How well could a deep learning model be trained to predict hERG (and other asssays to be added on plan) libiality?

Machine learning and artificial intelligence approaches are making major waves across science. There are some work done in this field[1]. By combining the Deep Learning techniques with data gained from previous study, we can gain information from a few more models. This allows a better understanding of how deep learning may work in hERG liability prediction. This research would be more centered on supervised machine learning.
We plan to take hERG as our start-up practice of this whole PhD project. Once the first modelling project is completed we will look at other assay targets from members of the GPCR, Ion Channel, tyrosine kinase, and transporter groups as commonly investigated assays in drug development.

In other areas of chemical sciences, statistical models from machine learning experienced growing popularity. They have been shown to reduce the cost of simulating chemical systems,[2–4] improve the accuracy of quantum methods,[5,6] generate force field parameters,[7,8] predict molecular properties and support the design of new materials.[9,10] Neutral networks have been a particularly powerful method,[11,12] used to generate high-quality potential energy surfaces and predict material properties.

Themes

Establishment of a Systematic Database

Normally, the data preparation is a time consuming portion of the Machine Learning related research. Given there are abundant data gained from former research[1]. Getting them well-organised and stored is a vital step before moving on further.
Data storage relies on the database, which is a collection of tables with typed columns. SQL Server supports different data types, including primitive types such as Integer, Float, Decimal, Char (including character strings), Varchar (variable length character strings), binary (for unstructured blobs of data), Text (for textual data) among others. Microsoft SQL Server also allows user-defined composite types (UDTs) to be defined and used. It also makes server statistics available as virtual tables and views (called Dynamic Management Views or DMVs). In addition to tables, a database can also contain other objects including views, stored procedures, indexes and constraints, along with a transaction log. A SQL Server database can contain a maximum of 2^{31} objects and can span multiple OS-level files with a maximum file size of 260 bytes (1 exabyte).

Investigations on recurrent neural network models

Project

A necessary pre-requisite of machine learning models is the availability of training data. Fortunately, there are 1547 molecular datasets on hERG at TOXRIC (https://toxric.bioinforai.tech/home). This data will provide a rich training and test dataset to enable the role of deep learning to be explored and evaluated in this PhD.

We can conceive of cheminformatics as a two-part problem: encoding chemical structure as features (molecular description), and mapping the features to the output property (usually using machine learning).

Molecular Description

To enable effective machine learning, molecules of different sizes must be "featurised" to produce a vector of fixed length that describes the molecule. The Morgan Fingerprint, as programmed within the RDkit Python package, is our initial proposed molecular featuriser. As this fingerprint contains details of the local molecular environments within the molecule, it will likely enable the prediction of at least the local vibrational frequency modes. However, predicting low frequency global vibrational modes and the rotational constants may require different molecular description. A PhD project enables this molecular featurisation problem to be extensively explored. An accurate description of molecular properties is necessary For instance, one of the key challenges in Cheminformatics is obtaining a description of a system's

potential energy surface (PES) that is accurate, transferable, computationally efficient, and as simple as possible[13,14]

Selection of Deep Learning (ML) models

Chemically accurate and comprehensive studies of the virtual space for a large number of molecules are severely limited by the computational cost of our research. A composite strategy that adds machine learning corrections could lead to computational inexpensive approximate methods.[15,16]

The second of these is most often the province of machine learning.[17] There are multiple relatively mature Machine Learning methodologies at present that we will consider, such as Regression,[17] Clustering,[18] Classification,[17–19] Dimensionality Reduction.[20]

An initial research question is what machine learning methods are most suitable for this type of data (in particular, what deep learning methods are most suitable for this portion of work); in general, this is not possible to predict without testing. Therefore, RNN, LSTM and GRU models would be trained on the computationally predicted data.

Proposed Timeframe for a PhD (scope of project needs to be reduced for Honours)

- Yr 1, Month 1-3: Extensive literature review of machine learning in Cheminformatics;
- Yr 1, Month 1-12: Explore and evaluate the suitability of deep learning for predicting frequencies in our study using experimental and computationally-generated datasets;
- Yr 2, Month 1-6: Explore and evaluate the suitability of machine learning for predicting intensities in our study using computationally-generated datasets;
- Yr 2, Month 7-12: Explore and evaluate the suitability of machine learning for predicting ... using computationally-generated datasets;
- Yr 3, Month 1-9: Explore how machine learning methods can be used in ... ;
- Yr 3, Month 10-12: Prepare and submit thesis.

Desired Outcome

This project will explore and evaluate the applicability of deep learning to predict hERG liability. This information will provide guidance to the field of Cheminformatics as to whether and how the new tools of deep learning can be integrated to replace, extrapolate or enhance existing methodologies.
This project will contribute a comprehensive approach to AI modelling of drug development assays to culminate in a useful array of machine learning models that can provide fast cheap answers to important drug development questions. The project is eminently viable and important as it will impact the future direction of a major research field internationally. Results will be disseminated by publications in strong international refereed journals.

REFERENCES

[1] L. S. K. Konda, S. Keerthi Praba, and R. Kristam, "herg liability classification models using machine learning techniques," *Computational Toxicology*, vol. 12, p. 100089, 2019.

[2] A. Lopez-Bezanilla and O. A. von Lilienfeld, "Modeling electronic quantum transport with machine learning," *Phys. Rev. B*, vol. 89, p. 235411, Jun 2014.

[3] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, "Accelerating materials property predictions using machine learning," *Scientific reports*, vol. 3, p. 2810, 09 2013.

[4] L. Mones, N. Bernstein, and G. Csányi, "Exploration, sampling, and reconstruction of free energy surfaces with gaussian process regression," *Journal of Chemical Theory and Computation*, vol. 12, no. 10, pp. 5100–5110, 2016. PMID: 27598684.

[5] K. Yao and J. Parkhill, "Kinetic energy of hydrocarbons as a function of electron density and convolutional neural networks," *Journal of Chemical Theory and Computation*, vol. 12, no. 3, pp. 1139–1147, 2016. PMID: 26812530.

[6] J. C. Snyder, M. Rupp, K. Hansen, L. Blooston, K.-R. Müller, and K. Burke, "Orbital-free bond breaking via machine learning," *The Journal of Chemical Physics*, vol. 139, no. 22, p. 224104, 2013.

[7] F. Fracchia, G. Frate, G. Mancini, W. Rocchia, and V. Barone, "Force field parametrization of metal ions from statistical learning techniques," *Journal of Chemical Theory and Computation*, vol. 14, 11 2017.

[8] K. Yao, J. E. Herr, D. W. Toth, R. Mckintyre, and J. Parkhill, "The tensormol-0.1 model chemistry: a neural network augmented with long-range physics," *Chem. Sci.*, vol. 9, pp. 2261–2269, 2018.

[9] E. Kim, K. Huang, A. Tomala, S. Matthews, E. Strubell, A. Saunders, A. Mccallum, and E. Olivetti, "Machine-learned and codified synthesis parameters of oxide materials," *Scientific Data*, vol. 4, p. sdata2017127, 09 2017.

[10] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, and A. Aspuru-Guzik, "The harvard clean energy project: High-throughput screening of organic photovoltaic materials using first-principles electronic structure theory," pp. 7001–, 02 2012.

[11] C. M. Handley and P. L. A. Popelier, "Potential energy surfaces fitted by artificial neural networks," *The Journal of Physical Chemistry A*, vol. 114, no. 10, pp. 3371–3383, 2010. PMID: 20131763.

[12] J.-P. Piquemal and K. D. Jordan, "Preface: Special topic: From quantum mechanics to force fields," *The Journal of Chemical Physics*, vol. 147, no. 16, p. 161401, 2017.

[13] K. N. Houk and F. Liu, "Holy grails for computational organic chemistry and biochemistry," *Accounts of chemical research*, vol. 50, no. 3, pp. 539–543, 2017.

[14] Z. E. Hughes, J. C. R. Thacker, A. L. Wilson, and P. L. A. Popelier, "Description of potential energy surfaces of molecules using fflux machine learning models," *Journal of Chemical Theory and Computation*, vol. 15, no. 1, pp. 116–126, 2019.

[15] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Big data meets quantum chemistry approximations: The machine learning approach," *Journal of chemical theory and computation*, vol. 11, no. 5, pp. 2087–2096, 2015.

[16] A. Fabrizio, B. Meyer, R. Fabregat, and C. Corminboeuf, "Quantum chemistry meets machine learning," *Chimia*, vol. 73, no. 12, pp. 983–989, 2019.

[17] J. B. O. Mitchell, "Machine learning methods in chemoinformatics," *Wiley interdisciplinary reviews. Computational molecular science*, vol. 4, no. 5, pp. 468–481, 2014.

[18] R. Xu and D. Wunsch, *Clustering*. IEEE series on computational intelligence, Hoboken: Wiley-IEEE Press, 1. aufl. ed., 2008.

[19] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[20] L. Sun, *Multi-label dimensionality reduction / Liang Sun, Shuiwang Ji, and Jieping Ye.* Chapman and Hall/CRC machine learning and pattern recognition series, CRC Press, 2014.