 doi:10.20944/preprints202005.0408.v1

Proposed Improvements For Automated Chemical Safety Evaluations Using In-Silico Techniques

Bryan Jordan[1]

Streamplate, 405/410 Elizabeth St, Surry Hills, NSW, 2010, Australia. Correspondence: bryan@streamplate.com

Abstract: The vastness of chemical-space constrains traditional drug-discovery methods to the organic laws that are guiding the chemistry involved in filtering through candidates. Leveraging computing with machine-learning to intelligently generate compounds that meet a wide range of objectives can bring significant gains in time and effort needed to filter through a broad range of candidates. This paper details how the use of Generative-Adversarial-Networks, novel machine learning techniques to format the training dataset and the use of quantum computing offer new ways to expedite drug-discovery.

Keywords: drug discovery; machine learning; in silico; pharmacology

Introduction

Computing provides researchers with attempts to virtualise biochemical interactions at a pace and volume that outperforms traditional in vitro or in vivo methods. Such technological advancements have made in-silico techniques increasingly effective in the drug-discovery phase, which can filter millions of molecules at a cost-efficient speed that would otherwise be unavailable. Virtual screening is, therefore, useful in simulating molecular interactions that can lead to a range of insights about molecular behaviour such as toxicity as shown by Wu and Wang (2018). With a resurgence in machine-learning (ML) due to advances in hardware and the dividends of networked research-communities coming into fruition in the form of enriched datasets, researchers can now correctly leverage in-silico methods. The most performant usage of ML in measuring toxicity come from Generative Adversarial Networks (GANs), but even then, these generally have provided sub-optimal results (Zhang et al. 2018). Much of this arises from the nature of how ML models exploit training data and then seek to maximise a return function — which is often over-simplified in the case of toxicity-based experiments. Such a fundamental reliance on training data shows how as it is the seed of many ML models, these models are compromised. Such problems become further compounded when they're expected to adequately navigate the chemical space while fulfilling a range of parameters involved in toxicity. Specifically, this review suggests improvements that focus on the training data, the network architecture and the usage of quantum computing to improve toxicity predictions. It is, therefore, the goal of this review to offer several proposals that improve ML techniques used in assessing toxicity.

History

One of the earliest phases of developing new drugs involves drug-discovery: a process where structures are identified that best bind to a drug target such as a particular protein or enzyme. It became evident amongst researchers that rather than using traditional assays to identify prospects, computers could simulate these experiments at much more optimal costs. In-silico techniques subsequently encroached on toxicity experiments in the wake of a growing number

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0408.v1

of datasets becoming available to researchers such as DeepChem, PyMol and the RCSB Protein Data Bank (Mayr et al. 2015). Virtualisation, however, meant that computational representations of molecules have been somewhat compressed and simplified, hindering the precision at which they can effectively virtualise the real biochemical realities of the properties they are examining. For instance, a one-dimensional molecular representation will typically only describe the molecular weight, solubility, charge and number of rotatable bonds atom type as highlighted by Karim et al. (2019). Two-dimensional representations take into account the graph of covalent and aromatic bonds but do not refer to spatial coordinates. Both dimensions naturally fail to account for three-dimensional space and their subsequent evolution across time.

Historically, in silico simulations have used these data representations for two components - generative and predictive simulations. Generative tasks lead to the synthesis of virtual molecules that are reproducible with a chemical formula that's returned by the computer, which is usually in the ORGANIC or SMILES format (Noorden, 2018). Predictive tasks compute the chemical properties of these generated virtual molecules like activity, toxicity and water solubility. Generative models such as GANs or auto-encoders have been used with varying degrees of success by using different physicochemical properties in their virtual molecular representations.

However, these systems collectively suffer from a range of problems, including non-convergence and mode-collapse. Underlying these problems is that while large training sets have been released, there is an inherent bias from researchers to pursue only profitable prospects which makes these datasets skewed mostly towards false positives (Karim et al., 2019. Researchers have catalogued only 130 million organic and inorganic substances with the Chemical Abstract Service, representing a mere fraction of the purported potential of 1063 molecules in the chemical space. This alludes to not only the limitation of the training data that is being used in ML techniques, but also the potential for undefined physiochemical insights that may be influencing results.

State of the Art

Standardised datasets give ML models the locality necessary to converge a solution towards a sensible result. In this way, ML-based predictions for protein folding are symptomatically similar to the scope of toxicity in that the domain of candidates is so large that intelligent mechanisms are needed to sift though high-dimensional information properly. ProteinNet represents one such dataset which provides a protein sequence, structures as secondary or tertiary along with supportive meta-details such as training, validation and test splits (Pu et al. 2019). Experiments using ProteinNet have highlighted how, despite this rich dataset, there remain inherent biochemical properties which can compromise the sensibility of results. For instance, as proteins share an evolutionary relationship, this virtually guarantees that every protein is somewhat related to another. Worse, Hui (2018) describes how a computer can interpret sequences of categorical variables of two identical proteins as identical proteins which limits the model's delineation between proteins.

Cutting edge ML methods for toxicity prediction have transitioned from typical reinforcement-learning models to more comprehensive GANs. Simpler ML models (Jeon, 2019) have been used with less success as seen in the cascade model of molecular mapping descriptors of compounds to their respective assay results which included:

1. ALogP,
2. Polarisability,

 doi:10.20944/preprints202005.0408.v1

3. Number of rotatable bonds,
4. Polar surface area,
5. Number of hydrogen bond donors and acceptors,
6. Molecular weight.

Through these failings, researchers identified that multitask ML models aiming to accurately model biophysical properties require information to be shared between datasets for successful performance (Zhang et al. 2018). Subsequently, the encoder/decoder duality in GANs allows for the innate details of datasets to be translated across separate models which theoretically aid with the construction of high-dimensional representations of the target. This has ensured that by encoding molecules in a SMILES representation, GANs have been used in the first stage of the drug discovery process by generating desired biological activity and the generation of non-fullerene electron acceptors for organic solar cells. GANs provide researchers with the ability to synthesise compounds than simply discriminate them as traditional ML methods allow. Even then, results from Karim et al. 2019 have shown there is typically a considerable amount of invalid molecules and even for those adequately structured, there is often low variance or a failure to reach all the necessary parameters.

Attempts to control the shortcomings of GANs have relied on tactical changes that have mostly failed to overcome the naturally occurring architectural flaws of these models. Typical tactics used to improve GAN performance as described by Hui (2018) include:

1. Normalising inputs between –1 and 1,
2. A modified loss function to optimise (G) by using max(logD) as the first formulation has vanishing gradients early on,
3. Using a spherical Z and interpolating via a great circle rather than straight lines,
4. Constructing mini-batches for real and fake-data,
5. Avoid using sparse gradients such as ReLu and MaxPool, and instead of using LeakyReLu.
    a. For downsampling use:

i. Average pooling, cov2d and stride,

    b. For upsampling use:

i. Pixel shuffle and convtranspose2d and stride,

6. Use soft and noisy labels using floating numbers instead of integers and sporadically flip labels during training,
7. Use SGD for discriminator and ADAM for generator,
8. Add noise to inputs and add gaussian noise to every layer of generator,
9. Use tanh as the last layer of the generator output,
10. Use two discriminators for maximising discrimination of ‘real’ data and ‘fake’ data.

Proposed Methods for Improvement

Each of the following methods are designed to support GAN-based techniques at different stages of the model.

Training Data: Iterative Complexity

As ML models leverage insights derived from datasets, this seed must be properly configured to germinate correctly. Unfortunately, the datasets used today for toxicity prediction are skewed and consequently misguide ML models. With this in mind, the idea is then to generate instead a separate model that can accurately generate the periodic table and iteratively get this model to regenerate datasets modestly.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0408.v1

In this manner, this type of unsupervised learning follows the trend of impressive results of unsupervised learning that was catapulted into the spotlight after the accomplishments of Alpha Zero. Rather than developing a specialised, high-performant model in a tailored domain — the model should self-teach itself from first-principles and learn so by generating molecules that not only are bound to biophysical laws but are also elements of the Periodic Table.

The reason for emulating the Periodic Table is that we want the model to be initially configured to the breadth of molecules available. The second stage of the model's evolution will be to construct compounds consisting of two-molecules, three-molecules and continuing with an increasing molecular count. Obviously, the way these models evolve is subject to further tailoring, with each iterative discriminator able to be progressively advance the complexity of the preceding model. Experimental results will provide insights into the most optimal type of configurations between different classifications of molecules. Subsequently, this lends itself also to be optimised by ML models which optimise slate configurations.

As for the type of molecular configuration, four-dimensional representations should be the golden standard for in-silico representations as counterpart techniques used for in-vivo or in-vitro methods already exploit this by default. This, of course, further inflates the computational space. In this way, to not only compress but predict how molecules will behave in time, a Monte-Carlo Search Tree (MCST) should be utilised to store results as developed by Silver et al. (2017). Sifting through all possible combinations is impossible, and so the searching algorithm requires a separate model that is in many ways, a proxy for the loss function in the principal model that generates the training data.

Subsequently, 'iterative complexity' means using a dataset for training a separate model that in turn, will produce the training data for the principal model. Datasets that have been manually composed should always be as imperfect representations of their problem domain. To counter this, three separate ML-based models are proposed to generate the training data.

1. Slate model:
    a. Organises how the succeeding, 'Principal' model should evolve once it learns the Periodic Table.
2. Principal model:
    a. Generates a training set in the form of a pre-configured set of weights that best represent molecules and compounds.
3. Searching model:
    a. A model for searching through a MCST that holds the molecular behaviour interactions.

In this way, it is hoped that by automating the training data generation, innate relationships that govern molecules and compounds can be represented appropriately and counter the void created by today's datasets.

Network Architecture: Concentrated Generalisers for Multi-Objective Completion

Deep networks have proven successful for a range of discrete challenges, namely in computer vision and natural language processing but have failed to generalise to more abstract concepts or less-defined classifications. In this way, the failure of many GANs that employ deep architectures to predict toxicity understandably fits the expected outcome of these models.

 doi:10.20944/preprints202005.0408.v1

The theoretical understandings of these failures aren't correctly understood, but experimentally these type of models have failed to deliver the results necessary for high-performance toxicity predictions. Deep networks are not reflective of the 'Universal Approximation Theorem' (Kumar, 2019) which describes how a recurrent feed-forward network with just a single hidden layer can approximate continuous functions on compact subsets of real numbers. Pragmatically, the number of neurons required for this hidden layer may outnumber the bounds of practical implementations which implies a ceiling may exist within practical applications.

As modern architectures have failed to reach a substantial level of general performance, the case for alternative architectures becomes evident. A proposed alternative then is to have a series of refined networks operate in unison with one another, each sharing the same hidden layer. This type of network shall be referred to as a form of a 'Concentrated Generaliser' (CG), which in essence, is about capturing the unifying similarities between varying learnt domains of knowledge.

CGs follows recent research from Principal Component Analysis which highlights how a few features capture most variance. Assuming the usage of non-linear functions, all functions within the hypothesis space are therefore theoretically possible. CGs can develop by aggressively pruning the network through Taylor expansion as developed by Nvidia (Anwar, 2015). This marks the beginning of a CG, a highly attenuated network that is hyper-localised to particular stimuli. The next step is to create another network that is tailored to a separate domain and then pruned upon successful performance.

To concentrate these two independent nodes, the network needs to learn how to successfully merge and appropriately adjust the weights of the hidden layer to be able to reach an adequate standard of performance. This standard can be defined as a measure of processing time or the number of changes, but it is the system's goal to minimise the number of changes necessary to reach comfortable performance.

As can be seen, this process can be repeated countless times with each iteration bridging together two independently pruned networks. It may be initially advantageous to allow these systems to algebraically add their hidden layer to the net total of hidden nodes in the network until CGs performance can be better analysed.

Quantum Computing

Recent developments in quantum computing have brought the technology closer towards the realisation of Richard Feynman's initial vision of using the machines for 'quantum physics and chemistry simulations'. In this way, biochemists need to begin adapting their skillsets to utilise better this hardware that is a total subversion to the typical methodologies employed for *in silico* experiments. In the past few years and even recently, there has been steady progress on the engineering frontier of quantum computing, including the realisation that standard microfabrication facilities (Francis, 2015) can be used to create quantum processor units and Google's recent 'quantum supremacy' breakthrough (Arute, 2019).

The ever-improving state of quantum computers places greater emphasis on the need for biochemists to begin understanding how they can adapt their problem domains to that of quantum computers. Over fifty quantum algorithms have been identified as seen in the Quantum Algorithm Zoo (2019), with the Harrow-Hassidim-Lloyd (HHL) algorithms promising an exponential speed-up against classical computers. In the way that machine learning algorithms offer biochemists an opportunity to intelligently sift through millions of

 doi:10.20944/preprints202005.0408.v1

data-points in much faster efficiency than a naïve, brute-force approach — a quantum computer running quantum algorithms for machine learning has obvious benefits (Niu, 2019).

The significance of quantum algorithms can be seen in how HHL algorithms can solve a system of linear equations in log time. Granted, that a number of pre-conditions are necessary to enable this speed-up — but such challenges exist as engineering problems and will undoubtedly become resolved in time. Similarly, a polynomial-time quantum algorithm was proposed for estimating certain topological features of data, most notably Betti numbers (Lloyd et al. 2016) which count the number of holes and voids across the multiple dimensions of a scatter plot.

Just as how pharmacology corporations reap lucrative rewards from the IP following a successful drug discovery, quantum computers can become the machines that can generate such lucrative IP in a much more efficient manner (Solenov, 2019). The need for biochemists, computer scientists and electrical engineers to begin the proper conversations across universities and research centres to start developing these quantum computers is only becoming more immediate. There exists substantial financial and scientific opportunities by properly leveraging these computers in the domain of chemistry.

Conclusion

Much of *in silico* methodologies remain ripe for disruption due in part to their reliance on cutting-edge technologies. It is therefore critical that those using them ensure their skillset remains relevant. Machine learning undoubtedly offers ample opportunity for biochemists to tackle toxicity in a far more efficient manner than before. However, new technologies are approaching that if successful, may even undermine classical machine learning approaches.

References

Anwar S, Hwang K, Sung Wonyong (2015). Structured Pruning of Deep Convolutional Neural Networks. Journal on emerging Technologies in Computing Systems – Special Issue on Hardware and Algorithms for Leaning On-a-chip. Volume 13, Issue 3: 32.

Anon (2018). ProTox-II – Prediction of Toxicity of Chemicals. Available at: http://tox.charite.de/protox_II/ (accessed 3rd October 2019).

Anon (2019). Polo Club of Data Science. Available at: https://poloclub.github.io/ganlab/ (accessed 5th October 2019).

Anon (2019). Quantum Algorithm Zoo. Available at: http://quantumalgorithmzoo.org/ (accessed 7th October 2019).

Arute F, Arya K, Babbush R, Bacon D, Bardin J.C., Barendas, et al. (2019). Quantum supremacy using a programmable superconducting processor. Available at: https://www.nature.com/articles/s41586-019-1666-5.pdf (accessed 30th October 2019.

Brownlee J (2019). How to identify and Diagnose GAN Failure Modes. Available at: https://machinelearningmastery.com/practical-guide-to-gan-failure-modes/. (accessed 1st October 2019).

Brownlee J (2019). How to Explore the GAN Latent Space When Generating Faces. Available at: https://machinelearningmastery.com/how-to-interpolate-and-perform-vector-arithmetic-with-faces-using-a-generative-adversarial-network/ (accessed 4th October 2019).

Francis H (2015). Australian researchers make quantum computing breakthrough, paving way for world-first chip. Available at: https://www.smh.com.au/technology/australian-researchers-make-quantum-computing-breakthrough-paving-way-for-worldfirst-chip-20151005-gk1bov.html (accessed 8th October 2019).

Hui J (2018). GAN – Why is it so hard to train Generative Adversarial Networks!. Available at: https://medium.com/@jonathan_hui/gan-why-it-is-so-hard-to-train-generative-advisory-networks-819a86b3750b. Accessed (3rd October 2019).

Hui J (2018). GAN – Ways to improve GAN performance. Available at: https://towardsdatascience.com/gan-ways-to-improve-gan-performance-acf37f9f59b. (accessed 3rd October 2019).

Hui J (2018). GAN – A comprehensive review into the gangsters of GAN (part 2). Available at: https://medium.com/@jonathan_hui/gan-a-comprehensive-review-into-the-gangsters-of-gans-part-2-73233a670d19. (accessed 1st October 2019).

Hui J (2019). GAN – What is wrong with the GAN cost function? Available at: https://medium.com/@jonathan_hui/gan-what-is-wrong-with-the-gan-cost-function-6f594162ce01. (accessed 2nd October 2019).

Jeon M (2019). Tensorflow-GAN: Basics of Generative Adversarial Networks. Available at: https://medium.com/@fabulousjeong/gan-with-tensorflow-basics-of-generative-adversarial-networks-d71bb9a4cae2 (accessed 4th October 2019).

Karim A, Mishra A, Hakim Newton M. A., Sattar A (2019) Efficient Toxicity Prediction via Simple Features Using Shallow Neural Networks and Decision Trees. ACS Omega 2019 (4) pp 1874 – 1888.

Kristiadi A (2019). Generative Adversarial Nets in TensorFlow. Available at: https://wiseodd.github.io/techblog/2016/09/17/gan-tensorflow/. (accessed 2nd October 2019).

Kumar N (2019). Illustrative Proof of Universal Approximation Theorem. Available at: https://hackernoon.com/illustrative-proof-of-universal-approximation-theorem-5845c02822f6 (accessed 4th October 2019).

Lloyd S., Garnerone S. (2015). Quantum algorithms for topological and geometric analysis of data. Nature, 7, 10138.

Mayr A, Klambaur G, Unterthiner T, Hochreiter S (2015). Deep Tox: Toxicity Prediction Using Deep Learning. Frontiers in Environmental Science: 10.3389/fenvs.2015.00080

Niu M. Y., Boixo S., Smelyanskiy V. N., Neven H. (2019). Universal quantum control through deep reinforcement learning. Quantum Information 5: 33.

Pu L, Naderi M, Liu T, Wu HC, Mukhopadhyay S, Brylinski (2019) eToxPred: A machine leanring-based approach to estimate the toxicity of drug candidates. BMC Pharmacology Toxicity. 2019 Jan 8; 20(1);2: doi: 10.1186/s40360-018-0282-6.

Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X (2016). Improved Techniques for Training GANs. Advances in neural Information Processing Systems 29 (NIPS 2016).

Silver D, Huber T, Schrittwieser J, Antonoglou I, Lai M, Guez A, Lanctot M ,et al. (2017). Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. Available at: https://arxiv.org/pdf/1712.01815.pdf (accessed 3rd October 2019).

Solenov D., Brieler J., Scherrer J. F. (2019). The Potential of Quantum Computing and Machine Learning to Advance Clinical Research and Change the Practice of Medicine. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6205278/pdf/ms115_p0463.pdf (accessed October 9th 2019).

Van Noorden R (2018). Software beats animal tests at predicting toxicity of chemicals. Available at: https://www.nature.com/articles/d41586-018-05664-2#ref-CR1 (accessed October 5th 2019).

Wu Y, Wang G (2018). Machine Learning Based Toxicity Prediction: From Chemical Structural Description to Transcriptome Analysis. International Journal of Molecular Sciences: 19(8):2358.

Zhang L, Zhang H, Ai H, Huan H (2018). Applications of Machine Learning Methods in Drug Toxicity Prediction. Current Topics in Medicinal Chemistry. 18(12).