Which modern AI methods provide accurate predictions of toxicological endpoints? Analysis of Tox24 challenge results.

Stephanie A. Eytcheson[1,2] and Igor V. Tetko[3,4,*]

[1]Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee 37830, United States; [2]Center for Computational Toxicology and Exposure, Great Lakes Toxicology and Ecology Division, U.S. Environmental Protection Agency, Office of Research and Development, Duluth, Minnesota 55804, United States; [3]Institute of Structural Biology, Molecular Targets and Therapeutics Center, Helmholtz Munich - Deutsches Forschungszentrum Für Gesundheit Und Umwelt (GmbH), 86764 Neuherberg, Germany; [4] BIGCHEM GmbH, Valerystr. 49, 85716 Unterschleißheim, Germany

The Tox24 challenge[1] was designed to evaluate the progress that has been made in computational method development for the prediction of *in vitro* activity since the Tox21 challenge, which was organised by National Institutes of Health National Center for Advancing Translational Sciences (NCATS).[2] In this challenge, participants were tasked with developing models to predict chemical binding to transthyretin (TTR), a serum binding protein, based on chemical structure. The chemicals tested for activity against TTR[3], which were a subset of the U.S. EPA's contribution to the multi-agency "Toxicology in the 21st Century" (Tox21)" initiative[4] were used as training and test sets for the challenge. The challenge results were announced during the last day of the ICANN2024 conference https://e-nns.org/icann2024, which took place in September 2024 in Lugano.

TTR is one of the serum binding proteins responsible for transport of thyroid hormones (TH) to target tissues. TTR also plays a role in maintaining the balance of free versus bound (*i.e.*, available vs. inactive) TH.[5,6] As such, TTR may be important for supporting TH homeostasis and proper functioning of the thyroid system. Endocrine disrupting chemicals (EDCs) may trigger adverse health effects by disrupting the endogenous hormone system; thus, the identification of such chemicals is a critical target for toxicology. The analysed dataset included chemicals that were screened in a competitive binding assay designed to measure the reduction in fluorescence due to displacement of 8-anilino-1-naphthalenesulfonic acid ammonium salt (ANSA) from TTR.

Challenge data

1813 unique chemicals selected from the ToxCast ph1_v2, ph2, and e1k libraries were screened through the TTR binding assay at a target concentration of 100 μM. Chemical activity was calculated against a thyroxine (T4) standard curve where the high concentration of T4 represented 100% activity (ANSA completely displaced from TTR) and low concentration T4 represented 0% activity (ANSA not displaced from TTR). Due to assay interference, autofluorescent chemicals were excluded leaving a total number of 1512 compounds kept for analysis. All data used for measurements were reported with only the chemical names. The structural information (Simplified Molecular Input Line Entry System, SMILES) for data was originally retrieved from U.S. EPA's CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard).[4] Several compounds (*e.g.*, oils and mixtures) in the original set did not have SMILES in the dashboard, and so, we retrieved missing structural information for them from PubChem based on CAS RN and names. These compounds were randomly split into a training set of 1012, a leaderboard set of 200 and a blind set of 300. Following suggestions from the challenge participants, who noticed discrepancies between SMILES, CAS RN and names for several molecules, we further reviewed the data by manually checking conflictual cases against CAS https://commonchemistry.cas.org and prioritised CAS suggestions for SMILES. In cases when CAS suggestions were ambiguous, we used structures retrieved from PubChem. The structural data were updated on 21.06.2024 on the challenge website and the participants were also asked to further validate the data.

Because multiple testing could, in principle, help the participants to identify the activity values of (some) compounds, we decided to release the leaderboard set 15 days before the challenge ended. This allowed participants to have an earlier access to these data that could be used to enhance their models.

Challenge setup

The challenge was hosted on the On-line CHEmical database and Modelling environment (OCHEM)[7] website https://ochem.eu. The users had the option to register and upload their predictions themselves or submit them directly to the challenge organisers via email. Only one team submitted their results via email.

Not including the organisers, a total of 78 teams took part in the competition. Based on the registration information 45, 13, 11 and 9 teams identified themselves as academic, commercial, non-profit - governmental and self-employed, respectively. The challenge attracted participants from all over the world, representing 27 countries. The largest numbers of teams were from the USA (15) and Germany (8) followed by Russia and Poland (6 each). Since country information was not required, several teams did not indicate it. In total, there were 1374 submissions by all teams.

Challenge rules

Each participant could only belong to one team. Each team was required to provide information on its composition before the challenge ended to ensure that this rule was upheld. The teams were encouraged to use supplementary data taken from experimental measurements from similar assays in order to enhance their models, *e.g.*, transfer learning or multitask approaches. However, the use of supplementary data from exactly the same assay type and for the same endpoint as was used for model development within the challenge was not allowed.

The Root Mean Squared Error (RMSE) for the blind test set, whose experimental values were kept secret before the challenge ended, was used for the final scoring of the participating teams. The RMSEs were rounded to one decimal digit and in case of equal scores, the team which submitted their final result earlier had a higher position in the final ranking. In addition to the winner of the challenge, we identified ten teams that had non-significantly different scores to the winning model using paired samples *t-test* (see Table 1), whom we collectively call the "group winners". All of these teams were also asked to send a short description of their methods to confirm adherence to the challenge guidelines on the use of data, reproducibility of models and team composition. These reports were used to prepare a summary of models in Table 1.

The participating teams could test their models' performance on the leaderboard set, which had two purposes: to guide the models' development and to verify that submissions were provided in the correct format. It should be noted that once the leaderboard set was publicly released (on 15.08.24), the performance on this set no longer reflected the prediction accuracy of the models because the set essentially becomes training data. Indeed, after its release, some participants submitted predictions for this set and obtained an RMSE=0 (as was the case for team #3, for example). Of course, the leaderboard set could be still used for model selection only, and team #7 (and potentially others) did in fact use it in this way. However, this was not obligatory and each team could rely on their own validation protocols to select the best model for their final submission.

Analysis of submissions before the release of the leaderboard set on 15.08

The RMSEs for both leaderboard and blind sets were strongly correlated with the Pearson correlation coefficient R = 0.97 and thus RMSEs on the leaderboard set were strongly predictive of the performance of the methods for the blind test set. RMSE values for the leaderboard set submitted before its public release provided good estimations of the errors associated with models' performances on the blind test set. There were in total 866 submissions by 49 teams before the leaderboard set was made publicly available.

Amid all these 866 submissions the lowest RMSE = 20.9 for the blind test set before release of the leaderboard set was achieved by team #2 (Table 1), which also corresponded to the lowest RMSE = 20.6 achieved by this team for the leaderboard set at that time. Hypothetically, if the challenge had ended on the 15th of August and all teams had submitted their models based on the best leaderboard performance, this team would have won the challenge. Under the same conditions, the second place would have gone to #9 with blind set RMSE=21.2, and #3 would have placed fourth with RMSE=21.3 (after team scottkdh, who submitted results with the same RMSE earlier than #3, however scottkdh was not among the "group winners" based on their final submission, therefore their results are not listed in Table 1).

Table 1. The "group winners" of the challenge

n	team: members	RMSE					Short description of several methods as provided by the authors (see Supplementary Information for several descriptions as provided by participants)	Publication reference
		Final blind test	Lowest	submissions before 15th August				
				Leaderboard	Blind test	hypothetic place[1]		
1	Amidoff: D. Makarov, A. Ksenofontov	20.5	20.5	20.5	21.4	5	Consensus of four models: CatBoost[8] method using Mold2[9] and ALOGPS[10,11] plus E-state indices descriptors[12,13] and two representation learning methods: Transformer CNN[14] and CNF2.[15] The authors analysed compounds as mixtures. Model is available as https://ochem.eu/article/162082	
2	tcirino: T. Cirino	20.7	20.7	20.6	20.9	1	Tautomer generation for molecules with multiple tautomeric forms was done. A consensus model was based on representation learning models, including Transformer CNF2 and CNN as well Graph Neural Network (GNN) models, AttFP, ChemProp and GIN provided by Keras Graph Convolution Neural Networks (KGCNN)[16] as implemented in OCHEM,[17] which were combined with CatBoost models based on EPA, Mold2, and 2D descriptors calculated with MORDRED, PaDEL2 and RDKIT.	
3	znavoyan: Z. Navoyan, A. Tevosyan, H. Yeghiazaryan, L. Khondkaryan, N. Abelyan, V. Atoyan, N. Babayan	20.7[2]	20.4	21.8	21.3	4	RF models were developed on bioassay data from ToxCast[4] and eMolTox[18] which formed a set of Bioassay descriptors. The final model was based on a consensus of RF models developed with a combination of bioassay descriptors with RDKit, E-state and Graph Neural Networks (GNN). GNN was used based on Principal Neighborhood Aggregation (PNA).[19]	
4	Microsomes: Y. Uesawa, Y. Iwashita, K. Kimura, T. Komasaka, K. Shishido, M. Asada	20.8	20.8	19.9	21.4	6	The model was developed by stacking prediction results from XGBoost, RF,[20] CatBoost, and LightGBM, using Mordred,[21] E-state, pH-dependent descriptors, as well as diverse nuclear receptor- and stress response pathway-related activity predictions generated by Toxicity Predictor[22] as explanatory variables. The robustness of the final model was ensured by nested cross-validation.	
5	YingkaiZhangLab X. Pan, Y. Zhang, Y. Gu, W.J. Zhou	20.8	20.5	-[3]	-	-	The initial dataset was extended with 215 TTR bioactivity endpoints collected from public databases (EquiVS[23] and Papyrus[24]). Structures of molecules were optimised. Consensus of three models based on sPhysNet-MT,[25] which used 3D descriptors based on	

							radial basis functions (RBFs), KANO,[26] which utilizes a chemical element-oriented knowledge graph, and GGAP-CPI (protein Graph and ligand Graph network with Attention Pooling for Compound-Protein Interaction prediction) which combines both 2D ligand molecular graphs and 3D protein structure graphs, were used.	
6	AntonijaBoss: A. Kraljevic, B. Lučić	21.2	21.2	-	-	-	Consensus of an RF model based on ALOGPS, E-state indices and CDK[27] descriptors and Transformer CNF2 model.	
7	SankalpJain: S. Jain, A. Zakharov	21.3	21.3	-	-	-	Deep Learning Consensus Architecture (DLCA),[28] which is a consensus model based on Deep Neural Networks (DNN) using Morgan, Avalon, and AtomPair along with RDKit[29] physicochemical descriptors combined with Convolutional Neural Network (CNN) based on SMILES.	
8	alx.dga: A. Dougha	21.4	21.4	21.5	22.0	13	Mixture of experts based on a Tanimoto similarity of tested compound to the training set. Expert models included Support vector Machines (SVM)[30], RF, Knowledge-guided Pre-training of Graph Transformer (KPGT)[31] and ChemProp.[32]	
9	luispintoc: L. Pinto	21.4	20.4	21.0	21.2	2	Consensus of four Foundation Chemistry Models fine-tuned: two MolFormer[33] models (which were also pre-trained on 50k compounds from ZINC database), SMI-TED,[34] and UniMol.[35]	
10	GAT_Wang: H. Wang, W. Liu, J. Chen	21.4	21.4	-	-	-	Data for 2403 chemicals against thyroid hormone receptor β (TRβ) were collected from PubMed. They were used for transfer learning based on graph attention network (GAT) architecture.	
11	vchupakhin: V. Chupakin	21.4	21.4	-	-	-	A feature set (AUX) of 180 descriptors was used. These were predictions of models developed for proteins that bind compounds that are structurally similar to thyroxine and retinol. Data were collected from PubChem and ChEMBL and had different units (active/inactive, Ki, IC50, %). AUX descriptors were used in combination with Mordred 2D to develop an ensemble of 50 stacked VotingRegressor models,[36] with each model developed using CatBoost.	

Final blind test: the final RMSE submitted by the team, which was used to score teams.
Lowest: the minimal RMSE for the blind test among all submissions of the respective team.
[1]The hypothetical place had the challenge closed on 15.08.24.
[2]Submission had the same RMSE as the submission from #2 (submitted 21-08-2024 08:53:04), but was submitted later (30-08-2024 07:16:19) than #2.
[3]Team submitted its first predictions after 15.08.24 and thus no results were available.

Analysis of RMSEs following release of the leaderboard set

The release of the leaderboard set, which extended the training set to 1212 molecules, allowed several teams to improve upon the RMSE obtained for the blind set. The lowest RMSE for the blind set (20.4) was obtained by teams #9 and #3, but these participants did not submit the respective predictions with these results as their final entries. In the case of #3, the results for the leaderboard set were not predictive of the blind set (i.e., this team had entries with, e.g., leaderboard RMSE=0) that could not be used to guide selection of the model. Team #9 had the lowest RMSE of 14.8 for the submission leaderboard results with the lowest RMSE for the blind entry that would have won the competition if selected for the final submission.

In general, the release of the leaderboard set allowed seven teams to obtain RMSEs lower than 20.9, thus surpassing the lowest RMSE for the blind set achieved by team #2 before release of the leaderboard set. This observation proved the benefit of releasing the leaderboard data to the participants: otherwise teams that could successfully identify compounds from this set would have had a significant advantage over participants who did not do this. Thus, there would be a risk of us inadvertently comparing the ability of teams to identify data from the leaderboard set, rather than comparing the performances of methods. Six out of seven teams which achieved an RMSE < 20.9 are in Table 1 (#1 to #5 and #9) and only one team, manbaritone, did not rank as one of the "group winners". The last submission from team manbaritone had an RMSE=21.6, narrowly falling short of the threshold to be considered a "group winner". Only three out of eleven models in Table 1 (#3, #5 and #9) did not select their best submission with minimal score as the final submission.

Models developed using OCHEM

The winner #1, as well as runners-up #2 and #6, developed their models using the OCHEM platform. This impressive result further confirms the high accuracy of OCHEM, which was previously used to develop the top model in the ToxCast EPA Challenge,[37] the best balanced accuracy[38] in the Tox21 challenge, and which also contributed the winning model in the Kaggle EUOS/SLAS Solubility Challenge.[17] In addition to providing a highly accurate and convenient way to model data, models developed using OCHEM for the challenge are made publicly available and can be used by other groups to predict new compounds. All three models used Transformer CNF2,[15] which is an extension of Transformer CNN.[14] The latter method was also used by winning and runner-up models developed within this challenge. Both of these methods are Natural Language Processing (NLP) methods that were pre-trained on a large corpus of chemical structures (SMILES from ChEMBL database; no activity data were used), which contributed to their high accuracy.

Use of mixtures

The winning team developed models using descriptors for mixtures. In the case of compounds that consisted of several components, descriptors were calculated for each component and averaged. These types of descriptors were added in OCHEM for analysis of properties of binary non-additive mixtures,[39] ionic liquids[15,40] but also were apparently useful for this study. The rationale behind using mixture descriptors is that it is not always the main component that gives rise to toxicity.

Use of tautomers

While the winning team used mixture descriptors, the runner-up team #2 separated mixtures into individual records, while also enumerating tautomers using the same activities reported for the original structure. This data augmentation strategy increased the dataset to 2400 records, which were used to develop the model. The idea behind this analysis was that it could not be known for certain that the tautomer provided in the training set was active and contributed to toxicity both in the training and test sets.

Data cleaning

Team #9 noticed that after desalting, there were some duplicates in the datasets and only 1165 unique molecules were available, for which the participant used averaged values. Team #2 also corrected several structures and excluded duplicates as well as outliers, which were identified as records with large errors. Team #7 removed inorganic compounds, salts, and any compounds containing metals, rare atoms, or other special atoms thus reducing the dataset size (training + leaderboard) to 1125 compounds. It should be mentioned that no one team used negative data (i.e., discarded compounds due to autofluorescence) or attempted to predict autofluorescence properties of compounds to improve the models. Possibly such modelling could further improve the quality of models which is worth exploring in the future.

Consensus modelling

With an exception of team #10, which used transfer learning, all models from Table 1 were consensus models. Some participants referred to some models as ensembles, but we reserve the term ensemble for a set of models developed using the same method, such as the ensemble of 50 models used by #11, while consensus is a more general term and can include models developed using different methods, or the same method. Note that DLCA,[28] which was used by team #7, inherently combined several different descriptors and learned representation based on SMILES, meaning it was also counted as a consensus model.

Use of other data

Several teams #10 and #11 used data collected from bioassays to develop models to be used as descriptors for the final model (#3, #11), while team #5 collected data to improve training for the GGAP-CPI (protein Graph and ligand Graph network with Attention Pooling for Compound-Protein Interaction Prediction) model and team #11 used additional data for transfer learning (#11). Team #9 developed a consensus of four Foundation Chemistry Models. Two MolFormer-based models were pre-trained on ~50k data points from the ZINC database, as well as Tox21 and Tox24 datasets. These models, in addition to the SMI-TED and UniMol models, were fine-tuned using predictions of boosted decision-tree algorithms during the last step.

Overall analysis of ML methods

In 8 out of 11 models, the authors used at least one representation learning method. They included Graph Neural Networks (ChemProp, Attentive Fingerprints, GIN) as well as Natural Language Processing (NLP) methods based on SMILES data processing (Transformer CNN, CNF2, Convolutional Neural Networks) or Foundation Chemistry Models (MolFormer, SMI-TED and UniMol). The use of fine-tuned Foundation Chemistry Models for TTR binding prediction reported here is, to our knowledge, one of the first reported uses of such models for toxicity prediction, which complements other studies highlighting the promise of using of Foundation Chemistry Models and Large Language Models[41] in chemistry and material sciences.

Among traditional methods, the most popular were decision trees (RF and boosted decision trees, such as CatBoost, LightGBM, XGBoost) but other traditional methods based on fully connected DNN and SVM were also used. The majority of the descriptor-based models were developed using 2D descriptors, such as Estate, Mordred, Mold2, CDK, Morgan, Avalon, and AtomPair. Only one team used 3D RBF descriptors.

Conclusions

There has been significant progress in development of advanced ML tools that can be used for toxicity predictions since the Tox21 challenge took place. In particular, representation learning methods have gained strong momentum and were used in the majority of studies done by the “group winners” of the Tox24 challenge, which included both winning and runner-up models. Fine-tuned Foundation Chemistry Models also demonstrated high predictive accuracy. Many of these approaches are less than five years old and did not exist during the Tox21 Challenge. These observations clearly demonstrate the high impact that advanced ML/AI methods have made on the field.

The Tox24 Challenge had 78 participating teams in total - almost twice the number of teams (40) that participated in Tox21.[2] This is indicative of an increasing interest and a wide engagement of the scientific community in the use of New Approach Methodologies (NAMs) for toxicity prediction.

Considering the high accuracy of novel methods, the OECD principles on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models[42] may need to be extended to describe/exemplify how these approaches and their intrinsic features (e.g., consensus, multitasking, pre-training, fine-tuning, transfer learning, etc.) can be reliably used in regulatory assessments. In combination with explainable AI, these methods could be used to create more accurate and interpretable risk assessments for chemicals.

Acknowledgements

The challenge was co-organised by Marie Sklodowska-Curie Innovative Training Network European Industrial Doctorate grant agreement No. 956832 “Advanced machine learning for Innovative Drug Discovery” (AIDD), Horizon Europe Marie Skłodowska-Curie Actions Doctoral Network grant agreement No. 101120466 “Explainable AI for Molecules” (AiChemist) as well as by Chemical Research in Toxicology journal and ICANN2024. IVT sincerely thanks Larisa Charochkina for help with data curation and Katya Ahmad for her remarks and corrections. We also thank the competition participants for their comments on the manuscript describing the approaches used by each team.

S. Eytcheson was supported in part by an appointment to the Research Participation Program at the Office of Research and Development Center for Computational Toxicology and Exposure, U.S. Environmental Protection Agency, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and EPA. The views expressed in this paper are those of the authors and do not necessarily reflect on the views or policies of the U.S. Environmental Protection Agency, nor does the mention of trade names or commercial products indicate endorsement by the federal government.

Supporting Information

The Supporting Information is available free of charge at xxxx.
Examples of descriptions of methods submitted by teams. The full descriptions can be found in the respective articles with challenge results.

References

(1) Tetko, I. V. Tox24 Challenge. *Chem. Res. Toxicol.* **2024**, *37* (6), 825–826. https://doi.org/10.1021/acs.chemrestox.4c00192.
(2) Huang, R.; Xia, M.; Nguyen, D.-T.; Zhao, T.; Sakamuru, S.; Zhao, J.; Shahane, S. A.; Rossoshek, A.; Simeonov, A. Tox21Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways as Mediated by Exposure to Environmental Chemicals and Drugs. *Front. Environ. Sci.* **2016**, *3*.
(3) Eytcheson, S. A.; Zosel, A. D.; Olker, J. H.; Hornung, M. W.; Degitz, S. J. Screening the ToxCast Chemical Libraries for Binding to Transthyretin. *Chem. Res. Toxicol.* **2024**, *37* (10), 1670–1681. https://doi.org/10.1021/acs.chemrestox.4c00215.
(4) Richard, A. M.; Huang, R.; Waidyanatha, S.; Shinn, P.; Collins, B. J.; Thillainadarajah, I.; Grulke, C. M.; Williams, A. J.; Lougee, R. R.; Judson, R. S.; Houck, K. A.; Shobair, M.; Yang, C.; Rathman, J. F.; Yasgar, A.; Fitzpatrick, S. C.; Simeonov, A.; Thomas, R. S.; Crofton, K. M.; Paules, R. S.; Bucher, J. R.; Austin, C. P.; Kavlock, R. J.; Tice, R. R. The Tox21 10K Compound Library: Collaborative Chemistry Advancing Toxicology. *Chem. Res. Toxicol.* **2021**, *34* (2), 189–216. https://doi.org/10.1021/acs.chemrestox.0c00264.
(5) Richardson, S. J. Cell and Molecular Biology of Transthyretin and Thyroid Hormones. In *International Review of Cytology*; Academic Press, 2007; Vol. 258, pp 137–193. https://doi.org/10.1016/S0074-7696(07)58003-4.
(6) Rabah, S. A.; Gowan, I. L.; Pagnin, M.; Osman, N.; Richardson, S. J. Thyroid Hormone Distributor Proteins During Development in Vertebrates. *Front. Endocrinol.* **2019**, *10*.
(7) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I. I.; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q.-Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information. *J. Comput. Aided Mol. Des.* **2011**, *25* (6), 533–554. https://doi.org/10.1007/s10822-011-9440-2.
(8) Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. In *Proceedings of the 32nd International Conference on*

Neural Information Processing Systems; Curran Associates Inc.: Montréal, Canada, 2018; pp 6639–6649.
(9) Hong, H.; Xie, Q.; Ge, W.; Qian, F.; Fang, H.; Shi, L.; Su, Z.; Perkins, R.; Tong, W. Mold2, Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics. *J. Chem. Inf. Model.* **2008**, *48* (7), 1337–1344. https://doi.org/10.1021/ci800038f.
(10) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (6), 1488–1493. https://doi.org/11749573.
(11) Tetko, I. V.; Tanchuk, V. Y. Application of Associative Neural Networks for Prediction of Lipophilicity in ALOGPS 2.1 Program. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (5), 1136–1145. https://doi.org/12377001.
(12) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35* (6), 1039–1045. https://doi.org/10.1021/ci00028a014.
(13) Huuskonen, null; Livingstone, null; Tetko, null. Neural Network Modeling for Estimation of Partition Coefficient Based on Atom-Type Electrotopological State Indices. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (4), 947–955. https://doi.org/10.1021/ci9904261.
(14) Karpov, P.; Godin, G.; Tetko, I. V. Transformer-CNN: Swiss Knife for QSAR Modeling and Interpretation. *J. Cheminformatics* **2020**, *12* (1), 17. https://doi.org/10.1186/s13321-020-00423-w.
(15) Makarov, D. M.; Fadeeva, Y. A.; Shmukler, L. E.; Tetko, I. V. Machine Learning Models for Phase Transition and Decomposition Temperature of Ionic Liquids. *J. Mol. Liq.* **2022**, *366*, 120247. https://doi.org/10.1016/j.molliq.2022.120247.
(16) Reiser, P.; Eberhard, A.; Friederich, P. Graph Neural Networks in TensorFlow-Keras with RaggedTensor Representation (Kgcnn). *Softw. Impacts* **2021**, *9*, 100095. https://doi.org/10.1016/j.simpa.2021.100095.
(17) Hunklinger, A.; Hartog, P.; Šícho, M.; Godin, G.; Tetko, I. V. The openOCHEM Consensus Model Is the Best-Performing Open-Source Predictive Model in the First EUOS/SLAS Joint Compound Solubility Challenge. *SLAS Discov.* **2024**, *29* (2), 100144. https://doi.org/10.1016/j.slasd.2024.01.005.
(18) Ji, C.; Svensson, F.; Zoufir, A.; Bender, A. eMolTox: Prediction of Molecular Toxicity with Confidence. *Bioinformatics* **2018**, *34* (14), 2508–2509. https://doi.org/10.1093/bioinformatics/bty135.
(19) Corso, G.; Cavalleri, L.; Beaini, D.; Liò, P.; Velickovic, P. Principal Neighbourhood Aggregation for Graph Nets. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Vancouver, BC, Canada, 2020; p Article 1112.
(20) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32. https://doi.org/10.1023/A:1010933404324.
(21) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminformatics* **2018**, *10* (1), 4. https://doi.org/10.1186/s13321-018-0258-y.
(22) *Toxicity Predictor*. http://mmi-03.my-pharm.ac.jp/tox1/prediction_groups/new (accessed 2024-12-06).
(23) Gu, Y.; Li, J.; Kang, H.; Zhang, B.; Zheng, S. Employing Molecular Conformations for Ligand-Based Virtual Screening with Equivariant Graph Neural Network and Deep Multiple Instance Learning. *Molecules* **2023**, *28* (16). https://doi.org/10.3390/molecules28165982.
(24) Béquignon, O. J. M.; Bongers, B. J.; Jespers, W.; IJzerman, A. P.; van der Water, B.; van Westen, G. J. P. Papyrus: A Large-Scale Curated Dataset Aimed at Bioactivity Predictions. *J. Cheminformatics* **2023**, *15* (1), 3. https://doi.org/10.1186/s13321-022-00672-x.
(25) Xia, S.; Zhang, D.; Zhang, Y. Multitask Deep Ensemble Prediction of Molecular Energetics in Solution: From Quantum Mechanics to Experimental Properties. *J. Chem. Theory Comput.* **2023**, *19* (2), 659–668. https://doi.org/10.1021/acs.jctc.2c01024.
(26) Fang, Y.; Zhang, Q.; Zhang, N.; Chen, Z.; Zhuang, X.; Shao, X.; Fan, X.; Chen, H. Knowledge Graph-Enhanced Molecular Contrastive Learning with Functional Prompt. *Nat. Mach. Intell.* **2023**, *5* (5), 542–553. https://doi.org/10.1038/s42256-023-00654-0.
(27) Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliazkova, N.; Kuhn, S.; Pluskal, T.; Rojas-Chertó, M.; Spjuth, O.; Torrance, G.; Evelo, C. T.; Guha, R.;

Steinbeck, C. The Chemistry Development Kit (CDK) v2.0: Atom Typing, Depiction, Molecular Formulas, and Substructure Searching. *J. Cheminformatics* **2017**, *9* (1), 33. https://doi.org/10.1186/s13321-017-0220-4.
(28) Zakharov, A. V.; Zhao, T.; Nguyen, D.-T.; Peryea, T.; Sheils, T.; Yasgar, A.; Huang, R.; Southall, N.; Simeonov, A. Novel Consensus Architecture To Improve Performance of Large-Scale Multitask Deep Learning QSAR Models. *J. Chem. Inf. Model.* **2019**, *59* (11), 4613–4624. https://doi.org/10.1021/acs.jcim.9b00526.
(29) Landrum, G. RDKit: Open-Source Cheminformatics. **2006**.
(30) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20* (3), 273–297. https://doi.org/10.1007/BF00994018.
(31) Li, H.; Zhang, R.; Min, Y.; Ma, D.; Zhao, D.; Zeng, J. A Knowledge-Guided Pre-Training Framework for Improving Molecular Representation Learning. *Nat. Commun.* **2023**, *14* (1), 7568. https://doi.org/10.1038/s41467-023-43214-1.
(32) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59* (8), 3370–3388. https://doi.org/10.1021/acs.jcim.9b00237.
(33) Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; Das, P. Large-Scale Chemical Language Representations Capture Molecular Structure and Properties. *Nat. Mach. Intell.* **2022**, *4* (12), 1256–1264. https://doi.org/10.1038/s42256-022-00580-7.
(34) *SMI-TED: A large-scale foundation model for materials and chemistry | OpenReview*. https://openreview.net/forum?id=Yq8At31hLi (accessed 2024-12-06).
(35) Zhou, G.; Gao, Z.; Ding, Q.; Zheng, H.; Xu, H.; Wei, Z.; Zhang, L.; Ke, G. Uni-Mol: A Universal 3D Molecular Representation Learning Framework; 2022.
(36) Wolpert, D. H. Stacked Generalization. *Neural Netw.* **1992**, *5* (2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1.
(37) Novotarskyi, S.; Abdelaziz, A.; Sushko, Y.; Körner, R.; Vogt, J.; Tetko, I. V. ToxCast EPA in Vitro to in Vivo Challenge: Insight into the Rank-I Model. *Chem. Res. Toxicol.* **2016**, *29* (5), 768–775. https://doi.org/10.1021/acs.chemrestox.5b00481.
(38) Abdelaziz, A.; Spahn-Langguth, H.; Schramm, K.-W.; Tetko, I. V. Consensus Modeling for HTS Assays Using In Silico Descriptors Calculates the Best Balanced Accuracy in Tox21 Challenge. *Front. Environ. Sci.* **2016**, *4*.
(39) Oprisiu, I.; Novotarskyi, S.; Tetko, I. V. Modeling of Non-Additive Mixture Properties Using the Online CHEmical Database and Modeling Environment (OCHEM). *J. Cheminformatics* **2013**, *5* (1), 4. https://doi.org/10.1186/1758-2946-5-4.
(40) Makarov, D. M.; Fadeeva, Yu. A.; Shmukler, L. E.; Tetko, I. V. Beware of Proper Validation of Models for Ionic Liquids! *J. Mol. Liq.* **2021**, *344*, 117722. https://doi.org/10.1016/j.molliq.2021.117722.
(41) Van Herck, J.; Gil, M. V.; Jablonka, K. M.; Abrudan, A.; Anker, A. S.; Asgari, M.; Blaiszik, B.; Buffo, A.; Choudhury, L.; Corminboeuf, C.; Daglar, H.; Elahi, A. M.; Foster, I. T.; Garcia, S.; Garvin, M.; Godin, G.; Good, L. L.; Gu, J.; Xiao Hu, N.; Jin, X.; Junkers, T.; Keskin, S.; Knowles, T. P. J.; Laplaza, R.; Lessona, M.; Majumdar, S.; Mashhadimoslem, H.; McIntosh, R. D.; Moosavi, S. M.; Mouriño, B.; Nerli, F.; Pevida, C.; Poudineh, N.; Rajabi-Kochi, M.; Saar, K. L.; Hooriabad Saboor, F.; Sagharichiha, M.; Schmidt, K. J.; Shi, J.; Simone, E.; Svatunek, D.; Taddei, M.; Tetko, I.; Tolnai, D.; Vahdatifar, S.; Whitmer, J.; Wieland, D. C. F.; Willumeit-Römer, R.; Züttel, A.; Smit, B. Assessment of Fine-Tuned Large Language Models for Real-World Chemistry and Material Science Applications. *Chem. Sci.* **2025**. https://doi.org/10.1039/D4SC04401K.
(42) *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*. OECD. https://doi.org/10.1787/9789264085442-en (accessed 2024-12-10).

Supporting Information

Examples of descriptions of methods submitted by teams. The full descriptions can be found in the respective articles with challenge results.

Team #1: Amidoff

Dmitriy M. Makarov, Alexander A. Ksenofontov

G. A. Krestov Institute of Solution Chemistry of the Russian Academy of Sciences, Ivanovo, Russia

At the outset of the Challenge, we constructed a baseline CatBoost model utilising filtered Mold2 descriptors and sought to enhance it by incorporating additional properties. In order to achieve this, preliminary models were created in order to predict water solubility, melting point and various toxicity endpoints. The predictions from these models were combined with the Mold2 descriptors in order to generate the final prediction of TTR binding activity. As the inclusion of these additional properties did not result in an improvement in model performance, a transition was made to modelling on the OCHEM web platform.

A variety of descriptor packages were employed in the OCHEM, including Mold2, ALogPS and OEstate, Mordred, PaDEL, and MACCS fingerprints. Following the generation of the descriptors, a filtration process was conducted to remove those deemed ineffective for describing structure-activity relationships. Given that approximately 10% of the dataset consisted of mixtures and salts, it was deemed inappropriate to remove small fragments and neutralise the molecules during the calculation of the descriptors.

In order to ascertain the most efficacious algorithms, a number of methods based on pre-computed descriptors were subjected to evaluation. The evaluation included Random Forest, XGBoost, CatBoost, and Associative Neural Networks. Additionally, several representation learning methods that do not necessitate the pre-calculation of descriptors were employed: the Transformer Convolutional Neural Network (TransCNN), the Transformer Convolutional Neural Fingerprint (TransCNF), and the Graph Convolutional Neural Network from Chemprop.

Initially, the best models were selected and their performance was assessed using 5-fold cross-validation on a combined dataset comprising the training set and leaderboard set. The top-performing models were integrated into a consensus model, which was then evaluated on the blind test set. Our team's winning model was based on the averaged predictions of four individual models: CatBoost/ALogPS and OEstate, CatBoost/Mold2, TransCNF, and TransCNN.

Team #2: tcirino

Thalita Cirino
Molecular Biotechnology and Health Sciences Department, University of Torino, Italy

In the Tox24 Challenge,[1] I secured 2nd position out of 79 participants, with a final predictions set submission reaching an RMSE of 20.7. My modeling strategy relied primarily on two key approaches: data augmentation through tautomer generation and consensus modeling. The integration of data augmentation provided a comprehensive representation of the dataset.

Data Preprocessing. Compounds were standardized through the OCHEM[2] preprocessing built-in tools, which includes (I) correction functional group (e.g., nitro and azido groups) chemical representation; (II) charge neutralization by attaching additional hydrogen atoms, and (III) converting structures into canonical SMILES before generating 3D structures, eliminating inconsistencies due to uploaded conformations and (IV) salt counterions and other small molecules remotion from chemical structures, leaving just the parent compound.

Data augmentation. To broaden the representation of the chemical space, tautomeric forms of each compound were generated using the TautomerEnumerator class in RDKit 2020.03.1.[3] Tautomerism, a form of structural isomerism where compounds can exist in different forms through the migration of a hydrogen atom and reorganization of double bonds, can significantly influence molecular properties and interactions.[4] To maintain computational feasibility, the number of tautomers was limited to 20 per molecule. This approach expanded the dataset from 1,212 to 3,438 compounds, providing a more comprehensive coverage of potential molecular conformations that could influence toxicological properties.

Model Development. A 5-fold cross-validation (5CV) protocol was used to evaluate the performance of various machine learning models available in OCHEM. Models with the lowest processing errors and the best 5CV metrics were selected for the final consensus model, which consisted in combining models selected through a non-weighted average to produce the final predictions. The models selected were based on the following methods:

- CatBoost[5] - Mold2[6], a gradient-boosting algorithm combined with a set of 2D molecular descriptors;
- Message Passing Neural Network,[7] a graph-based algorithm that has each node receiving messages from its neighbours to then update its representation based on the aggregated message, which promotes it adaptabilidy in learning from a hybrid representation, merging task-specific encodings and fixed descriptors;
- Transformer Convolutional Neural Networks (TransCNN),[8] a Natural Language Processing method that was pre-trained over 1.7M molecules from the ChEMBL database[9] to learn the task of canonisation of chemical structures. The learned latent representation is used as input to one-dimensional Convolutional Neural Network (CNN) and its output is correlated with the target properties of molecules using fully connected neural networks;
- Transformer Convolutional Neural Network Fingerprint (CNF2)[10] extends the previous method and uses a combination of several CNNs with different receptive fields to provide a richer representation (fingerprint) to be correlated with target properties of molecules.

References

(1) Tetko, I. V. Tox24 Challenge. Chemical Research in Toxicology 2024, 37, 825–826.

(2) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I. I.; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de Sousa, J.; Zhang, Q.-Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. Journal of Computer-Aided Molecular Design 2011, 25, 533–554.

(3) Landrum, G. RDKit: Open-source cheminformatics. 2013; http://www.rdkit.org.

(4) De Oliveira, C.; Yu, H. S.; Chen, W.; Abel, R.; Wang, L. Rigorous Free Energy Perturbation Approach to Estimating Relative Binding Affinities between Ligands with Multiple Protonation and Tautomeric States. Journal of Chemical Theory and Computation 2019, 15, 424–435.

(5) Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; Gulin, A. CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems 2018, 31.

(6) Hong, H.; Xie, Q.; Ge, W.; Qian, F.; Fang, H.; Shi, L.; Su, Z.; Perkins, R.; Tong, W. Mold 2 , Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics. Journal of Chemical Information and Modeling 2008, 48, 1337–1344.

(7) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. 2017; https://arxiv.org/abs/1704.01212, Version Number: 2.

(8) Karpov, P.; Godin, G.; Tetko, I. V. Transformer-CNN: Swiss knife for QSAR modelling and interpretation. Journal of Cheminformatics 2020, 12, 17.

(9) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large scale bioactivity database for drug discovery. Nucleic Acids Research 2012, 40, D1100–D1107.

(10) Makarov, D. M.; Fadeeva, Y. A.; Shmukler, L. E.; Tetko, I. V. Machine learning models for phase transition and decomposition temperature of ionic liquids. Journal of Molecular Liquids 2022, 366, 120247.

Team #3: znavoyan

Algorithm Description

Z. Navoyan[1], A. Tevosyan[1], H. Yeghiazaryan[1], L. Khondkaryan[1,3], N. Abelyan[2], V. Atoyan[1], N. Babayan[1,3]

1. Toxometris.ai, Glendale, CA 91204, USA
2. Biocentric.ai, Yerevan 0075, Armenia
3. Institute of Molecular Biology, NAS RA, Yerevan 0014, Armenia

In order to achieve superior results, several techniques described below are used, each of which improves RMSE by a different factor. Techniques that were tried but did not lead to a significant improvement in results are not included.

- Train Random Forest (RF) models on bioassay data from ToxCast and eMolTox, then use these models to make predictions on data from the Tox24 Challenge. Subsequently, use these predictions as descriptors in the main models. The bioassay data preprocessing and model training are fully automated, e.g. automatically detecting and filtering nan containing data points, low variance data etc. In cases where the bioassay data is unbalanced, different balancing techniques are automatically applied based on the degree of imbalance. For highly imbalanced cases (imbalance ratio < 10%), undersampling is used. For moderate imbalance, the SMOTE balancing technique is applied. When the number of data points in both classes is nearly equal, no balancing is used. The trained models are then validated to ensure they have an AUC greater than a specified threshold.
- The best-performing descriptors for RF were selected using a Genetic Algorithm.
- Different models are used in an ensemble, and the final prediction is calculated by averaging the predictions of the individual models within the ensemble. This allows for capturing different aspects of molecular representations. We used the following models:
    - Random Forest based on RDKit descriptors + Bioassay descriptors,
    - Random Forest based on oestate descriptors + Bioassay descriptors,
    - Graph Neural Network.
- Use of Specialized GNN: We utilize a GNN with a residual block architecture, similar to ResNet. Up to 30 different GNN kernels were tested, and ultimately, PNA (Principal Neighborhood Aggregation) was chosen as the best solution for this problem. Node, edge, and graph features were meticulously selected from RDKit and Bioassay descriptors and used in GNN training.
- Multitask setup: With the help of our partners molecular docking was performed for compounds from the dataset and the calculated binding score was used along with other bioassay descriptors.

Team #8: alx.dga

Alexis Dougha[1]

[1] BFA, Université Paris Cité, CNRS UMR 8251, INSERM U1133, Paris, France

1. Data preprocessing and featurisation

First, drop salt from SMILES using RDKit SaltRemover. Then, 2 types of features were compared:

a) RDKit physicochemical features
b) Latent features generated with graph transformer KPGT (Li et al. 2023)

2. Data split (Train/Val)

2 splitting strategies with the initial 999 compounds dataset to find best parameters for each model:

a) "Hard split": For each compound, computational of Tanimoto Similarity for all pairs of compounds. Compounds with max similarity to any other compound < 1/3 are in Val set, the other compounds in Train set
b) "Easy split": Random 774/225 split (to obtain same ratio as hard split)

3. Benchmarked models

I chose to compare and optimise the hyperparameters of the following models: Support Vector Machine (SVM) and Random Forest (RF) with a grid search, the D-MPNN Chemprop (Yang et al. 2019) with Bayesian Optimisation.

➔ Some of the best performing models (criterium is RMSE on leaderboard set):
A. SVM with rbf, C=100, epsilon=0.5, KPGT features
B. RF with n_estimators=500, max_features=1/6, min_samples_split=6, RDKit features
C. SVM with rbf, C=65, epsilon=1.5, KPGT features selected with recursive feature elimination using RF
D. Chemprop (with hard split for training and validation sets)

Note that for **A** and **C**, the hyperparameters were best performing with the easy split while for **B** and **D**, they were best performing with the hard split.

4. Model selection in similarity intervals

I tried to find the best combination of models A, B, C, D to predict different subsets of the leaderboard set based on their Tanimoto similarity to the 999 compounds dataset.

Max similarity vs training set	[0, 0.25[	[0.25, 0.4[	[0.4, 0.7[	[0.7, 1]
	C	0.75***C** + 0.25***B**	0.5***C** + 0.5*(0.75***A** + 0.25***B**)	**D**

Table 1: Partition of the leaderboard set into 4 subsets depending on the Tanimoto similarity

5. Retrain adding the leaderboard dataset

For **A**, **B**, **C** I simply use 100% of available data for training. For **D**, I add the leaderboard compounds to the training set if max similarity to any other compound (initial 999-training set + leaderboard) is > 1/3, else it is added to validation set.

Finally, I divide the blind set into 4 subsets following the same partition as presented in Table 1 (criterium: max similarity vs any compound of the full 1199 compounds training set, same intervals), I predict the activity for the blind set with models **A**, **B**, **C**, **D** and I weight the predicted values by each model following Table 1. Ensembles of 1000 RF (B) and 50 Chemprop (D) models were used because of their sensitivity to random seeds.

Team #9: luispintoc

Luis Pinto
Independent Researcher

Abstract:

In this competition, I developed a solution that combined an ensemble of four models: two MolFormer models, SMI-TED, and UniMol. This approach secured 9th place, achieving an RMSE of **21.4**. Notably, with a slightly different ensemble, my blind test set RMSE could have dropped to 20.4, lower than the first place, but this lower submission was not selected as my final submission.

Data Preprocessing:

The molecules were cleaned using the *molvs* package, which led to some molecules sharing the same SMILES representation. To handle this, I averaged the target values for these duplicates, reducing the dataset to 1165 unique molecules. It might be beneficial to retain salt/neutralized pairs as separate data points, as their target values can vary significantly.

Model Training and Internal Test Set Creation:

After the release of leaderboard data, I retrained the models to incorporate this new information. I separated 20 molecules as an internal test set, with a final train set of 1145 molecules. Consistent 5-fold cross-validation with the same seed and data ordering was applied across all models. The ensemble was built on the out-of-fold (OOF) predictions and validated on my internal test set.

Use of Auxiliary Targets:

In an effort to enhance the performance of the deep learning models, I incorporated the OOF predictions from tree-based models (LGBM and CatBoost) trained on different chemical descriptors as auxiliary targets. This approach allowed the deep learning models to learn from the implicit chemical knowledge captured by the tree-based models. However, this method is prone to overfitting, making it essential to establish a robust cross-validation (CV) strategy. Additionally, a strong correlation between CV scores and leaderboard performance is crucial to ensure that the models generalize well. Ideally, the leaderboard and test set distributions should align closely, which fortunately seemed to be the case in this competition. Training on these auxiliary targets did decrease the leaderboard RMSE (and OOF score) by a few points on each model individually.

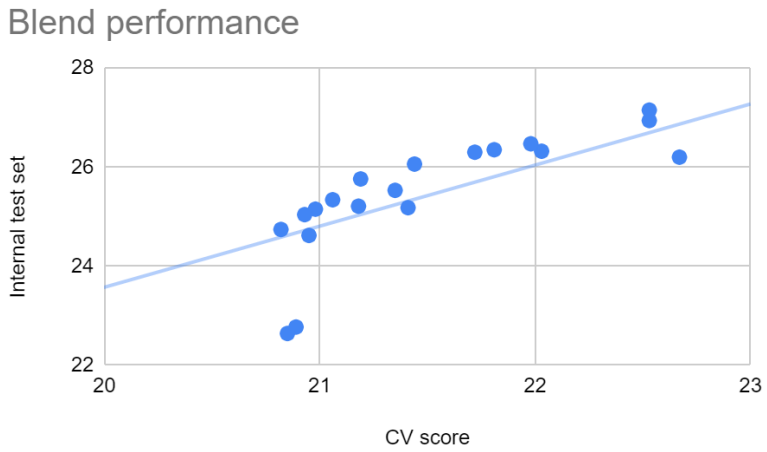
Model-Specific Details:

- **MolFormer**: Two versions of the MolFormer model were included in the ensemble. Both were pretrained on a subset of the ZINC dataset (~50,000 data points closest to the Tox24 molecules using *ECFP*), Tox21, and Tox24 datasets using multitask regression (MTR) on 27 *Mold2* descriptors. This pretraining improved both CV and leaderboard performance by a few RMSE points.
    - The first model was trained with the predictions of an LGBM model using OEState descriptors as auxiliary targets.
    - The second model was trained with the predictions from a CatBoost model using Mold2 descriptors.
- **SMI-TED**: This transformer model was fine-tuned without additional pretraining. It was trained with LGBM predictions using OEState descriptors as auxiliary target.
- **UniMol**: The UniMol model was only fine-tuned and included in the final ensemble. It was trained with LGBM predictions using Mold2 descriptors as auxiliary target.

Challenges and Alternative Approaches:

Interestingly, a previously submitted ensemble version that did not include the MolFormer model trained with CatBoost predictions achieved an RMSE of 20.4, outperforming the eventual 1st place.

Other models tested included XGB, LGBM, CatBoost, ChemBERTa2, and Giraffe, with descriptors from RDKit, Mordred, fingerprints, Mold2, and OEState. Additionally, I experimented with adding docking scores obtained via PyRx as auxiliary outputs and compiled various datasets related to TTR binding as auxiliary tasks, though these efforts didn't improve results.



The two data points at the bottom of the plot correspond to the results that secured 9th place (lowest internal test set RMSE) and the potential 1st place in the competition (second lowest internal test set RMSE). In retrospect, selecting my best model solely based on the performance of 20 internal test data points proved to be a suboptimal approach due to the limited sample size.

Team #10

Team #11: vchupakhin

Tox24 challenge solution

Vladimir Chupakhin, Simulations Plus (vlad.chupakhin@simulations-plus.com), 2024-09-04

Data

Since transthyretin (TTR) is an understudied target, we developed a feature set by leveraging models for related targets. We identified them from the PDB (Burley et al., 2017) that bind to compounds structurally similar to thyroxine and retinol, and cross-referenced these targets with PubChem (Kim et al., 2021) and ChEMBL (Gaulton et al., 2012) e.g. plasma retinol-binding protein. TTR protein structures revealed that the TTR binding site is β-sheet enriched, thus we included amyloid, tau and prions as related targets. Additionally, a literature review uncovered a set of thyroid-related toxicity studies for iodothyronine deiodinases (Olker et., 2020), thyrotropin-releasing hormone receptor, thyroid-stimulating hormone receptor, thyroperoxidase and sodium/iodide symporter (Dracheva et al., 2022), whose associated data were incorporated into the auxiliary data pool. This led to the creation of AUX, a set of 180 features derived from classification (e.g. active/inactive) or regression models (Ki, IC50, %) based on various descriptors (up to 4).

Before modeling all compounds went thru established RDKit-based pipeline for compound standardization.

Modeling

Models composing AUX descriptor were based on RDKit 2D (Landrum 2006), Mordred 2D (Moriwaki et al., 2018) and RDKit Morgan fingerprint (R=3, 2048 bits). 5-fold cross-validation was used as a default modeling setup for most of the developed models.

For the initial set of models for TTR we used the CatBoostRegressor algorithm (Prokhorenkova et al., 2018) with the following descriptors: AUX, RDKit 2D, Mordred 2D, ChemBERTA (Chithrananda et al., 2020), Reduced Graph Fingerprint (Stiefl et al., 2006), and RDKit Morgan fingerprints (Rogers & Hahn, 2010) with varying radii and folded sizes.

We selected the top four descriptors - AUX, Mordred 2D, RDKit 2D, and Morgan fingerprint—based on their RMSE rankings for ensemble modeling. Utilizing the scikit-learn framework (Pedregosa et al., 2011), we exhaustively tested combinations of these descriptors with either the Voting or Stacking Regressor (Wolpert, 1992). The most effective heterogeneous ensemble consisted of CatBoostRegressor models based on Mordred 2D and AUX descriptors combined with a VotingRegressor as a final model. To further enhance performance, we introduced an ensemble involving 50 distinct VotingRegressor models with multiple random initializations with median of those predictions as our final submission.

Insights gained from this approach will be considered for future integration into the ADMET Predictor® platform by Simulations Plus, Inc. A full list of used sources and references will be provided upon request.

References

Burley SK, Berman HM, Kleywegt GJ, Markley JL, Nakamura H, Velankar S. Protein Data Bank (PDB): the single global macromolecular structure archive. Methods in Molecular Biology. 2017;1607:627-641. doi:10.1007/978-1-4939-7000-1_26.

Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res. 2021 Jan 8;49(D1). doi: 10.1093/nar/gkaa971.

Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012 Jan;40(Database issue). doi: 10.1093/nar/gkr777.

Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems. 2018;31.

Moriwaki H, Tian YS, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. Journal of Cheminformatics. 2018;10(1):4. doi: 10.1186/s13321-018-0258-y.

Chithrananda S, Grand G, Ramsundar B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885. 2020.

Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, Gómez-Bombarelli R, Hirzel T, Aspuru-Guzik A, Adams RP. Convolutional networks on graphs for learning molecular fingerprints. Advances in Neural Information Processing Systems. 2015;28.

Rogers D, Hahn M. Extended-connectivity fingerprints. Journal of Chemical Information and Modeling. 2010;50(5):742-754. doi: 10.1021/ci100050t.

Wolpert DH. Stacked generalization. Neural Networks. 1992;5(2):241-259. doi: 10.1016/S0893-6080(05)80023-1.

Dracheva, E., Norinder, U., Rydén, P., Golosovskaia, E., Örn, S., & Ahrens, L. (2022). In silico identification of potential thyroid hormone system disruptors among chemicals in human serum and chemicals with a high exposure index. Environmental Science and Technology, 56(12), 8363-8372.