

USPOREDBA I OBJAŠNJIVOST MODELA STROJNOG UČENJA U PREDVIĐANJU NASTAJANJA FARMACEUTSKIH KOKRISTALA

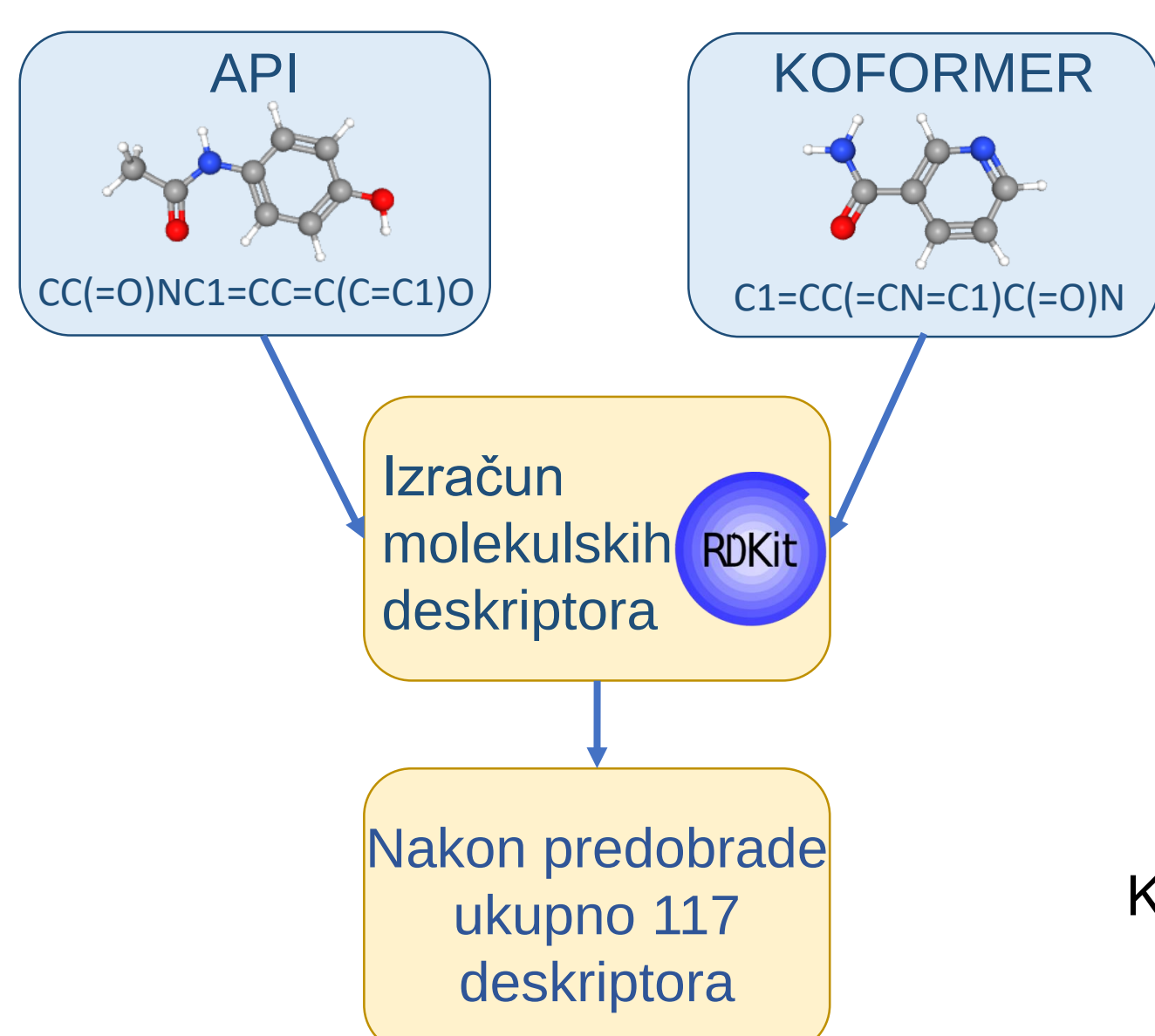
Filip Vretenar, Željka Ujević Andrijić

Sveučilište u Zagrebu Fakultet kemijskog inženjerstva i tehnologije

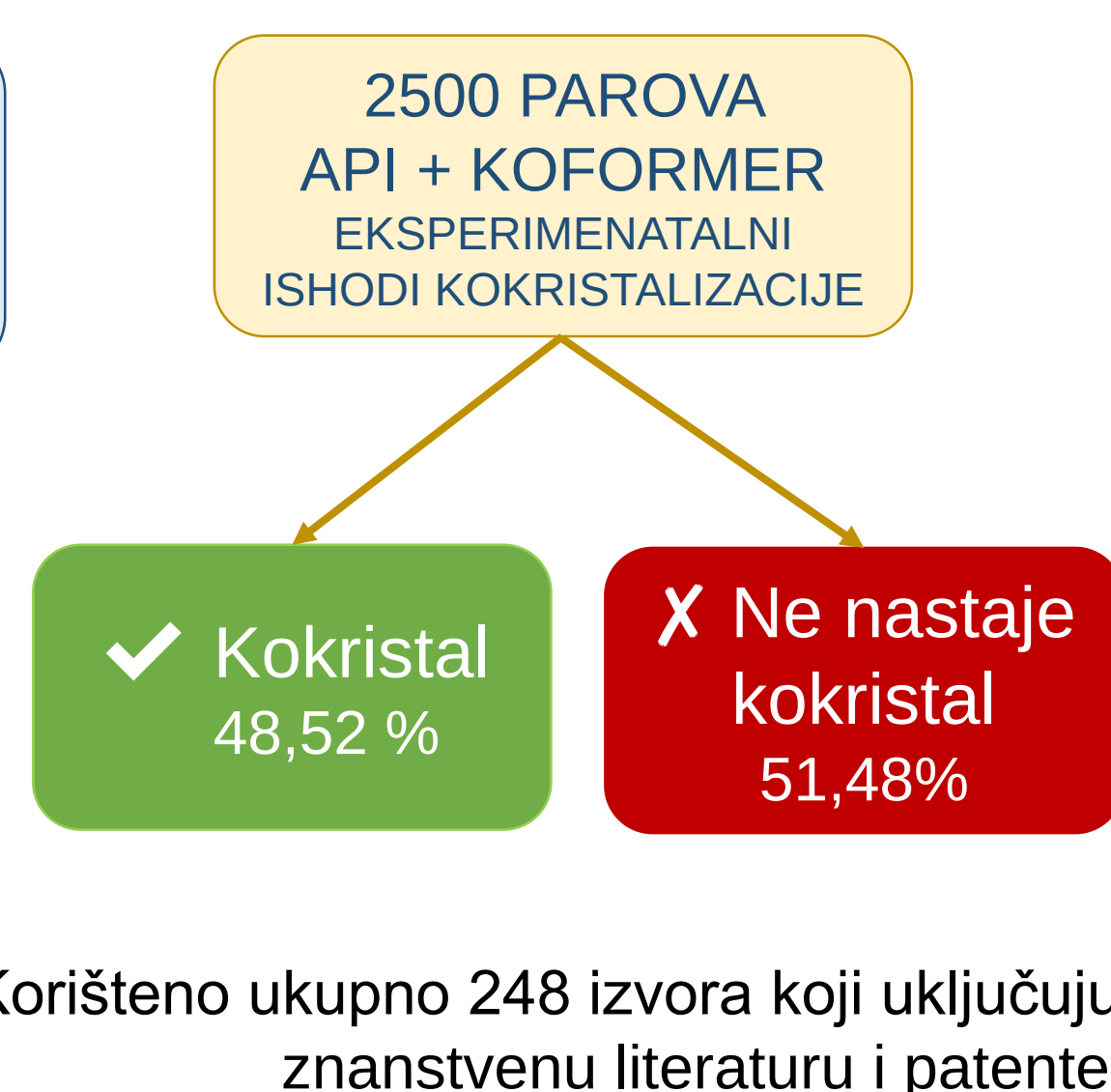
fvretenar@fkit.unizg.hr
zujevic@fkit.unizg.hr

Razvoj farmaceutskih kokristala važna je strategija za poboljšanje fizikalno-kemijskih svojstava farmaceutskih djelatnih tvari (engl. *active pharmaceutical ingredient*, API), ali zahtijeva dugotrajno ispitivanje velikog broja potencijalnih koformera. Taj se proces može ubrzati primjenom klasifikacijskih modela strojnog učenja. U ovom radu uspoređeni su rezultati dobiveni generativnim transformerskim modelom TabPFN, unaprijed naučenim na velikom broju sintetičkih podataka, s modelima gradijentnog pojačavanja stabala odlučivanja. Modeli su razvijeni koristeći molekulske deskriptore API-ja i koformera te eksperimentalne ishode kokristalizacije iz literature. Za interpretaciju predviđanja i određivanje važnosti pojedinih deskriptora primijenjena je SHAP (engl. *SHapley Additive exPlanations*) analiza.

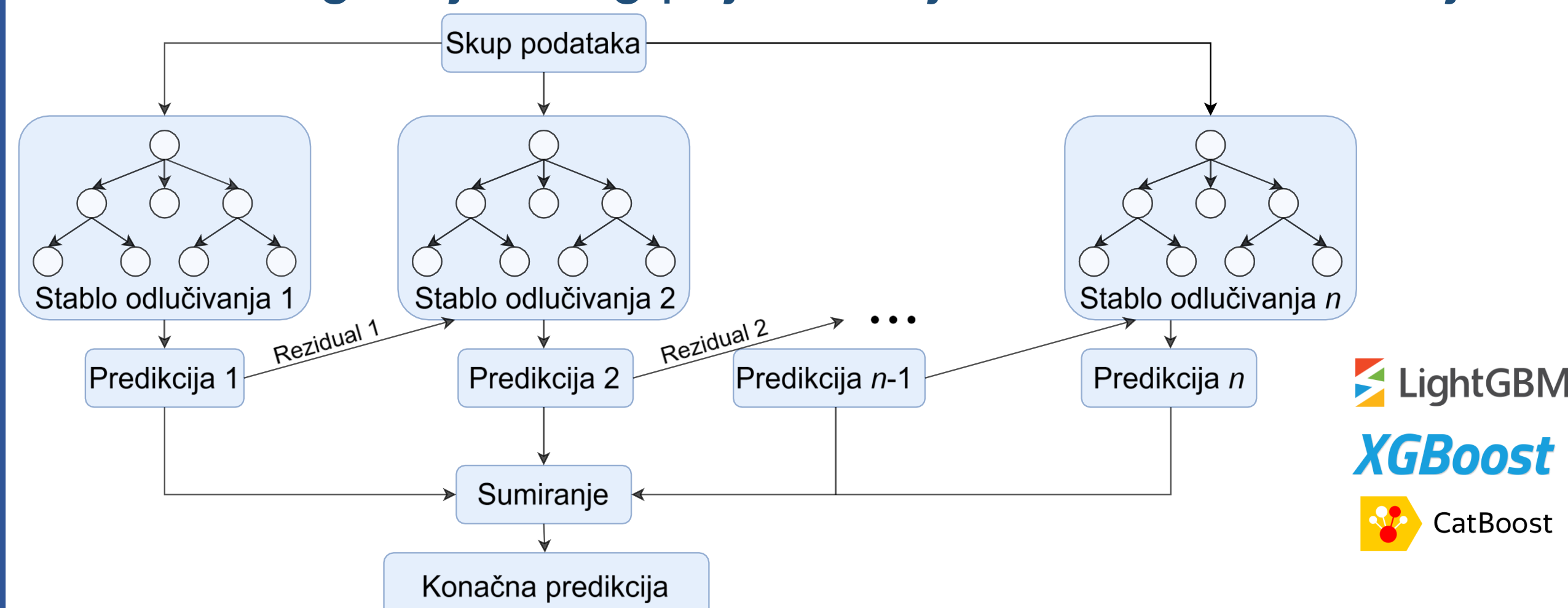
Molekulski deskriptori



Podaci za razvoj modela



Modeli gradijentnog pojačavanja stabala odlučivanja

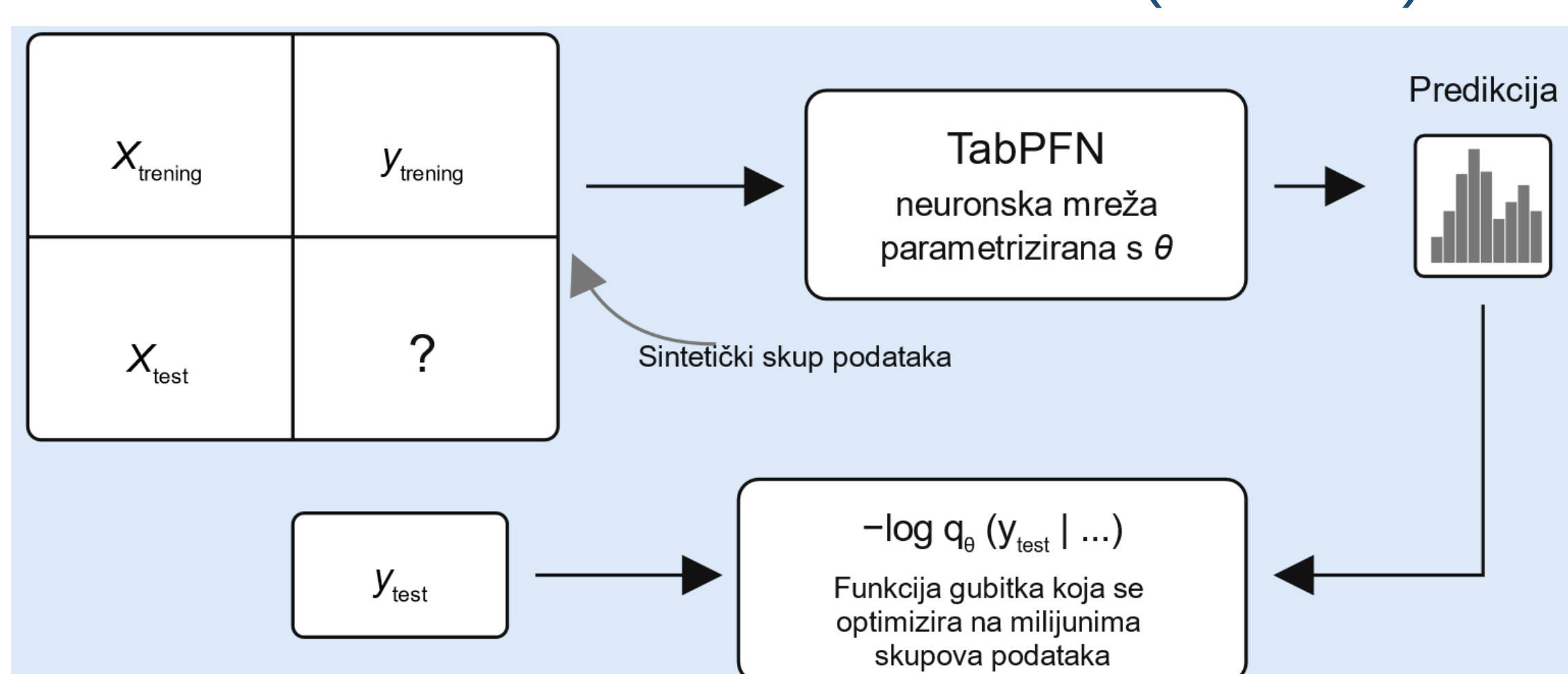


Posljednjih dvadesetak godina smatraju se vodećim modelima za tablične skupove podataka.

Rezultati modela na skupu za testiranje

Model	Točnost	ROC-AUC	F1-MJERA	Osjetljivost	Specifičnost
CatBoost	83,20 %	88,42 %	82,86 %	83,54 %	82,88 %
XGBoost	82,60 %	87,06 %	82,57 %	84,77 %	80,54 %
LightGBM	82,40 %	88,25 %	83,13 %	83,13 %	81,71 %

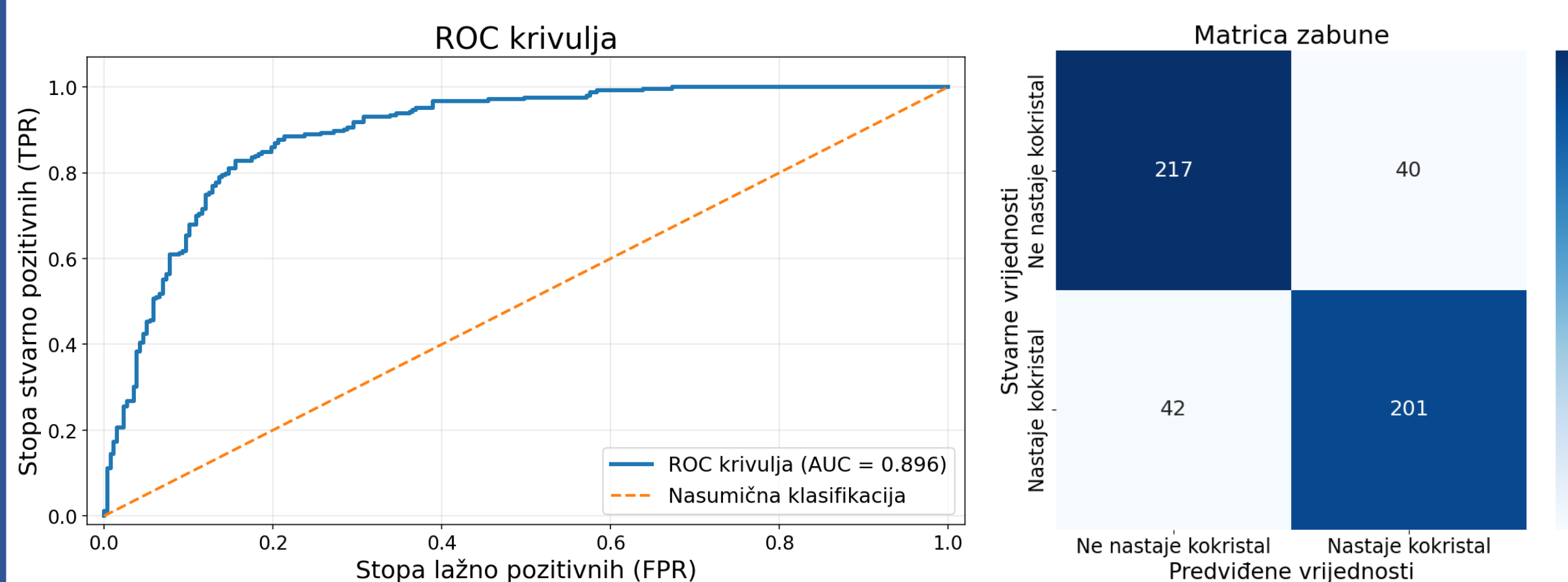
Tabular Prior-data Fitted Network (TabPFN)



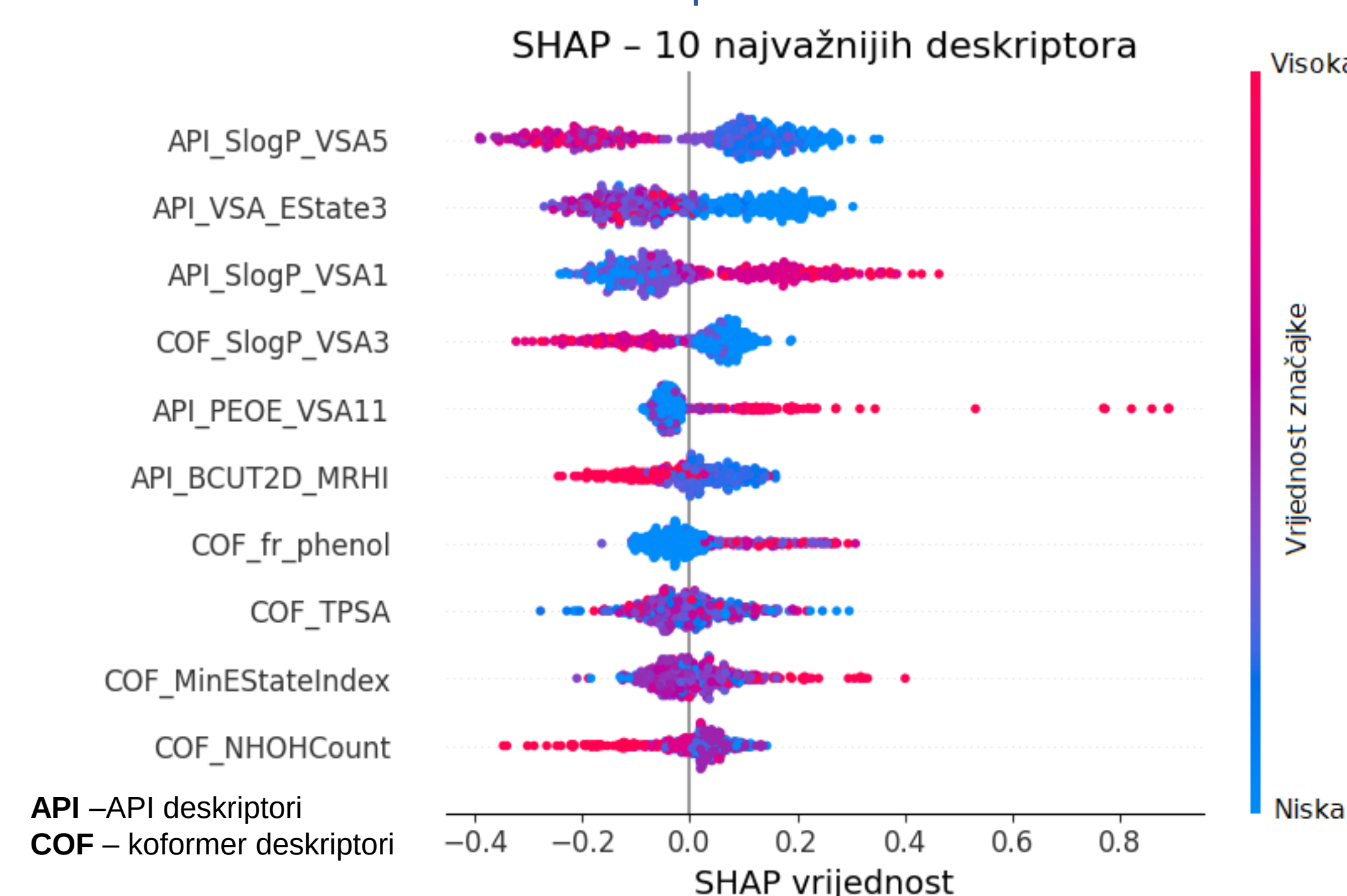
TabPFN model je unaprijed naučen na velikom broju sintetičkih tabličnih skupova podataka te primjenom *in-context learninga* stečeno znanje koristi za predviđanje na novim podacima bez dodatnog podešavanja parametara modela.

Rezultati TabPFN-a na skupu za testiranje

Model	Točnost	ROC-AUC	F1-MJERA	Osjetljivost	Specifičnost
TabPFN	83,60 %	89,60 %	83,06 %	82,72 %	84,44 %



SHAP beeswarm prikaz CatBoost model



Najučestaliji deskriptori među SHAP analizama

Deskriptor	Opis	Modeli
API_SlogP_VSA5	Σ vdW površina atoma s lipofilnošću (logP 0,10–0,15)	4/4
API_SlogP_VSA1	Σ vdW površina atoma s lipofilnošću (logP 0,25–0,30)	4/4
COF_MinEStateIndex	Najniža elektrotopološka vrijednost atoma	3/4
API_PEOE_VSA11	Σ vdW površina atoma s parcijalnim nabojem 0,15–0,20	3/4
COF_SlogP_VSA3	Σ vdW površina atoma s lipofilnošću (–0,20–0,00)	3/4

Modeli = broj modela u kojima je deskriptor među 10 najvažnijih prema SHAP analizi.

U ovom radu uspoređeni su modeli gradijentnog pojačavanja stabala odlučivanja i TabPFN za predviđanje nastajanja farmaceutskih kokristala. Ispitani modeli pokazali su slične i zadovoljavajuće rezultate na skupu za testiranje, pri čemu je najveću točnost ostvario TabPFN (83,60 %). SHAP analizom izdvojeni su važni molekularni deskriptori koji utječu na predviđanja modela, prvenstveno povezani s lipofilnošću i elektroničkim svojstvima molekula. Dobiveni rezultati ukazuju na mogućnost primjene objašnjivih modela strojnog učenja za učinkovitiji odabir potencijalnih koformera i brži razvoj farmaceutskih kokristala.