.93**	**100.00**	**100.00**

První místo **Druhé místo** **Třetí místo**

Podmíněné generování nových molekul



AI CENTER
FEE CTU

www.aic.fel.cvut.cz

papezmil@fel.cvut.cz

Centrum Umělé Inteligence

Fakulta Elektrotechniky

České Vysoké Učení Technické v Praze

18. 9. 2025