Home
Teaching
Research
Publications
People
Blackboard
Software
Contact
Bachelor Study Bioinformatics
Master Study Bioinformatics

Tox21 Machine Learning Data Set

Training and test data contain both dense and sparse features

The Tox21 data set comprises 12,060 training samples and 647 test samples that represent chemical compounds. There are 801 "dense features" that represent chemical descriptors, such as molecular weight, solubility or surface area, and 272,776 "sparse features" that represent chemical substructures (ECFP10, DFS6, DFS8; stored in Matrix Market Format). Machine learning methods can either use sparse or dense data or combine them. For each sample there are 12 binary labels that represent the outcome (active/inactive) of 12 different toxicological experiments. Note that the label matrix contains many missing values (NAs). The original data source and Tox21 challenge site is https://tripod.nih.gov/tox21/challenge/.

- Training data DENSE features: tox21_dense_train.csv.gz (51 MB)
- Training data SPARSE features:
    - tox21_sparse_train.mtx.gz (8 MB)
    - column names (4 MB)
    - row names (<1 MB)
- Training data labels: tox21_labels_train.csv.gz (<1 MB)

- Test data DENSE features: tox21_dense_test.csv.gz (1 MB)
- Test data SPARSE features:
    - tox21_sparse_test.mtx.gz (<1 MB)
    - column names (4 MB)
    - row names (<1 MB)
- Test data labels: tox21_labels_test.csv.gz (< 1MB)

Code to run RandomForest:

- Python code: sampleCode.py (<1 MB)
- R code: sampleCode.R (<1 MB)

Tox21 Package

- tox21.zip(~36 MB): contains all data files and sample codes

Additional information on the data set

- Tox21 compound data in tabular format: tox21_compoundData.csv (1MB)
- Tox21 compound data in chemical format: tox21.sdf.gz (3MB)

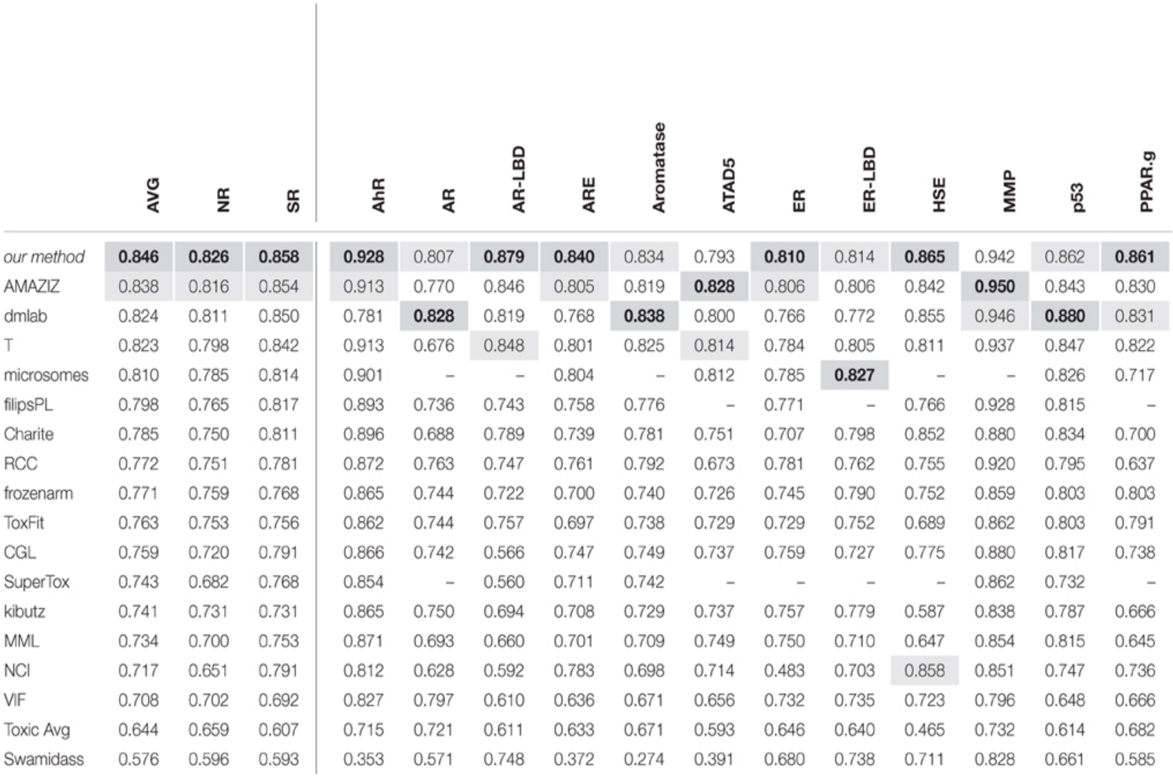
References

If you use this data set please cite the following publications:

[Mayr2016] Mayr, A., Klambauer, G., Unterthiner, T., & Hochreiter, S. (2016). DeepTox: Toxicity Prediction using Deep Learning. *Frontiers in Environmental Science*, **3**:80.

[Huang2016] Huang, R., Xia, M., Nguyen, D. T., Zhao, T., Sakamuru, S., Zhao, J., Shahane, S., Rossoshek, A., & Simeonov, A. (2016). Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Frontiers in Environmental Science*, **3**:85.

Benchmarking Results



	AVG	NR	SR	AhR	AR	AR-LBD	ARE	Aromatase	ATAD5	ER	ER-LBD	HSE	MMP	p53	PPAR.g
our method	**0.846**	**0.826**	**0.858**	**0.928**	0.807	**0.879**	**0.840**	0.834	0.793	**0.810**	0.814	**0.865**	0.942	0.862	**0.861**
AMAZIZ	0.838	0.816	0.854	0.913	0.770	0.846	0.805	0.819	**0.828**	0.806	0.806	0.842	**0.950**	0.843	0.830
dmlab	0.824	0.811	0.850	0.781	**0.828**	0.819	0.768	**0.838**	0.800	0.766	0.772	0.855	0.946	**0.880**	0.831
T	0.823	0.798	0.842	0.913	0.676	0.848	0.801	0.825	0.814	0.784	0.805	0.811	0.937	0.847	0.822
microsomes	0.810	0.785	0.814	0.901	–	–	0.804	–	0.812	0.785	**0.827**	–	–	0.826	0.717
filipsPL	0.798	0.765	0.817	0.893	0.736	0.743	0.758	0.776	–	0.771	–	0.766	0.928	0.815	–
Charite	0.785	0.750	0.811	0.896	0.688	0.789	0.739	0.781	0.751	0.707	0.798	0.852	0.880	0.834	0.700
RCC	0.772	0.751	0.781	0.872	0.763	0.747	0.761	0.792	0.673	0.781	0.762	0.755	0.920	0.795	0.637
frozenarm	0.771	0.759	0.768	0.865	0.744	0.722	0.700	0.740	0.726	0.745	0.790	0.752	0.859	0.803	0.803
ToxFit	0.763	0.753	0.756	0.862	0.744	0.757	0.697	0.738	0.729	0.729	0.752	0.689	0.862	0.803	0.791
CGL	0.759	0.720	0.791	0.866	0.742	0.566	0.747	0.749	0.737	0.759	0.727	0.775	0.880	0.817	0.738
SuperTox	0.743	0.682	0.768	0.854	–	0.560	0.711	0.742	–	–	–	–	0.862	0.732	–
kibutz	0.741	0.731	0.731	0.865	0.750	0.694	0.708	0.729	0.737	0.757	0.779	0.587	0.838	0.787	0.666
MML	0.734	0.700	0.753	0.871	0.693	0.660	0.701	0.709	0.749	0.750	0.710	0.647	0.854	0.815	0.645
NCI	0.717	0.651	0.791	0.812	0.628	0.592	0.783	0.698	0.714	0.483	0.703	0.858	0.851	0.747	0.736
VIF	0.708	0.702	0.692	0.827	0.797	0.610	0.636	0.671	0.656	0.732	0.735	0.723	0.796	0.648	0.666
Toxic Avg	0.644	0.659	0.607	0.715	0.721	0.611	0.633	0.671	0.593	0.646	0.640	0.465	0.732	0.614	0.682
Swamidass	0.576	0.596	0.593	0.353	0.571	0.748	0.372	0.274	0.391	0.680	0.738	0.711	0.828	0.661	0.585

Performance of methods in terms of AUC. This table reports the performance of methods that participated in the Tox21 Data Challenge in terms of area under ROC curve. The first row, "our method", displays the performance of the Deep Learning method "DeepTox". For details see [Mayr2016].

Contact

Guenter Klambauer.

Acknowledgements

We thank Ruili Huang and Menghang Xia at NIH for organizing the challenge, collaboration and permission to share the data.