December 13, 2024

Dreaming Molecules from Latent Ichor with Equivariant and Recurrent Graph Neural Networks

Jon Potter, Max Shushkovsky, Shivank Sadasivan

Specific Aims

The objective of this project is to design novel molecules that can selectively activate androgen receptors using graph based deep learning. We aim to train an Equivariant Graph Neural Network (HEGNN) to learn molecular embeddings from SMILES strings and decode these embeddings into valid molecular graphs using a Graph Recurrent Neural Network (GRNN) and output them as SMILE strings. By refining model parameters and enforcing chemical constraints, we seek to improve the validity of generated molecules while addressing the significant class imbalance in the Tox21 dataset. The generated molecules will be evaluated for chemical plausibility, AR activation potential, toxicity, and synthesizability, creating a robust and scalable framework for drug discovery.

Background

The Tox21 Challenge was an initiative that has transformed toxicity testing by integrating high throughput screening and computational modeling to predict the toxicological effects of chemicals. This collaboration of scientists aims to develop innovative methods to assess the health impacts of chemicals by understanding how they disrupt biological pathways, potentially leading to adverse health effects. The Tox21 2014 Challenge marked a significant milestone, emphasizing the assessment of biological pathway disruptions by chemicals. For this study we

looked at their impact on nuclear receptors such as androgen receptors (AR). AR mediates the effects of androgens by regulating gene expression, playing a dual role in various diseases. In breast cancer, AR can act as a tumor suppressor in some contexts while promoting tumor growth in others. These dynamic interactions underscore AR's complexity, particularly in endocrine therapy, where targeting AR alongside other nuclear receptors is crucial. Beyond cancer, AR plays a significant role in bone health by regulating osteoblast and osteoclast differentiation, ensuring skeletal integrity. Recent studies have linked unusual AR expression to head and neck cancers, highlighting its broad impact across physiological systems and disease pathways.[1][2][3][4]

The Tox21 datasets provide a comprehensive resource for investigating the interactions of various compounds with AR. Each compound is represented by its SMILES (Simplified Molecular Input Line Entry System) notation, facilitating effective computational modeling and enabling cheminformatics analyses. The datasets specify the activation status for each compound, with unique compound IDs indicating whether a compound activates (1) or deactivates (0) the AR. This information is critical for the objectives of our project, as it allows for targeted exploration of compounds that can modulate AR signaling. The detailed activation labels enable the development of predictive models to identify novel compounds with desired activities, enhancing the potential for discovering selective AR modulators[7]

The Tox21 dataset comprises 8,982 compounds for AR, of which 380 are labeled as activating and 8,602 as deactivating in terms of their effect on AR. These compounds span various classes, including organic, ionic, and organometallic molecules. However, after preprocessing to focus on the main connected component of the drug molecules, the dataset contains 5,726 deactivating

compounds and 306 activating compounds. This significant disparity between deactivating and activating compounds presents challenges in developing balanced and effective computational models, particularly concerning class imbalance in machine learning algorithms.[7]

To facilitate the computational analysis of these compounds, a variety of atomic and edge level features were generated using RDKit[6], a cheminformatics tool. These features include:

- **Atomic Features:** Degree, Formal charge, Number of radical electrons, Hybridization, Aromaticity, Atomic mass, Atomic radius, Van der Waals radius, Chirality, Electronegativity, and whether the atom is in a ring.
- **Edge Features:** Bond type, Bond angle, Bond length, Rotatability, Polarity, whether the bond is in a ring, Conjugation, and Stereochemistry.

The inclusion of both atomic and edge level characteristics ensures that subtle molecular variations influencing AR interactions are captured, aiding in the design of molecules that can specifically activate or deactivate AR. These features were critical for capturing subtle molecular variations influencing AR interactions. The SMILES notation served as both input and output in this workflow, allowing iterative molecular generation and evaluation to refine predictions. The combination of RDKit generated features and custom functions provided a framework for designing and analyzing molecules that can specifically activate or deactivate AR.

Significance

This project aims to advance drug discovery by utilizing graph based deep learning to generate selective AR-activating molecules, focusing on the generation of organic and chemically valid compounds, both toxic and non-toxic. By integrating HEGNN and GRNN, the study

demonstrates the application of advanced graph neural network architectures in molecular design, offering a novel approach to addressing complex challenges in drug discovery. The project also seeks to improve therapeutic design while highlighting and addressing key limitations in existing datasets, such as class imbalance.

Experimental Design

Equivariant Graph Neural Networks

A key component to modeling molecules as graphs for representation learning purposes is the equivariance property. Should a molecule be rotated, translations, reflected, or permuted, its resulting embedding from a graph neural network should still be the same; the molecular properties do not change as a result of this transformation. Thus, new equivariant graph neural networks (EGNNs) needed to be developed. In the seminal paper on EGNNs, the following relevant modifications to standard message passing for the needed equivariant properties are explained, along with a spatial coordinate update (Satorras, Hoogeboom, and Welling). Let h_i^l refer to the node features of h at layer l, x_i^l refer to the spatial positions of the nodes at layer l, and Φ be a nonlinear learnable function (ie a neural network). By using the squared distance between spatial coordinates in (3), message passing is equivariant to translations and. In the spatial coordinate update in (4) the difference term is a representation of the relative position vector difference, so this is equivariant to rotations. When multiplying by a rotation matrix, the coordinates will get updated accordingly, by their relative positions. Reflection equivariance is similarly ensured by these update rules as well. Equation (5) ensures permutation equivariance is upheld because the order in which a node's neighbors pass their message does not matter. As a

result, an EGNN is used to embed the molecules into a higher dimension and then perform baseline classification (whether the molecule is activating the nuclear androgen receptor or not).

$$m_{ij} = \Phi_e(h_i^l, h_j^l, \left|\left|x_i^l - x_j^l\right|\right|^2, a_{ij}) \quad (1)$$

$$x_i^{l+1} = x_i^l + C\sum_{j\neq i}(x_i^l - x_j^l)\Phi_x(m_{ij}) \quad (2)$$

$$m_i = \sum_{j\neq i} m_{ij} \quad (3)$$

Recurrent Graph Neural Networks

In order to reconstruct graphs from embeddings, we crafted a Graph Recurrent Neural Network from scratch. The Graph Recurrent Neural Network (GRNN henceforth), is designed as a composition of two sub-models, a Node Recurrent Neural Network (NRNN) and an Edge Recurrent Neural Network (ERNN). The NRNN predicts, given a hidden state (or graph embedding if at the first step) and a map of allowed atomic elements, the next most likely atom to be included in a molecule. The ERNN is then run between the newest atom and all previously predicted atoms to predict whether or not a bond exists between the pair.

The process then repeats with the updated hidden state, until either the NRNN predicts a special End of Sequence (EOS) token, marking the end of the atom, or the maximum number of atoms have been added to the molecule. Through tuning, we found that setting the maximum number of atoms to 20 resulted in realistic and accurate predicted molecules.

The GRNN then uses the predicted nodes and edges to reconstruct a graph representing a molecule. Often, this results in multiple disconnected components, so only the largest connected component is retained, with all other components discarded. What remains is likely to be a valid molecule, but we use RDKit to verify chemical validity before accepting the generated molecule. If it fails this check, we discard the molecule, select a new initial embedding, and try again.

NodeRNN

The Node Recurrent Neural Network (NRNN) is responsible for generating the sequence of atoms in the molecule. At each sequence/"time" step t, the NRNN takes in a hidden state from the previous step h_{t-1} (or, if it is the first step, the initial graph embedding h_0) and predicts the next atom to add to the molecule.

Formally, the NRNN updates its hidden state and outputs using the following equations:

1. **Input Embedding:**

$$x_t=\text{Embedding}(y_{t-1}) \quad (4)$$

Where y_{t-1} is the atomic number index predicted at the previous time step.

2. **Hidden State Update**

$$h_t=\text{GRU}(x_t,h_{t-1}) \quad (5)$$

Where GRU denotes a classic Gated Recurrent Unit (Cho 2014).

3. **Atom Prediction**

$$o_t=Wh_t+b \quad (6)$$

$$y_t=\text{softmax}(o_t) \quad (7)$$

Where W and b are learned parameters, and y_t represents the probability distribution over possible atoms.

At each "time" step, the NRNN predicts the next atom by sampling from the probability distribution y_t. The sampled atom's embedding and updated hidden state is then used as input for the next time step. This process continues until an EOS token is predicted or the maximum number of atoms is reached.

To incorporate chemical knowledge, we use a map of possible elements represented by one-hot encoding. This encoding maps each atomic number to a unique index, ensuring that the NRNN predicts valid atoms.

EdgeRNN

After the NRNN predicts the next atom, the Edge Recurrent Neural Network (ERNN) determines how this new atom connects to the existing molecular structure. For each existing atom j (where $j=1,2,...,t-1$), the new atom t and atom j.

The ERNN operates as follows:

1. **Initialization**

$$s_{t,0}=h_t \quad (8)$$

Where $s_{t,0}$ is the initial hidden state for the ERNN at time t.

2. **Edge Prediction for Each Atom:**

For each existing atom j:

$$e_{t,j}=\text{GRU}(e_{t,j-1},s_{t,j-1}) \quad (9)$$

$$p_{t,j}=\sigma(Ue_{t,j}+c) \quad (10)$$

Where $e_{t,j}$ is the hidden state for the bond between atom t and atom j, U and c are learned parameters, and σ is the sigmoid function.

3. Bond Decision

A bond is predicted to exist if $p_{t,j}$ exceeds a certain threshold or by sampling from the probability. To help ensure chemical validity, we incorporated valence constraints during bond prediction. Each atom has a maximum allowed valence based on its chemical properties. Before adding a bond, we check whether adding it would exceed the valence of either atom. If it would, we skip the bond addition. Together, the NRNN and ERNN were trained on the activating ligands from the dataset using Cross-Entropy Loss and Binary Cross-Entropy Loss respectively, achieving a 98% reduction in loss over its training.

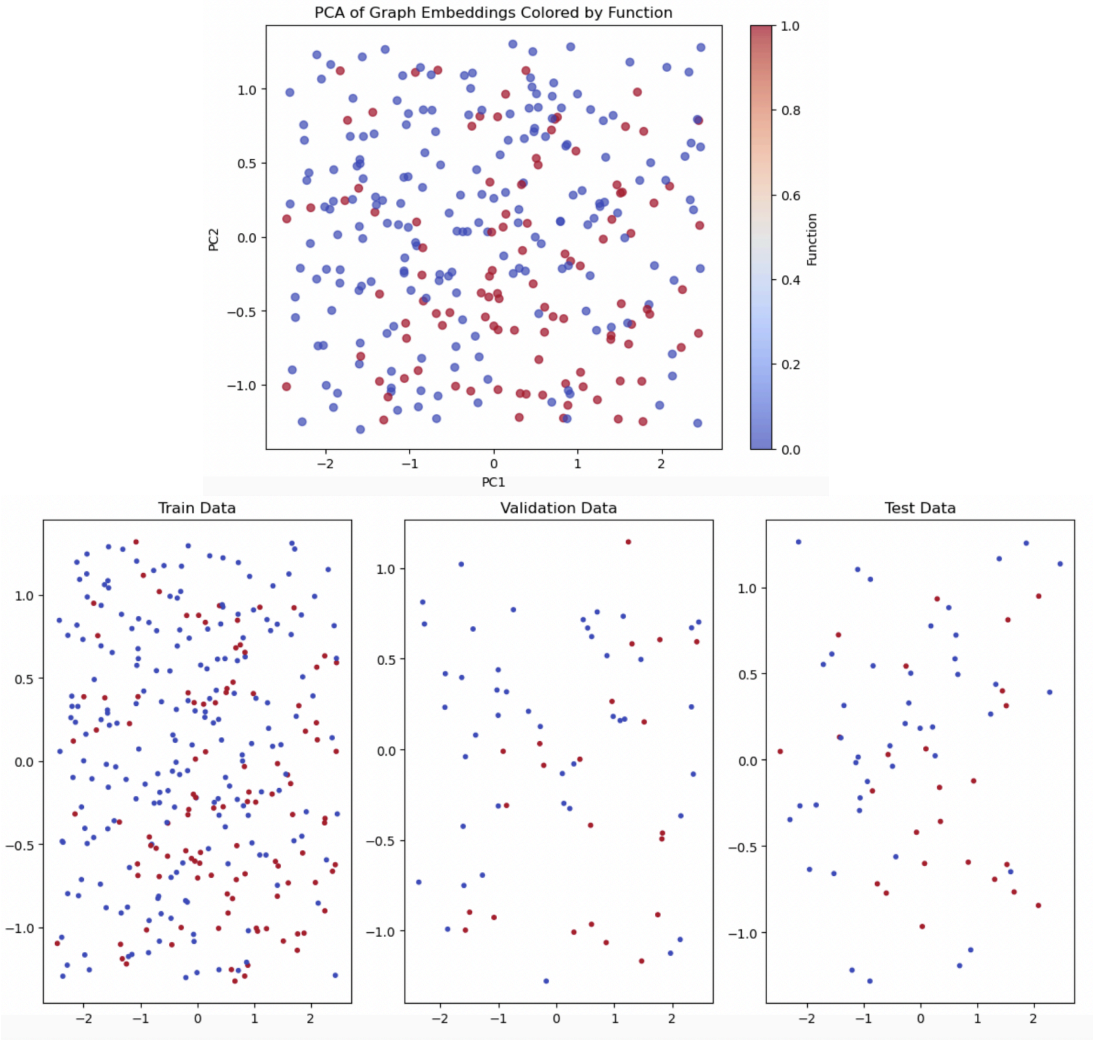
Methodology

An outline of the methodology is presented. Node and edge features were added to the molecules and processed into graph objects. They were split into training, validation, and testing, and passed into a higher order equivariant graph neural network (HEGNN) for two tasks: embedding the molecules into a higher dimension and adding a classification head to train the model to predict if the molecule is AR activating or not. Then, the AR activating embeddings are used for GRNN to learn from. The GRNN samples from nearby the centroid of this higher dimensional space (with random perturbations) and learns to generate a new graph following various molecular and structural properties. Afterwards, the graph is processed into a molecule, converted into a SMILES string, evaluated for validity as a molecule, and added to a set. This valid molecule generation is repeated 100 times, and ultimately, added back to the training set. The HEGNN is again trained on this "better" balanced dataset to see if testing accuracy and

AUC improves, which would support the evidence of the novel molecules truly being AR-activating.

Results

The performance of the HEGNN in embedding molecules into a higher dimensional space for AR activation classification was suboptimal. PCA visualization of the graph embeddings revealed overlap between AR-activating and non-activating molecules, indicating that the embeddings were not well separated for effective classification. This lack of separation likely



impacted the model's ability to accurately differentiate between the two classes and also due to class imbalance, highlighting the need for improved embedding strategies.

In the first iteration of molecule generation, 80 SMILES strings were produced, but only 20 were valid when processed into molecules using RDKit. To address this limitation, several model parameters were tuned, including increasing the number of hidden layers from 32 to 128, reducing the number of training epochs to mitigate overfitting, and incorporating additional edge and node features to provide better structural representations of molecules. These changes significantly improved the percentage of valid molecules generated. Additionally, a custom function was implemented to enforce a bias toward organic compounds by restricting the atomic distribution within generated molecules. This function ensured that 99% of the atoms in each compound were organic (e.g., carbon, hydrogen, nitrogen, oxygen, phosphorus), shifting the output distribution away from invalid organometallic compounds, which had been predominant in earlier iterations, toward simpler and more chemically plausible organic molecules.

Initially, the model favored metallic compounds, likely due to the high valency of metals such as aluminum and zinc. However, these compounds were structurally invalid. The introduction of constraints shifted the generation process toward organic molecules. Nitrogen-containing compounds were particularly notable, as they were more planar and structurally representative compared to carbon-based molecules, which tended to be more complex due to their single-bond properties. Despite these improvements, the model exhibited a notable limitation: it could only generate single bonds, which restricted the diversity of the generated compounds.

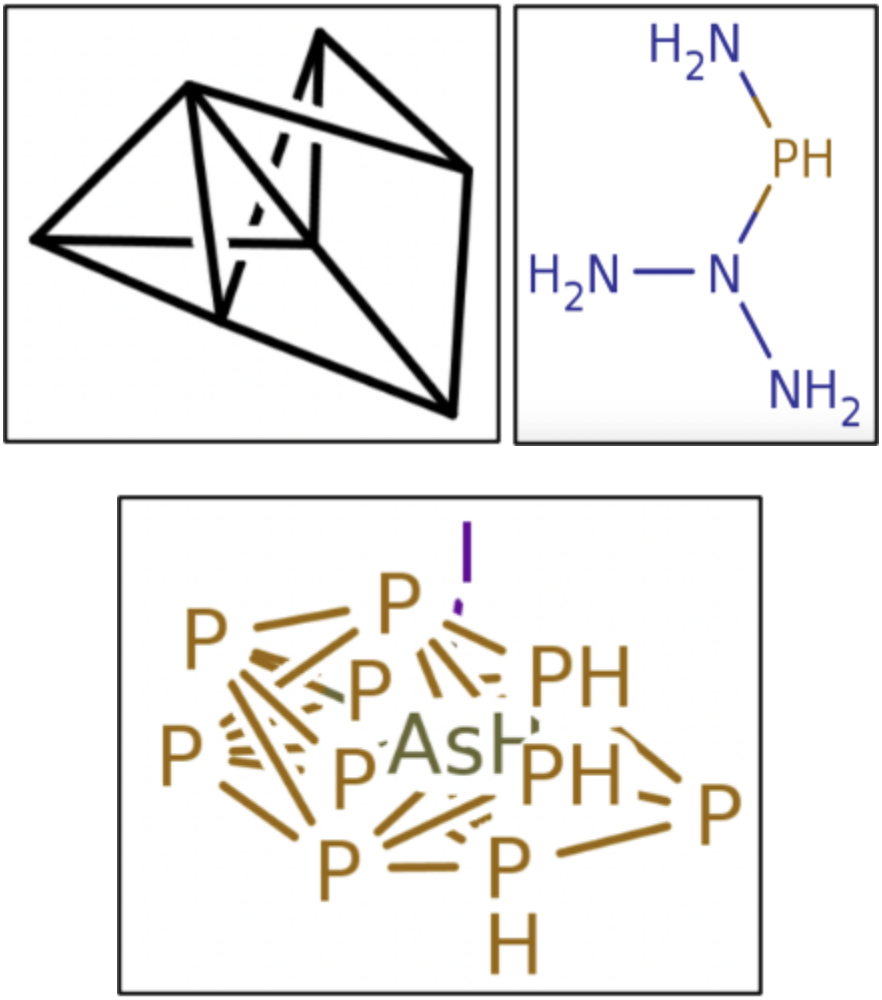


Fig. Complex carbon compound (C12C3C4C15C1C24C315), Simple and structurally valid Nitrogen chains (NPN(N)N), Invalid Complex Organo–metallic compound.

These refinements enhanced the model's ability to produce valid chemical structures. After parameter tuning and applying the organic compound restriction, 100 SMILES strings were generated, of which 56 were valid. While this represented a significant improvement in valid molecule generation, challenges remained in ensuring that the generated molecules were both chemically diverse and functionally relevant.

However, incorporating the "valid" generated molecules into the training set led to a decline in model performance. Initially, the model achieved a testing accuracy of 0.70 and an AUC of 0.76. After integrating the augmented dataset, the testing accuracy dropped slightly to 0.66, and the AUC decreased to 0.60. This decline suggests that the inclusion of generated molecules introduced noise or discrepancies into the dataset, potentially due to structural or molecular inconsistencies in the newly generated compounds.

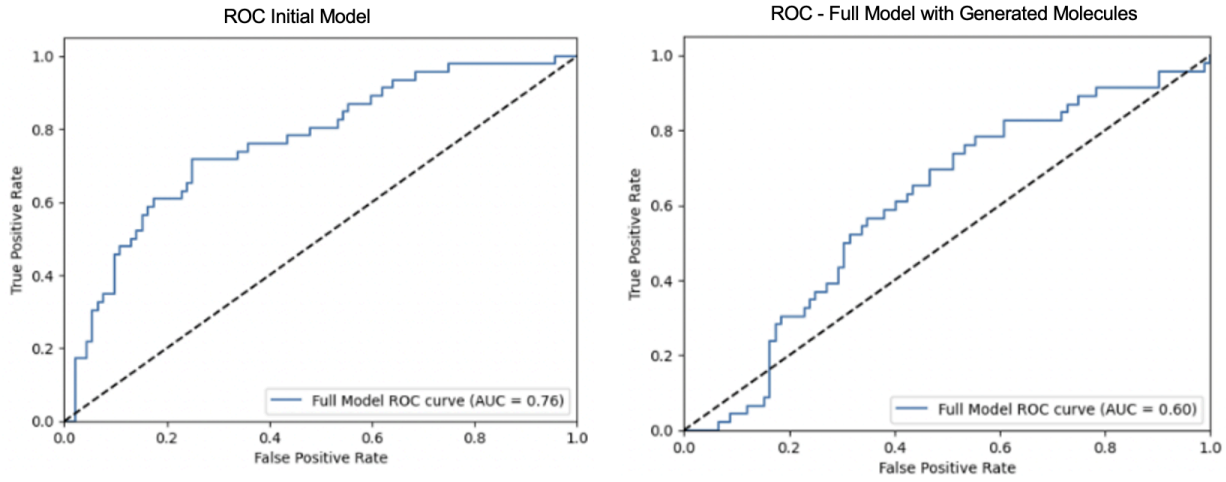


Fig. (Left) Initial model AUC and (Right) Full model AUC after adding generated molecules

To further evaluate the practical viability of the generated molecules, eToxPred, a machine learning based tool designed to estimate the toxicity and synthetic accessibility of small organic compounds was used. It utilizes algorithms trained on molecular fingerprints to evaluate drug candidates.[9] These assessments provided insights into the real-world applicability of the generated molecules and their potential for further optimization.

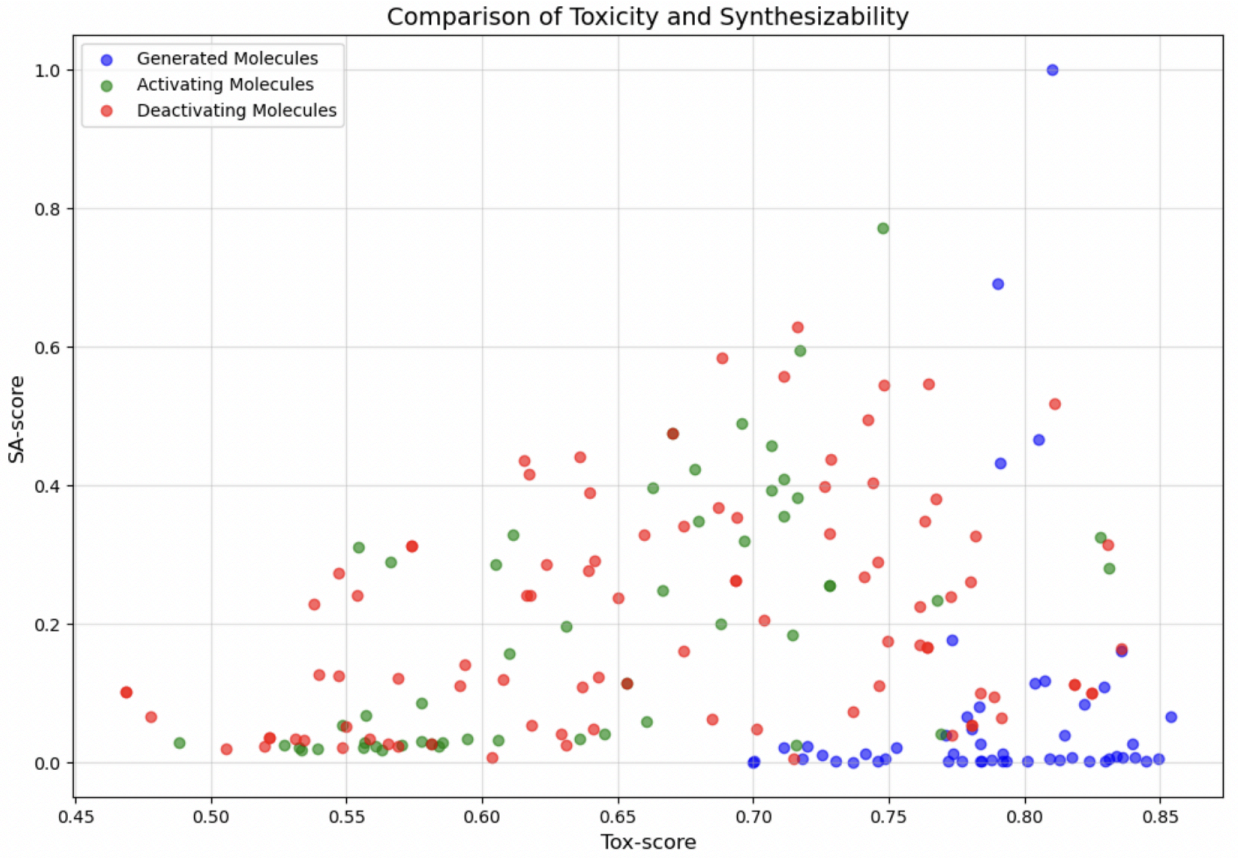


Fig. eToxPred output (Comparison of Toxicity and Synthesizability of generated activating molecules with randomly chosen training dataset)

Despite the observed challenges, the results indicate progress in generating valid organic molecules and underscore the need for additional refinements to improve classification performance and molecule diversity. There were limitations to the generation that likely had a large influence on these results, such as the enforcement of only single bonds and the difficulty in capturing whole molecule structural properties. Valency was an enforced property, but many others play an important role too. However, given the massive search space for molecule generation, along with the ability to overcome these limitations with more robust models, there is a lot of promise for graph approach learning in drug discovery/design.

References

1. Dai, C., & Ellisen, L. W. (2023, May 8). *Revisiting androgen receptor signaling in breast cancer*. The oncologist. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10166165/
2. Kensler, K. H., Regan, M. M., Heng, Y. J., Baker, G. M., Pyle, M. E., Schnitt, S. J., Hazra, A., Kammler, R., Thürlimann, B., Colleoni, M., Viale, G., Brown, M., & Tamimi, R. M. (2019, February 22). *Prognostic and predictive value of androgen receptor expression in postmenopausal women with estrogen receptor-positive breast cancer: Results from the Breast International Group Trial 1–98 - breast cancer research*. BioMed Central. https://breast-cancer-research.biomedcentral.com/articles/10.1186/s13058-019-1118-z
3. Manolagas, S. C., O'Brien, C. A., & Almeida, M. (2013, September 17). *The role of estrogen and androgen receptors in bone health and disease*. Nature News. https://www.nature.com/articles/nrendo.2013.179
4. Wei, L., Gao, H., Yu, J., Zhang, H., Nguyen, T. T. L., Gu, Y., Passow, M. R., Carter, J. M., Qin, B., Boughey, J. C., Goetz, M. P., Weinshilboum, R. M., Ingle, J. N., & Wang, L. (2023, February 3). *Pharmacological targeting of androgen receptor elicits context-specific effects in estrogen receptor-positive breast cancer*. Cancer research. https://pmc.ncbi.nlm.nih.gov/articles/PMC9896025/
5. You, C.-P., Tsoi, H., Man, E. P. S., Leung, M.-H., & Khoo, U.-S. (2022, December 5). *Modulating the activity of androgen receptor for treating breast cancer*. International journal of molecular sciences. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9739178/
6. Chen, Nan et al. Androgen receptor agonism in advanced oestrogen receptor-positive breast cancer. The Lancet Oncology, Volume 25, Issue 3, 269 - 270
7. *Papers with code - tox21 benchmark (molecular property prediction)*. The latest in Machine Learning. (n.d.).
8. RDKit: Open-source cheminformatics. http://www.rdkit.org
9. Pu, L., Naderi, M., Liu, T. et al. eToxPred: a machine learning-based approach to estimate the toxicity of drug candidates. BMC Pharmacol Toxicol 20, 2 (2019). https://doi.org/10.1186/s40360-018-0282-6
10. Cho, Kyunghyun; van Merrienboer, Bart; Bahdanau, DZmitry; Bougares, Fethi; Schwenk, Holger; Bengio, Yoshua (2014). "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". arXiv:1406.1078
11. Satorras, Vıctor Garcia, Emiel Hoogeboom, and Max Welling. "E (n) equivariant graph neural networks." *International conference on machine learning*. PMLR, 2021.