SYNTELLY
AI FOR ORGANIC CHEMISTRY

Prediction of toxicity endpoints as a pathway towards minimising risks in drug development

Conference participant: Dmitrii Shkil, MSc, Syntelly LLC

SYNTELLY
AI FOR ORGANIC CHEMISTRY

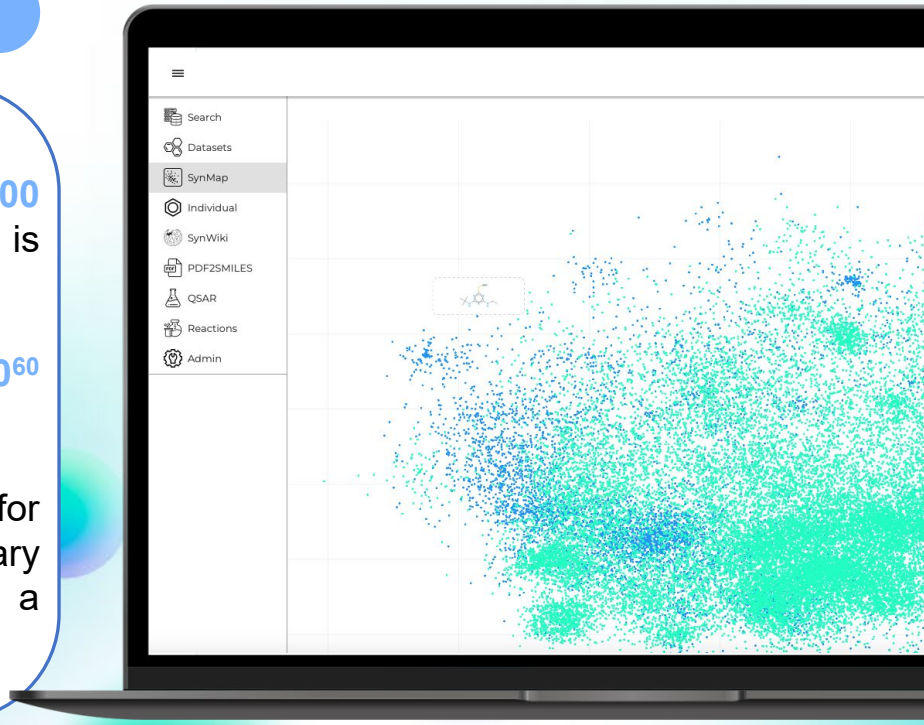
Scientific problems in drug discovery

CHEMICAL SPACE IS ENORMOUS

The scientific world is now aware of about **300 million** chemical compounds and their number is constantly growing.

Potential amount of chemical space estimated at 10^{60} molecules.

It is extremely difficult to manually search for chemical structures that have the necessary physicochemical or biological properties in such a huge array.



SynMap – example of chemical space map

SYNTELLY
AI FOR ORGANIC CHEMISTRY

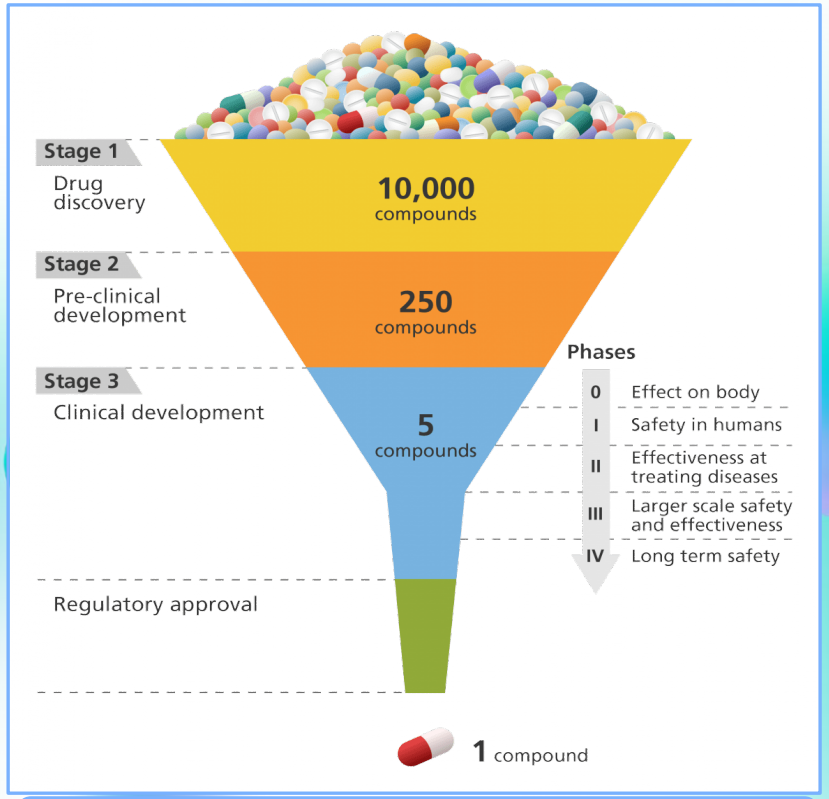
Scientific problems in drug discovery

COST OF TOXICITY INFORMATION FOR MOLECULES

In the global pharmaceutical market, the process from lead identification to clinical trials takes more than **12 years** and costs approximately **$1.8 billion USD** on average.

The need for research on toxicological properties is one of the main expenses of drug life cycle. Experiments use a large number of laboratory animals and can last for **several years**.

The limitations of *in vivo* and *in vitro* approaches to characterize chemical compounds have contributed to the development of *in silico* approaches.



From drug discovery to the market campaign

SYNTELLY
AI FOR ORGANIC CHEMISTRY

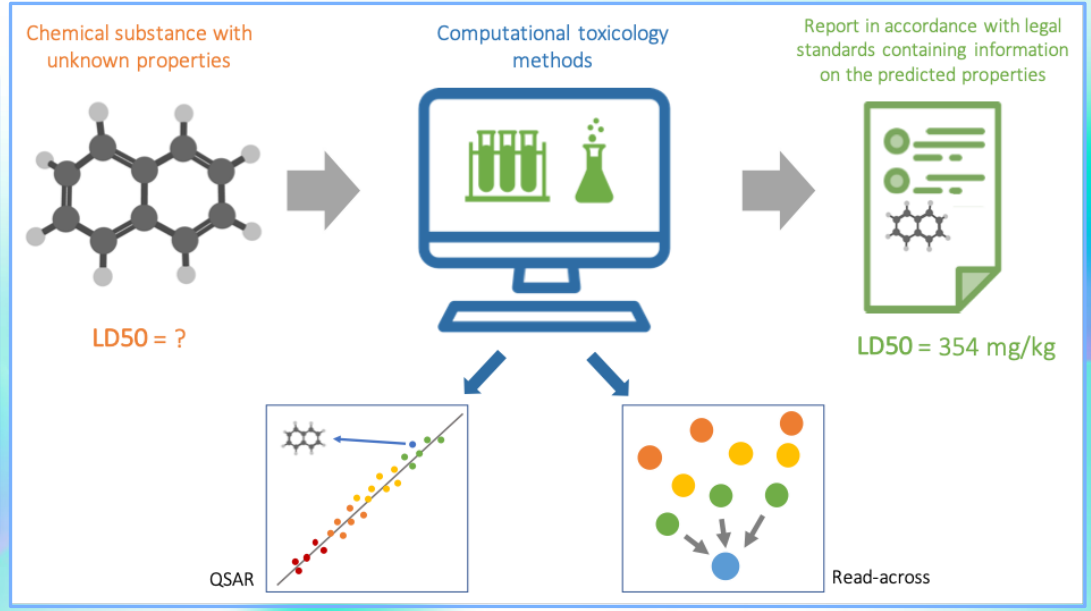
Scientific problems in drug discovery

NARROW APPLICABILITY DOMAIN OF EXISTING MODELS

Although more than **300 million** compounds are currently known, toxicity has only been studied for no more than approximately **300,000** compounds for **all endpoints**.

For the most part of studied compounds, molecular toxicity data are not aggregated, standardized, and well-curated.

Most of QSAR (quantitative structure-activity relationship) models have only **20-1000** samples in their training set, limiting its application for explored and unexplored chemical space.



Computational toxicology

SYNTELLY
AI FOR ORGANIC CHEMISTRY

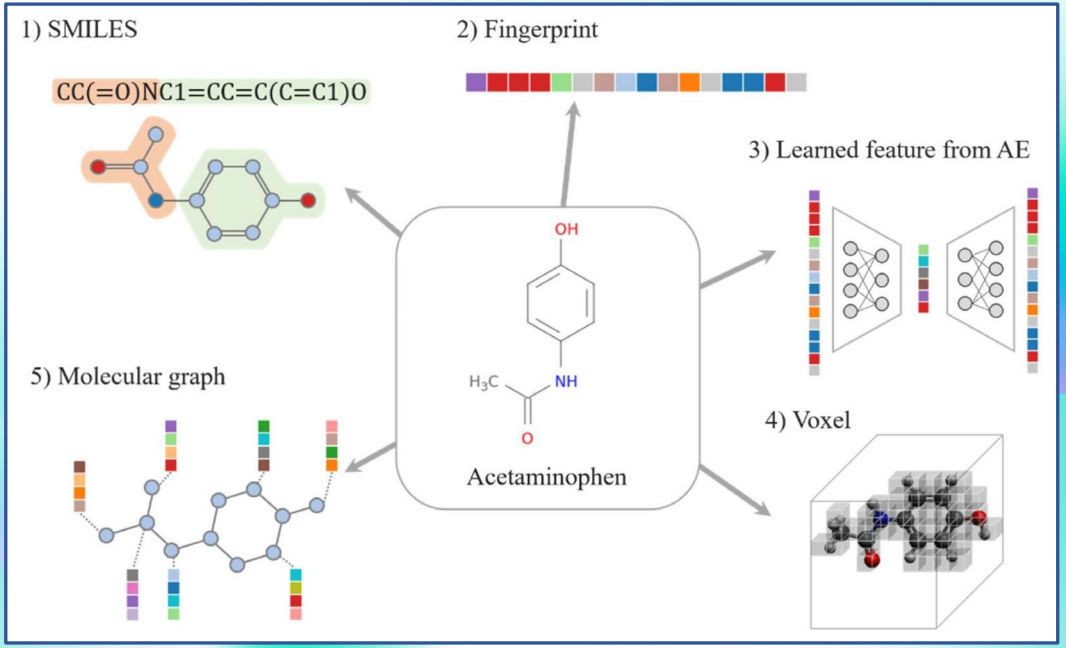
Hackathon

FROM THE STRUCTURE TO PREDICT TOXICITY ENDPOINTS

In Syntelly, we decided to organize the hackathon which was dedicated to the toxicity problem.

The task was as follows: based on open data, collect your own original dataset and build a machine learning model based on it to predict the toxicity of chemical compounds.

This was a non-standard hackathon, as it developed its own evaluation criteria and at the same time manual verification was carried out. Unlike competitions like Kaggle, our competition required attention not only to the models, but also to the data being collected, because without good quality data it is difficult to get a good model.



Vectorization of the structure

SYNTELLY
AI FOR ORGANIC CHEMISTRY

Solutions from the hackathon

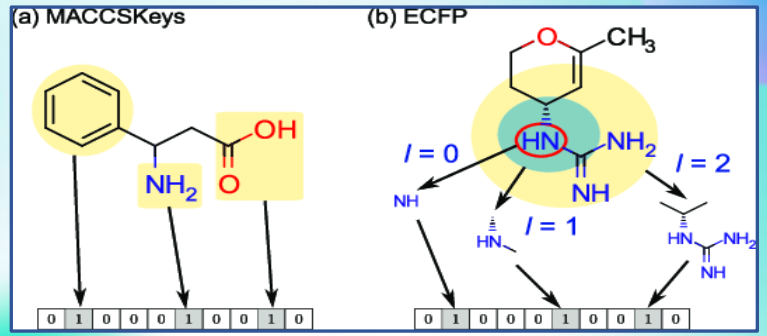
CATBOOST ON MORGAN ECFP or/and MACCS KEYS AND DESCRIPTORS

Name of target	Billy QSAR	Smi2Vec-LSTM	Smi2Vec-BiGRU	TranGRU
NR.AhR	**0.904**	0.678	0.879	0.833
NR.AR	0.771	0.691	0.714	**0.824**
NR.AR.LBD	0.747	0.748	0.824	**0.847**
NR.Aromatase	**0.802**	0.496	0.699	0.784
NR.ER	**0.787**	0.623	0.736	0.691
NR.ER.LBD	0.763	0.531	**0.868**	0.843
NR.PPAR.gamma	0.767	0.566	0.749	**0.838**
SR.ARE	**0.795**	0.641	0.761	0.701
SR.ATAD5	**0.806**	0.5	0.763	0.727
SR.HSE	**0.796**	0.612	0.785	0.736
SR.MMP	**0.951**	0.743	0.86	0.816
SR.p53	**0.818**	0.518	0.732	0.81

ROC AUC avg. values on 5-CV for Tox21 dataset

Name of target	MML	Syntelly
Mouse Oral LD_{50}	**0.43**	0.49
Rat Oral LD_{50}	**0.47**	0.68
Mouse Intraperitoneal LD_{50}	**0.45**	0.54
Rat Intraperitoneal LD_{50}	**0.58**	0.63
Mouse Intravenous LD_{50}	**0.46**	0.52
Rat Intravenous LD_{50}	**0.59**	0.63

RMSE values for regression task



Examples of chemical fingerprints

SYNTELLY
AI FOR ORGANIC CHEMISTRY

Our solution

Target name	CatBoost FPs	XGBoost Fragments	Benchmark	N samples
Cardiotoxicity (hERG binary)	0.888	0.926	**0.930** (CardioTox)	324010
Ames test	0.845	**0.894**	0.88 (Syntelly)	14168
SR-HSE	**0.839**	0.836	0.736 (TranGRU)	7281
NR-AR	0.724	0.797	**0.824** (TranGRU)	7263
NR-AR-LBD	0.843	0.835	**0.847** (TranGRU)	7133
NR-PPAR-gamma	0.727	0.809	**0.838** (TranGRU)	6942
NR-Aromatase	0.831	**0.875**	0.784 (TranGRU)	6929
NR-ER-LBD	0.858	**0.878**	0.843 (TranGRU)	6920
SR-ATAD5	0.75	**0.862**	0.727 (TranGRU)	6893
SR-ARE	0.825	**0.845**	0.701 (TranGRU)	6822
SR-p53	0.748	**0.877**	0.81 (TranGRU)	6749
NR-ER	0.852	**0.866**	0.691 (TranGRU)	6585
NR-AhR	0.816	**0.871**	0.833 (TranGRU)	6446
SR-MMP	0.853	**0.896**	0.816 (TranGRU)	6361

ROC AUC avg. values on 5-CV for datasets from the hackathon for binary classification task

SYNTELLY
AI FOR ORGANIC CHEMISTRY

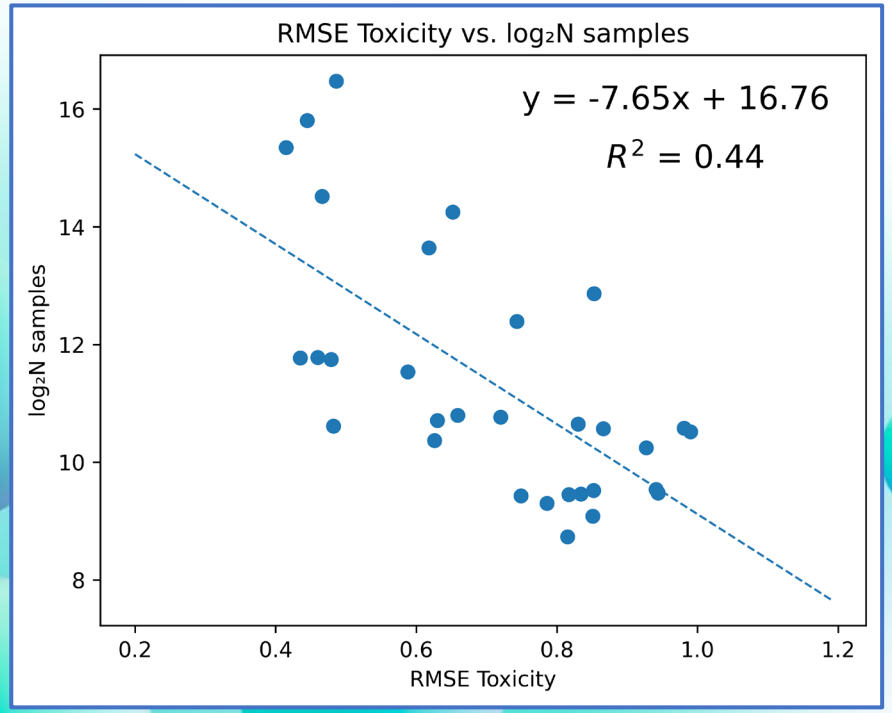
Our solution

Target name	CatBoost FPs	XGBoost Fragments	Benchmark	N samples
Mouse Intraperitoneal LD_{50}	0.562	**0.486**	0.54 (Syntelly)	91162
Mouse Oral LD_{50}	0.543	**0.445**	0.49(Syntelly)	57307
Mouse Intravenous LD_{50}	0.498	**0.415**	0.59 (Syntelly)	41630
Rat Oral LD_{50}	0.589	**0.466**	0.68 (Syntelly)	23409
Mouse Subcutaneous LD_{50}	0.696	0.652	**0.55** (Syntelly)	19457
Rat Intraperitoneal LD_{50}	0.71	**0.618**	0.63 (Syntelly)	12769
Rat Subcutaneous LD_{50}	0.829	**0.743**	0.69 (Syntelly)	5376
Tetrahymena pyriformis IGC_{50} 40 h	0.524	**0.46**	0.518 (TOXRIC)	3516
Mouse Intraperitoneal LD_{Lo}	0.495	**0.435**	0.52 (Syntelly)	3500
Rabbit Skin LD_{50}	0.521	**0.479**	0.58 (Syntelly)	3429
Rabbit Oral LD_{50}	0.626	**0.588**	**0.588** (Syntelly)	2969
Guinea Pig Oral LD_{50}	0.703	**0.659**	0.69 (Syntelly)	1778
Fathead Minnow LC_{50} 96 h	0.78	**0.72**	0.864 (TOXRIC)	1739
Rat Skin LD_{50}	0.665	0.63	**0.62** (Syntelly)	1673
Rat Intraperitoneal LD_{Lo}	0.512	**0.482**	0.63 (Syntelly)	1568

RMSE avg. values on 5-CV for datasets from the hackathon for regression task

Final part of the work

INCREASE OF EXPERIMENTAL DATA IS THE KEY TO HIGH PERFORMANCE



We hypothesized that data quantity is a significant factor in the quality of QSAR models. On the one hand, this statement is almost axiomatic within the framework of machine learning - the more data, the better quality can be achieved. On the other hand, this issue has not been widely addressed within machine learning in chemistry.

We built a linear regression model of the dependence of RMSE on the number of samples in the datasets and saw that there is a significant relationship between these variables. It is worth noting that the datasets are aggregated and the trend is observed at the level of all toxicity regression endpoints. Thus, we believe that we would like to inspire experimenters to obtain more experimental data with high statistical reliability of the data.

Thank you for attention!

If you want a token, please write to admin@syntelly.com