Supplementary data - "Using Molecular Embeddings in QSAR Modeling: Does it Make a Difference?"

María Virginia Sabando,[*,†,‡] Ignacio Ponzoni,[†,‡] Evangelos E. Milios,[¶] and Axel J. Soto[†,‡]

†*Institute for Computer Science and Engineering (UNS-CONICET), Bahía Blanca, Argentina*

‡*Department of Computer Science and Engineering, Universidad Nacional del Sur, Bahía Blanca, Argentina*

¶*Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada*

E-mail: virginia.sabando@cs.uns.edu.ar

Contents of the Repository

- *DeepSMILES formulas*: a folder containing one .csv file per labeled dataset (SR-ARE, SR-MMP, SR-ATAD5, HIV, PCBA-686978, ESOL, FreeSolv, Lipophilicity), each corresponding to the *DeepSMILES*[1] canonicalization of the compounds and the activity values.

- *RDKit SMILES formulas*: a folder containing one .csv file per labeled dataset (SR-ARE, SR-MMP, SR-ATAD5, HIV, PCBA-686978, ESOL, FreeSolv, Lipophilicity), each corresponding to the RDKit[2] canonicalization of the compounds and the activity values.

- *Trained embedding models*: a folder containing five .zip files. Each .zip file contains trained embedding models for each of the five embedding methods analyzed in our paper: *SMILESVec*,[3] *Mol2Vec*,[4] *Seq2Seq*,[5] *SA-BiLSTM*[6] and *PaccMann*.[7] These trained models can be loaded and used to extract embeddings for new datasets besides the ones covered in our paper, except for the supervised models *SA-BiLSTM* and *PaccMann* which, because of having undergone a supervised training, are only meant to be used with the datasets used in this study.

- *Source code*: a folder containing all the source code needed in order reproduce our approach. We provide scripts for training the supervised methods *SA-BiLSTM* and *PaccMann*; scripts for training all classification and regression methods tested in our paper (Naïve Bayes, Support Vector Machine, Random Forest, Ridge regression, Gradient Boosting Regression and Feed-Forward Neural Networks); simple scripts showing how to extract embeddings from the pre-trained models provided in this repository and scripts for computing traditional molecular representations.

  Please note that we do not provide the source code to train the unsupervised models *SMILESVec*[i], *Mol2Vec*[ii] and *Seq2Seq*[iii], since we neither own nor have contributed to the development of such source code. Instead, please refer to the corresponding GitHub repositories provided by the original authors of such techniques.

- *Traditional_molecular_representations.zip*: a .zip file containing all traditional molecular representations computed for each of the eight labeled datasets under study (SR-ARE, SR-MMP, SR-ATAD5, HIV, PCBA-686978, ESOL, FreeSolv, Lipophilicity).

[i] `https://github.com/hkmztrk/SMILESVecProteinRepresentation/tree/master/source/word2vec`.
[ii] `https://github.com/samoturk/mol2vec`.
[iii] `https://github.com/XericZephyr/seq2seq-fingerprint`.

Datasets

- *SR-ARE*, a bioassay for small molecule agonists of the antioxidant response element (ARE) signaling pathway[iv]. This assay of data is contained in the *Tox21 challenge* dataset, consisting of qualitative toxicity measurements on twelve biological targets[v].

- *SR-MMP*, a stress response assay for small molecule disruptors of the mitochondrial membrane potential (MMP)[vi]. This dataset is also included in the *Tox21 challenge* dataset.

- *SR-ATAD5*, a set of small molecules that induce genotoxicity in human embryonic kidney cells expressing luciferase-tagged ATAD5[vii]. This dataset also belongs to the *Tox21 challenge* dataset.

- *HIV*, a dataset introduced by the Drug Therapeutics Program (DTP) AIDS Antiviral Screen containing information about molecular ability to inhibit HIV replication. Screening results were categorized as confirmed inactive (CI), confirmed active (CA) and confirmed moderately active (CM). We merged the compounds categorized under the latter two labels, which yielded two classes: inactive (CI) and active (CA and CM)[viii].

- *PCBA-686978*, a PubChem bioassay containing information about molecular ability to inhibit the human tyrosyl-DNA phosphodiesterase 1 (TDP1)[ix].

- *ESOL*, a dataset consisting of water solubility data for 1128 compounds, used to train models that estimate solubility directly from chemical structures.[8]

[iv] `https://pubchem.ncbi.nlm.nih.gov/bioassay/743219`
[v] `https://tripod.nih.gov/tox21/challenge/about.jsp`
[vi] `https://pubchem.ncbi.nlm.nih.gov/bioassay/720637`
[vii] `https://pubchem.ncbi.nlm.nih.gov/bioassay/720516`
[viii] `https://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Data`
[ix] `https://pubchem.ncbi.nlm.nih.gov/bioassay/686978`

- *FreeSolv*—the *Free Solvation Database*—, a dataset comprising experimental and calculated hydration free energy values for small molecules in water.[9]

- *Lipophilicity*, a dataset curated from ChEMBL database[10] including experimental results of octanol/water distribution coefficient (*LogD 7.4*) for 4200 compounds [x].

Contact us

Please refer to our paper for further information:

- "**Using Molecular Embeddings in QSAR Modeling: Does it Make a Difference?**", authored by M.V. Sabando, I. Ponzoni, E.E. Milios and A .J. Soto. Preprint available at arXiv.org.

For further details or inquiries regarding the resources in this repository please contact the corresponding author.

References

(1) Dalke, A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. 2018.

(2) Landrum, G. RDKit: open-source cheminformatics http://www. rdkit. org. 2016.

(3) Öztürk, H.; Ozkirimli, E.; Özgür, A. A novel methodology on distributed representations of proteins using their interacting ligands. *Bioinformatics* **2018**, *34*, i295–i303.

(4) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of Chemical Information and Modeling* **2018**, *58*, 27–35.

[x] `https://www.ebi.ac.uk/chembl/document_report_card/CHEMBL3301361/`

(5) Xu, Z.; Wang, S.; Zhu, F.; Huang, J. Seq2seq fingerprint: An unsupervised deep molecular embedding for drug discovery. Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. 2017; pp 285–294.

(6) Zheng, S.; Yan, X.; Yang, Y.; Xu, J. Identifying structure–property relationships through SMILES syntax analysis with self-attention mechanism. *Journal of Chemical Information and Modeling* **2019**, *59*, 914–923.

(7) Oskooei, A.; Born, J.; Manica, M.; Subramanian, V.; Sáez-Rodríguez, J.; Rodríguez Martínez, M. PaccMann: Prediction of anticancer compound sensitivity with multi-modal attention-based neural networks. *arXiv e-prints* **2018**, arXiv:1811.06802.

(8) Delaney, J. S. ESOL: estimating aqueous solubility directly from molecular structure. *Journal of chemical information and computer sciences* **2004**, *44*, 1000–1005.

(9) Mobley, D. L.; Guthrie, J. P. FreeSolv: a database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design* **2014**, *28*, 711–720.

(10) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S., et al. The ChEMBL bioactivity database: an update. *Nucleic acids research* **2014**, *42*, D1083–D1090.