Journal of High School Science

Leveraging Deep Learning models to identify structurally adequate drug candidates for Parkinson's Disease

Garimella RV

Submitted: February 26, 2024, Revised: version 1, June 9, 2024
Accepted: July 4, 2024

Abstract
Parkinson's Disease (PD) is a neurodegenerative disease that causes the gradual impairment of movement primarily by mutations in the SNCA (synuclein alpha) gene, excessively producing the alpha-synuclein protein. This protein aggregates in formations known as Lewy Bodies which correlate with neurotransmitter dysfunction, hence leading to the common side-effects of PD, such as loss of balance and memory. Currently, there exist no widely available drugs for PD, partly due to the long process for drugs to be identified, FDA certified, and then marketed. This process, however, can be expedited through the use of drug repurposing by Machine Learning models. The model constructed in this paper utilized the Simplified Molecular Input Line Entry System (SMILES) encoding and vector representations of specific, FDA-approved drugs in order to extract important features of a drug's molecular structure. Then, disease representation vectors for PD were combined with a drug's vector representations and were used as inputs to both Deep Learning and Regression models to predict the efficacy of various drug-disease pairs. The use of these models was then applied specifically to PD, identifying the drugs that held promise for therapeutic efficacy in PD. Valproic Acid emerged as a potential drug for further research in its applications to PD, with the best predictive models utilizing Deep Learning. Drugs that are folate or pyridoxine modulators, carbonic anhydrase inhibitors, PPAR-γ modulators, anti-inflammatory agents (prostaglandin modulators, antioxidants, SIRT1 modulators), antifibrinolytics (plasminogen modulators), or those used to treat hyperthyroidism (thyroid hormone inhibitors), were scored high by the model with respect to their ability to modulate brain GABA levels, as were antimetabolites or alkylating agents that modulate nucleotides like adenosine and purine. Deep Learning models with low validation loss can predict drugs that are structurally adequate to provide therapeutic effects for PD and other medicinally under-served diseases.

Keywords
SMILES encoding, Machine Learning, Deep learning, Regression models, Parkinson's Disease, Lewy bodies, Drug repurposing, Valproic acid, Synuclein alpha, Drug Disease Association Matrix

Raghav Garimella, Evergreen Valley High School, 3300 Quimby Rd, San Jose, CA 95148, USA.
raghav.garimella@gmail.com

1. Introduction

Parkinson's Disease (PD) is a neurodegenerative disease that is known to interfere with motion assistive nerve cells. Therefore, many with PD experience shakiness and a loss of balance, in addition to memory loss. This disease is mainly caused by mutations in the SNCA gene, which aids in the creation of the protein called alpha-synuclein (a-syn). Mutations in this gene can lead to excessive production of a-syn, therefore leading to the formation of masses called Lewy Bodies (1). Lewy Bodies correlate with neurotransmitter dysfunction and decreased nerve-cell communication, hence leading to many of the debilitating symptoms of neurodegenerative diseases such as PD (2).

One possible method to search for therapies for neurodegenerative diseases such as PD is to use repurposed drugs. Drug repurposing is the process of finding if previously FDA-approved drugs (for any disease) are - or may be predicted to be - effective against diseases for which there yet exist no drugs or drugs that are marginally effective. It involves the identification and analysis of many drug-disease interactions in order to identify a new clinical use of these FDA-approved drugs (3). Drug repurposing may lessen the time needed for regulatory approval. FDA-approval for new chemical entities (NCEs) can take from ten to seventeen years (4).

Repurposed drugs can be found using Machine Learning models and need to be tested in clinical trials. The Drug Repurposing Knowledge Graph (DRKG) is a database which relates genes, compounds, and diseases. The Graph also contains pretrained models for drug repurposing and is an area of active research, especially for drug repurposing for COVID-19 (5, 6). However, the knowledge graph remains underdeveloped and does not contain adequate amounts of data to analyze multiple drug-disease associations for PD. Additionally, drug repurposing for neurodegenerative diseases must be nuanced, as different drugs have been found to be effective for different stages of these diseases. Moreover, it is important to have flexibility in the models used, as compared to pre-trained models, in order to repurpose drugs specifically for PD. Research by Stott *et.al*, into drug repurposing for PD utilized clinical trials to determine the efficacy of various agents such as Simvastin and Ambroxol as possible drugs for therapy for PD (7). These trials, however, take a long time to conduct, and the resultant drugs only aimed to limit the symptoms of PD as opposed to preventing its progression. Lasso logistic regression techniques were applied to repurpose drugs for neurodegenerative diseases (8). It is important, however, to test multiple multilayer Machine Learning models so that more effective drugs may be found. A stage-based analysis of PD, examining each stage of the disease, is also necessary to find more efficacious drugs.

Therefore, our research explored Machine Learning methods to identify candidate drugs for PD, with lower validation loss signifying stronger drug candidates. We hypothesized that a stage-based Deep Learning model trained on drug-disease associations would be able to identify drug candidates structurally similar to existing ones to provide therapeutic effects for PD.

To obtain drug-disease associations, the model was trained on a Drug-Disease Association matrix (DDA matrix) containing scored relationships between drugs and diseases based on molecular features. Drugs were evaluated structurally through the use of SMILES-encoded vector representations, used to extract features of the molecular composition of the drugs. Additionally, the model was trained to identify drugs for: PD (general), autosomal recessive early-onset PD, and late-onset PD. This approach eliminates the time needed for drug-identification in the drug creation process, and allows for key, targeted research into a specific drug for different stages of PD.

2. Materials and Methods

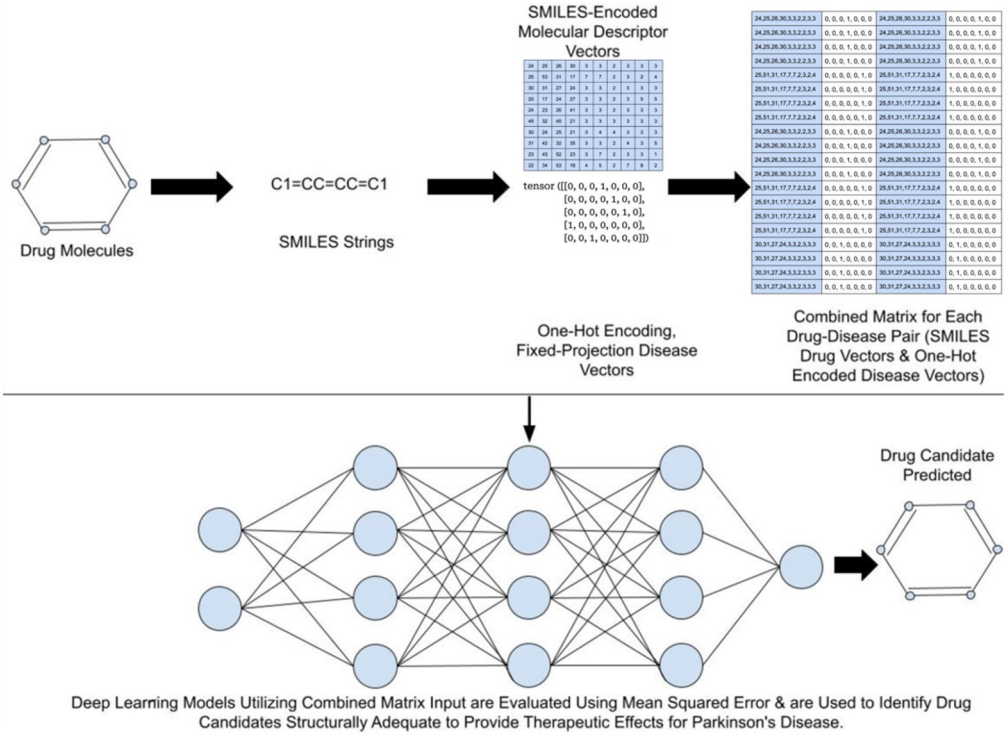

Figure 1. Overview of process and methodology utilized for SMILES encoding, construction of combined matrices, Deep Learning model architecture, and drug candidate prediction.

2.1 *Data preprocessing*

To obtain drug-disease associations, the model was trained on a Drug-Disease Association (DDA) matrix (9). Values in a DDA matrix are calculated using the phenotypes and genotypes of diseases, in accordance with known drug features as well as clinically proven drug-disease similarities. Relationships closer to 1 are indicative of a "positive" correlation between a drug and disease, while relationships closer to 0 are not. In order to remain consistent in feature extraction for the molecular composition and structures of various drugs, we utilized SMILES encoding as a standard form of molecular representation. SMILES encoding converts 3-D

representations of chemicals into strings that are readable by a computer. Data was initially imported from the DDA matrix. Then, the column denoting the 729 drug Compound IDs (CIDs) was converted into SMILES encoded strings, capturing the molecular composition and structures of each drug into computer readable format. This was performed through PubChem queries for each drug, utilizing the PubChemPy library, after which these SMILES strings were converted to decimal vector-representations of drugs based on 200 molecular descriptors (10). These vectors were formed utilizing Molecular Fingerprinting, a function from the rdkit Python library allowing for molecular feature extraction through unique identifiers on SMILES-encoded vectors. These vectors are stored corresponding to each FDA-approved drug. Before the models could be created, the DDA matrix underwent a "cleaning" process, where incomplete data was removed. From an original 5957 list of diseases, 5952 remained, after data cleaning and setting aside the three instances of PD in this dataset. The 5952 diseases were then vectorized using one-hot encoding, essentially assigning each disease a random value from 0 to 1. These vectors are then given a fixed-projection of size 100 after both the SMILES and disease vectors were normalized. These preprocessing steps ensured that the sizes of the vectors would be compatible for concatenation. Both the drug-related vectors and the disease-related vectors were then concatenated into a combined matrix utilizing PyTorch's built-in concatenation command (Figure 1). This resultant matrix contained every drug and disease pair, with 729 * 5952 rows of data, and was comprised of each vector concatenation, totaling to 300 columns (length 200 SMILES vectors and length 100 fixed-projected disease vectors).

2.2 *Deep Learning model architecture*

The Sequential model and MLP model were created after the combined matrix were split into training and testing data, utilizing the DDA matrix's values as labels for testing. Both models were created with six layers: linear, ReLU, linear, ReLU, linear, and then a final sigmoid layer**.** The loss in the MLP model decreased after including one batch normalization layer to the model's architecture, in order to stabilize the optimization process. The learning rate for both models was set to 0.0001 and 128 hidden layers were used for the MLP model. The data was then grouped into batch sizes of 27, in order to evenly split with the 729 FDA-approved drug instances. The function used to calculate the loss was the Mean Squared Error (MSE). Smaller values of MSE indicated lesser differences from the actual DDA matrix's values. Additionally, overfitting of the model to the training data was prevented by using a 75-25 split of data between training and testing. Since the disease vectors were given random values initially, these vectors needed to be updated as the models learnt more about each drug-disease relation, therefore an optimizer and gradient descent were used to minimize the loss, which was then utilized to update the one-hot encoded disease vectors with convergence tending values.

2.3 *Regression model architecture*

The Linear Regression model was constructed using a single linear layer, which produced optimal results. The Linear Regression model with L1 Regularization, or lasso regression,

which reduces the weights of the model every iteration as they approach or reach zero, was also created using a single linear layer. Finally, the Linear Regression model with L2 Regularization, or ridge regression, which reduced the weights of the model every iteration but never reached zero, was created using a single linear layer. This model architecture provided an optimal, low validation loss from each regression model. The Polynomial Regression Model was created differently. In order to reach a low validation loss, this model was a quadratic function (degree 2) and was constructed utilizing both a linear and sigmoid layer. To produce optimal results, this model also was programmed with a slower learning rate of 0.001, as opposed to the 0.0001 used for all previous models.

2.4 *Predictions for PD*

In order to use this model to predict which drugs were structurally similar to provide therapeutic efficacy for PD, data preprocessing was applied to a combined matrix specific to PD. The combined matrix would now concatenate the same SMILES encoded vectors but would now include one-hot encoded vectors including only the three forms of PD found in the dataset. This would mean the new combined matrix would have 729 * 3 rows of data with columns of length 300. The model was trained on the original combined matrix, with all of the diseases, excluding the cases of PD, but would now also be additionally tested using the data in the combined matrix for PD. The model was run for ten epochs, or iterations, through the data. On the final run through the data, when the model had learnt the maximum about the drug-disease interactions, the predicted (output) values of the model were stored in a list. By iterating through this list, where each element represented an interaction between one form of PD and a specific FDA-approved drug, the largest positive interactions (scaled from 0 to 1) were stored in three variables. Each variable corresponded to the largest positive interaction between a specific form of PD and a specific drug. Finally, by storing the index at which these maxima occurred, we were able to cross-reference these indices with the original indices in the list of FDA-approved drugs, therefore allowing us to find the drugs predicted as being the most efficacious for each form of PD by our model in the drug list.

3. Results

To investigate which drugs would be the best candidates to provide therapeutic efficacy for PD, we created six different Machine Learning models: a Deep Learning-based Feed Forward Neural Network and MLP Classifier, Regression-based Linear Regression, Lasso Linear Regression, Ridge Linear Regression, and Polynomial Regression. The models were trained over ten epochs of the data, utilizing an accuracy metric of Mean Squared Error (MSE). MSE was used because it penalizes smaller errors from the actual labels of the DDA less than it does larger errors. This increases the accuracy of the model as the model attempts to predict precise decimal values between 0 and 1, therefore making smaller errors acceptable, and larger discrepancies indicative of a poorly constructed model.

Table 1. Results of 12 trials with seeded (random) values for a MLP Model. PD, early onset PD, and late onset PD columns contain drug CIDs (Compound Identifiers) predicted drugs From 12 trials with seeded values for MLP model

Seed Value	PD	Earlv Onset PD	Late Onset PD
5	3121	3121	3767
7	3121	3121	3121
22	3121	3121	3121
30	3121	4075	3121
42	3121	3121	3121
43	1678	1678	1678
44	3121	3121	3121
50	3767	3121	4075
60	3121	3121	3121
70	3121	3121	3121
80	3121	3767	3121
100	1678	1678	3121
	Mode: 3121	Mode: 3121	Mode: 3121

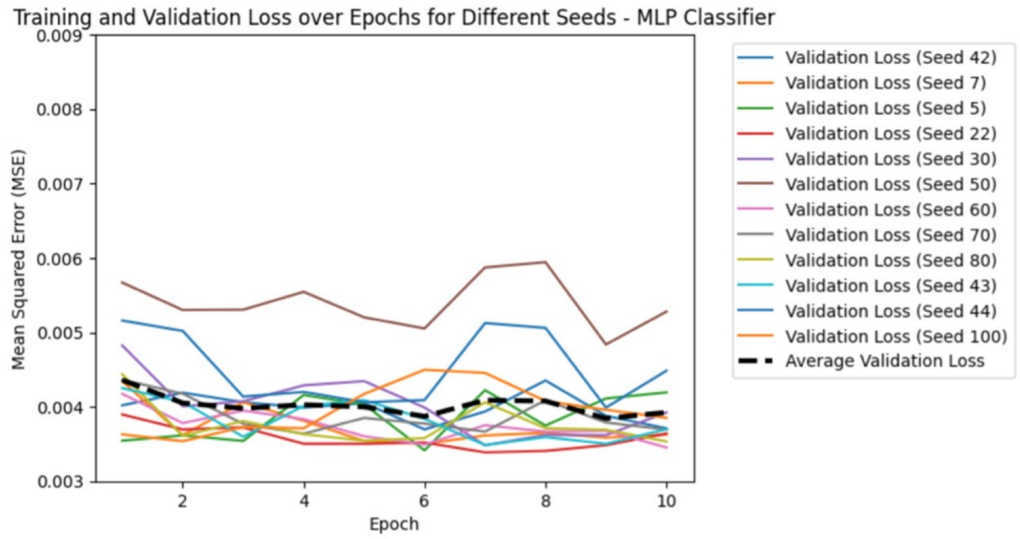

Figure 2. Results of MSE validation loss over 12 trials with seeded (random) values for a MLP model.

3.1 *Multi-Layer Perceptron classifier (MLP)*

The MLP model utilized the inputs of a combined matrix containing one-hot encoded vectors representing data of the three forms of PD analyzed and the SMILES-encoded vectors representing the molecular features of the FDA-approved drugs. This model was run twelve times, each with a different random

seed value representing the generation of random values for reproducible results in a Machine Learning model. The validation loss, representative of the difference between the values predicted by the model and actual labels of the DDA (calculated by MSE), reached a low of 0.0034 in Epoch 8 of the fourth seed value (22). The average validation loss at the tenth (final) epoch for this model was 0.0039, with the best-fit line representing the average validation loss per epoch plotted as a dashed line (Figure 2). Overall, the MLP Classifier predicted Valproic Acid (Drug CID: 3121) as the best candidate drug for PD, early-onset PD, and late-onset PD (Table 1). At the fourth seed value, presenting the lowest validation loss, Valproic Acid was predicted with a score of 0.908, 0.885, and 0.900 for PD, early-onset PD, and late-onset PD respectively.

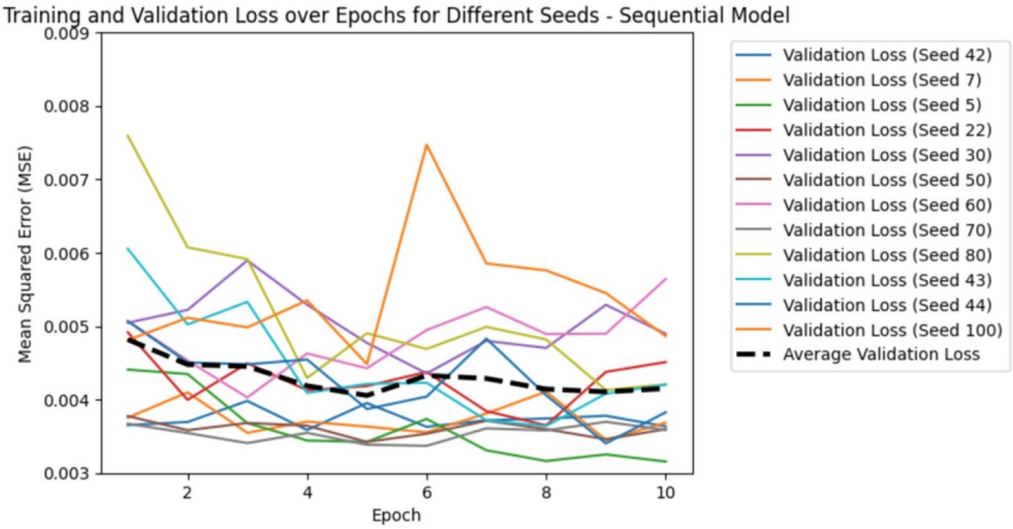

Figure 3. Results of MSE validation loss over 12 trials with seeded (random) values for a Sequential model.

3.2 *Sequential (feed-forward) neural network*

The Sequential model utilized the same inputs and accuracy metric as the MLP Classifier, using one-hot encoded vectors representing data of the three forms of PD, SMILES-encoded vectors representing the molecular features of the FDA-approved drugs, and MSE as the accuracy metric. The model was run twelve times using seed values to produce reproducible results (Table 2). The average validation loss at the tenth (final) epoch for this model was 0.0043, with the average validation loss per epoch graphed as a dashed line (Figure 3). This model reached its lowest validation loss of 0.0032 in Epoch 8 of its 3rd seeded value (5). This run predicted Valproic Acid as the best drug for PD with a score of 0.940. Contrary to the mode of the dataset, however, this run predicted Streptozocin as the best candidate drug for early-onset PD with a relatively lower score of 0.856, and Isoniazid as the best candidate drug for late-onset PD with a score of 0.892. The table depicts the results for the twelve runs and shows that the mode candidate drug of each category of PD was Valproic Acid.

Table 2. Results of 12 trials with seeded (random) values for a Sequential model. PD, early onset PD, and late onset PD columns contain drug CIDs (Compound Identifiers), predicted drugs From 12 trials with seeded values for Sequential model

Seed Value	PD	Early Onset PD	Late Onset PD
5	3121	29327	3767
7	3121	2708	3767
22	3121	38853	2708
30	3121	2554	3121
42	2708	2708	2708
43	3121	3609	5526
44	3950	2244	3121
50	2244	3121	3121
60	3121	3121	2244
70	3121	4075	3121
80	3121	445154	3121
100	3121	3121	3121
	Mode: 3121	Mode: 3121	Mode: 3121

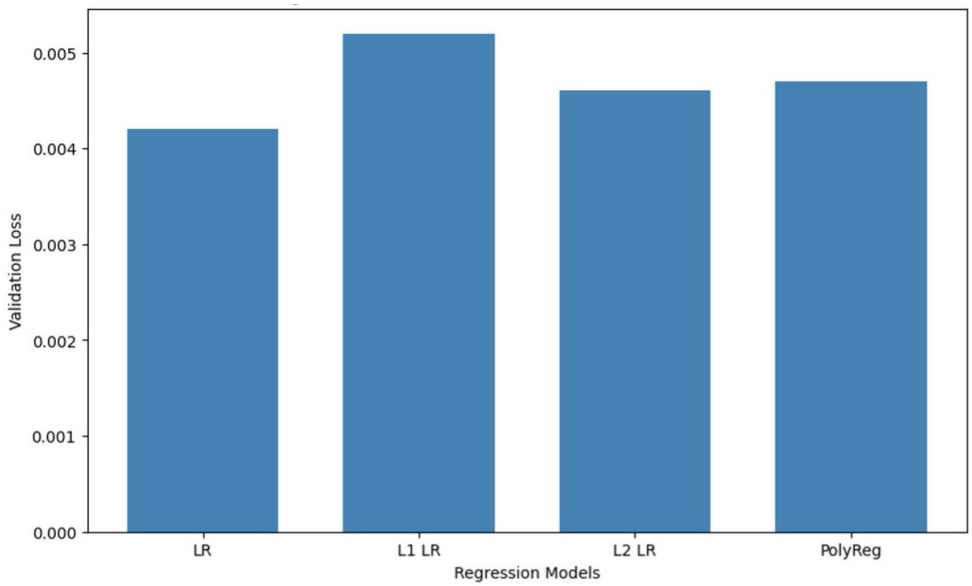

Figure 4. Results of MSE validation loss for all Regression models.

3.3 *Regression models*

All regression models utilized the same inputs and accuracy metrics as the Sequential and MLP models, utilizing one-hot encoded vectors for disease representation and SMILES-encoded vectors for drug representation. Regression models aimed to fit single graphical models to a relationship between independent

and dependent variables. Therefore, as the method of fitting to data did not change from run to run, these models only needed to be run once unlike the MLP and Sequential models with Neural Networks. Four regression models were compared: Linear Regression, Linear Regression with L1 Regularization, Linear Regression with L2 Regularization, and Polynomial Regression. The Linear Regression model resulted in the lowest validation loss of 0.0042, whereas the L1, L2, and Polynomial Regression models output higher validation losses of 0.0052, 0.0046, and 0.0047 respectively (Figure 4). These higher validation losses were likely due to the Regression models' sensitivity to outliers in the data, therefore showcasing the benefits of a more nuanced Deep Learning-based approach in finding potential drug candidates for PD. This model, still, corroborated the outputs of the Sequential and MLP models, showing Valproic Acid (Drug CID: 3121) as a key drug candidate for PD. The Linear Regression model predicted Valproic Acid as the best drug candidate for PD, with a score of 0.890 for PD (general), a score of 0.881 for early-onset PD, and a score of 0.883 for late-onset PD.

4. Discussion

4.1 *Biological analysis*

The MLP model resulted in the lowest average validation loss, or lowest difference between the model's predicted values and the actual DDA labels (Figure 2). This model consistently predicted Valproic Acid as the best candidate drug for all three instances of PD used to test the model (Table 2). Valproic Acid is often used for many neurological and psychiatric disorders, commonly to prevent seizures (11). This is because Valproic Acid inhibits histone deacetylases (HDAC), enzymes which prevent DNA transcription (12). HDAC6 is concentrated in Lewy Bodies, formations that correlate with neurotransmitter dysfunction, leading to many of the memory and motor skills lost in patients with neurodegenerative diseases like PD (13). These Lewy Bodies are formed due to the excessive production of a-syn, a cardinal impact of mutations in the SNCA gene which causes PD (14, 15). By limiting HDAC6 production by the use of Valproic Acid, many of the effects and harmful progressions of PD can potentially be prevented (16). Additionally, Valproic Acid has been shown to counteract the death of nigral neurons in the brain, and as the substantia nigra, the part of the brain controlling movement, is deteriorated in PD and by Lewy Body formation, Valproic Acid may have neuroprotective properties necessary to mitigate the effects of PD (17-20). Therefore, it is not unreasonable to assume that Valproic Acid should be further investigated as a drug that can provide therapeutic effects for PD. Since Lewy Body formations effect the progression of PD, particularly in late-onset PD, as Lewy Bodies spur dementia in later-stages of PD in patients, Valproic Acid may mitigate the negative effects of PD (21). Studies show that the inhibition of HDAC may reverse PD's effects on both motor and non-motor skills. Mice carrying the LRRK2 R1441G mutation saw an improvement in both motor and non-motor behavior after Valproic Acid was administered, as seen by increased cell survival, decreased inflammatory responses, and reduced expression of HDAC (22).

Furthermore, another drug candidate predicted, specifically for late-onset PD, was Isoniazid. This drug has been shown to aid in the treatment of TB, and individuals with TB have a 1.38 fold risk of PD. This drug candidate can aid in the prevention of PD as well (23). Isoniazid is known to act as an inhibitor of Gamma-aminobutyric acid (GABA) aminotransferase, and since GABA inhibition is necessary to prevent inflammation of blood vessels, a common side-effect of PD, Isoniazid can potentially aid in the prevention of the neurodegenerative process (24). Isoniazid can inhibit GABA production in many ways including by inhibiting the production of glutamic acid decarboxylase which converts glutamate into GABA (25). Additionally, in late-onset PD, individuals typically see increased problems with gait impairment. This is often due to a condition called chorea, where gait impairment is induced by unpredictable movements. Isoniazid has been shown to mitigate some of the effects of chorea through its regulation of GABA levels, and since mutations in the PRKN gene (gene associated with late-onset PD) often cause chorea, Isoniazid has the capability to lessen some of the negative impacts of this neurodegenerative disease (26).

4.2 *Deep Learning model analysis*

Both the MLP model and Sequential model predicted Valproic Acid as the candidate drug best suited to provide therapeutic efficacy for PD. These Deep Learning models achieved a low validation loss (0.0034 for MLP and 0.0032 for Sequential, compared to 0.0042 for Linear Regression) and could predict candidate drugs structurally similar to those in the training set to provide therapeutic efficacy for PD (Figures 2-4). Drug repurposing has the potential to reduce the time needed for regulatory approval for the development of new drugs for diseases such as PD. The use of Machine Learning expedites the search for structurally similar drugs. Additionally, drug repurposing may decrease the cost needed to bring new drugs to the market, an expenditure of over $300 million (27). Specific to PD, in America alone, approximately 500,000 individuals suffer from this disease (28). More research needs to be performed to definitively validate Valproic Acid as being suitable to progress into clinical trials for PD. Overall, Deep Learning models, such as the MLP model, may be able to utilize structural similarities, molecular properties, and disease genotypes and phenotypes to predict the drug likely to provide therapeutic effects for any given disease for which minimally effective drugs exist. These predictions may serve as a foundation for future, targeted research into Valproic Acid's applications to the mitigation of PD progression or causation.

4.3 *Relevance of the model in finding repurposed drugs*

Out of the 38 drugs listed in the appendix with a score > 0.852, Ten (colored red) (26%) act 'directly' on the CNS; primarily on GABA levels, being either neurotransmitter and/or receptor agonists / antagonists / mimetics, replenishers, modulators (reuptake, synthesis or degradation) or analogs; one of them being Valproic acid. Sixteen (colored green) (42%) drugs that are folate or pyridoxine modulators, carbonic anhydrase inhibitors, PPAR-γ modulators, anti-inflammatory agents (prostaglandin modulators, antioxidants, SIRT1 modulators), antifibrinolytics (plasminogen

modulators), or treat hyperthyroidism (thyroid hormone inhibitors), act indirectly to modulate brain GABA levels. Four (colored yellow) (11%) are antimetabolites or alkylating agents that can modulate nucleotides like adenosine and purine. Extra/Intracellular imbalances, or dysfunction of the latter are known correlates of PD. Three (blue) (8%) are voltage-gated ion-channel modulators. The remaining drugs modulate lipids, chelate metal ions or are toxins. It is evident that the model does return drugs that have established mechanisms of action with regard to interference with GABA or dopamine receptors or levels in the brain, thereby validating the core competency of the model. It is also of interest that the model does not seem to distinguish between positive or negative interference regarding GABA or dopamine levels. This may be construed as an advantage because the neurological effects of either increased or decreased expression of neurotransmitters are often neuron/ pathway/ brain region specific. The classification of drugs listed in the appendix according to the mechanism of action is heuristic, because of the varied degrees of overlap between mechanisms.

Valproic acid is attributed with indirectly influencing GABA or dopamine levels by modulating activities for folate, PPAR-γ, prostaglandins, TSH, plasminogen, and voltage gated ion channels, in addition to its direct effects of increasing the activity of the GABA-synthetic enzyme glutamic acid decarboxylase and inhibiting GABA transaminase. Such as wide spectrum of action may – in part – be responsible for the model assigning it a high score.

The fact that there were no anticholinergics, MAO or COMT inhibitors and NMDA antagonists among the found drugs may be due to the fact that the DDA matrix does not specifically describe mechanisms of action; even so, these are widely used drugs to treat a variety of neurological ailments, and the non-assignment of a relatively high score for these categories of drugs is perplexing. We do not have an explanation for this artifact at the present time. It is also instructive that the ubiquitin, mitochondrial and axonal transport pathways that comprise a large portion of the Parkinson's disease KEGG pathway database (https://www.genome.jp/kegg/pathway.html#drug) are under-represented among the drugs found using the model.

4.4 *Limitations & future research*

One of the limitations of this research originate in the dataset and model specificity. The DDA matrix includes associations between only 729 drugs and 5955 diseases. This means that some FDA-approved drugs remain unexplored as possible drug candidates for PD. Additionally, some diseases also remain unexplored, therefore implying that the model can be improved through training on multiple more drug-disease pairs. The model constructed in this study was itself based on the molBERT Machine Language transformer architecture wherein SMILES strings were converted to decimal vector-representations of drugs based on 200 molecular descriptors. The molBERT model obtained a maximum AUCROC of 0.743 (10). Therefore, it is important to understand that the ML model used in this study was itself based on a less-than-accurate ML model, hence errors could have been compounded in the process. Additionally,

although multiple hyperparameter changes were evaluated, model specifications such as learning rates and numbers of layers were primarily optimized through trial and error, meaning that the model specifications may not have been indicative of a fully optimized Deep Learning model for PD. Finally, it must also be noted that in long-term, high-dose use of antiepileptic drugs (AEDs), such as Valproic Acid have been shown to cause the formation of a reversible form of PD. Additionally, Valproic Acid, specifically, can cause the exacerbation of Parkinson's symptoms due to the promotion of a-syn (29). This however, must be further investigated, as recent research indicates that endogenous a-syn may possess neuroprotective qualities due to improving the function of dopaminergic neurons in the substantia nigra (30). Therefore, Valproic Acid's effect to cause PD requires further investigation (the dose defines the poison), as do the potential benefits of utilizing Valproic Acid to treat PD.

Multiple pathways including different proteins, as well as upstream and downstream processing, cellular components, and biological processes may affect the cause and progression of PD. This implies that structurally dissimilar molecules can mitigate PD. A-syn inhibitors, such as curcumin, polyphenols, nicotine, EGCG, and stilbene, all inhibit multiple different proteins, therefore leading them to have structural differences. Some of the drug candidates in this dataset will contain some of the same inhibition qualities, meaning that there is a great amount of variance in the structures considered to be "adequate" to break apart a-syn aggregates (31-33) or act *via* other mechanisms. Additionally, other pathways of disease prevention such as NLRP3 inhibitors, NURR1 activators, and RXR agonists similarly all target different proteins and have different structures (34). Future work could assign structures with similarities to certain compounds as relevant data points before utilizing the Deep Learning model. These compounds can include NPT200-11 (minzasolmin), MCC950, IZD174 (emlenoflast), BRF110, IRX4204, IRL790 (mesdopetam), and SCFA (short chain fatty acids). These compounds were found, de-novo, to suppress symptoms or progression of PD. Valproic Acid is structurally similar to SCFAs as it is a branched short-chain fatty acid (SCFA). Valproic Acid has two propyl groups which are linked to an acetic acid moiety (35).

More research needs to be performed by the inclusion of more diseases into the DDA matrix, as well as more FDA-approved drugs. This will increase the number of drug-disease pairs that the model is trained on, resulting in a wider range of drugs that may be therapeutic efficacious in PD. Additionally, more targeted research into the molecular structure and properties of Valproic Acid must be performed, including its dosage necessary for PD patients in different stages of the disease, as well as its potential long-term benefits or adverse effects for PD patients.

Finally, a more rational approach into optimizing model parameters such as learning rates and layers can help increase the accuracy of this model's predictions. The hyperparameters currently utilized were chosen through repeated testing, but could still be further improved. Lowering the learning rate can cause convergence but also unstable

results, while increasing the learning rate can create more stabilized results over a larger number of epochs. Increasing hidden layer sizes and number can cause increased complexity in the associations made by the Deep Learning models, but also pose the risk of overfitting, while decreasing hidden layer sizes and number can reduce overfitting but will simplify the model. Increasing the batch sizes requires more memory and runtime, but can help cause convergence of values, while decreasing batch sizes may cause illogical generalizations by the model, but can also help prevent it from predicting on local minima. Activation functions, such as RelU, and weight initiation can be tweaked to aid in model specificity and stability when making predictions. Regularization, seen in the Regression Models as well with L1 and L2 (lasso and ridge) aids in reducing noise during training and the construction of a more robust model. Utilizing these different hyperparameters and tweaking them can cause differences in the drugs output. Regularization is used to construct a robust model, but is not utilized excessively in order to prevent decreased performance on the training set. Finally, after testing through multiple epochs, the learning rate was increased to cause more stable results over the 120 total epochs tested on for the Deep Learning models (12 seeds each with 10 epochs).

5. Conclusion

The MLP model demonstrated that a multi-layered Deep Learning model could predict structurally adequate drug candidates for PD. The utilization of SMILES and One-Hot Encoding extracted molecular information and composition, while reformatting the data into a vector-representation. Subsequently, the combined matrix provided a compacted input to the models. Valproic Acid emerged as the drug candidate with the greatest drug-PD score. Drugs that are folate or pyridoxine modulators, carbonic anhydrase inhibitors, PPAR-γ modulators, anti-inflammatory agents (prostaglandin modulators, antioxidants, SIRT1 modulators), antifibrinolytics (plasminogen modulators), or those used to treat hyperthyroidism (thyroid hormone inhibitors), were scored high by the model with respect to their ability to modulate brain GABA levels, as were antimetabolites or alkylating agents that modulate nucleotides like adenosine and purine. Drugs such as Valproic Acid may hence be included into, and form a part of, key, targeted research and eventual clinical trials into drug repurposing for PD. Overall, utilizing important molecular features, drug candidates with particularly high drug-PD scores can be researched further, expediting the drug-identification process for PD. Machine Learning models such as Deep Learning may be effective in the field of drug repurposing, identifying potential drug candidates that can provide therapeutic effects for PD and other medicinally under-served diseases.

Acknowledgement

I greatly appreciate Annamarie Bair for her gracious tutelage and assistance throughout the drafting and programming process of this research project.

Abbreviations

PD: Parkinson's Disease, SMILES: Simplified Molecular Input Line Entry System, FDA: Food & Drug Administration, SNCA: Synuclein Alpha, a-syn: alpha-synuclein, MLP: Multi-Layer Perceptron, DDA: Drug-Disease Association, HDAC: Histone-Deacetylases, AED: Antiepileptic Drug, MSE: Mean Squared Error, ReLU: Rectified Linear Unit, GABA: Gamma-aminobutyric acid, MAO: Monoamine oxidase, COMT: Catechol-O-methyl-transferase, KEGG: Kyoto Encyclopedia of Genes and Genomes, NLRP3: NOD-, LRR- and pyrin domain-containing protein 3, NURR1: Nuclear receptor related 1 protein, RXR: Retenoid X receptor, EGCG: Epigallocathechin gallate, PPAR-γ: Peroxisome proliferator activated receptor gamma, TSH: Thyroid stimulating hormone, CNS: Central nervous system, LRRK2 R1441G: leucine-rich repeat kinase 2, PRKN: Parkin RBR E3 Ubiquitin Protein Ligase, OXPHOS: Oxidative phosphorylation, SIRT1: Sirtuin (silent mating type information regulation 2 homolog) 1, AUCROC: Area Under the Curve of the Receptor Operator Curve.

6. References

1. Srinivasan, E.; Chandrasekhar, G.; Chandrasekar, P.; Kannan, A.; Vickram, A. S.; Karunakaran, R.; et al. Alpha-Synuclein aggregation in Parkinson's disease. *Frontiers in Medicine* (2021), 8. https://doi.org/10.3389/fmed.2021.736978.

2. *Causes*. Stanford Health Care. https://stanfordhealthcare.org/medical-conditions/brain-and-nerves/dementia-with-lewy-bodies/causes.html#:~:text=Lewy%20bodies%20in%20the%20brain,with%20the%20production%20of%20dopamine.

3. Pareek, S.; Huang, Y.; Nath, A.; Huang, R. S. The success story of drug repurposing in breast cancer. In *Elsevier eBooks*; (2020); pp 173–190. https://doi.org/10.1016/b978-0-12-819668-7.00006-3.

4. Krishnamurthy, N.; Grimshaw, A.; Axson, S. A.; Choe, S. H.; Miller, J. E. Drug repurposing: a systematic review on root causes, barriers and facilitators. *BMC Health Services Research* (2022), 22 (1). https://doi.org/10.1186/s12913-022-08272-z.

5. Gogineni, A. K. *Analysis of Drug repurposing Knowledge graphs for Covid-19*. arXiv.org. https://arxiv.org/abs/2212.03911.

6. Negro, A.; Montagna, F.; Teng, M. N.; Neal, T.; Thomas, S.; King, S.; Khan, R. R. Analysis of the evolution of COVID-19 disease understanding through temporal knowledge graphs. *Frontiers in Research Metrics and Analytics* (2023), 8. https://doi.org/10.3389/frma.2023.1204801.

7. Stott, S.; Wyse, R.; Brundin, P. Drug Repurposing for Parkinson's Disease: The International Linked Clinical Trials experience. *Frontiers in Neuroscience* (2021), 15. https://doi.org/10.3389/fnins.2021.653377.

8. Courtois, É.; Nguyen, T.; Fournier, A.; Carcaillon-Bentata, L.; Moutengou, E.; Escolano, S.; et al. Identifying protective drugs for Parkinson's disease in Health-Care databases using machine learning. *Movement Disorders* (2022), *37* (12), 2376–2385. https://doi.org/10.1002/mds.29205.

9. Moridi, M.; Ghadirinia, M.; Sharifi-Zarchi, A.; Zare-Mirakabad, F. The assessment of efficient representation of drug features using deep learning for drug repositioning. *BMC Bioinformatics* (2019), 20 (1). https://doi.org/10.1186/s12859-019-3165-y.

10. Fabian, B. *Molecular representation learning with language models and domain-relevant auxiliary tasks*. arXiv.org. https://arxiv.org/abs/2011.13230.

11. Rahman, M. *Valproic acid*. StatPearls - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK559112/#:~:text=Valproic%20acid%20(VPA)%20ia%20as,disorder%20management%2C%20and%20migraine%20prophylaxis.

12. Göttlicher, M.; Minucci, S.; Zhu, P.; Krämer, O. H.; Schimpf, A.; Giavara, S.; et al. Valproic acid defines a novel class of HDAC inhibitors inducing differentiation of transformed cells. *The EMBO Journal* (2001), 20 (24), 6969–6978. https://doi.org/10.1093/emboj/20.24.6969.

13. Miki, Y.; Mori, F.; Tanji, K.; Kakita, A.; Takahashi, H.; Wakabayashi, K. Accumulation of histone deacetylase 6, an aggresome-related protein, is specific to Lewy bodies and glial cytoplasmic inclusions. *Neuropathology* (2011), 31 (6), 561–568. https://doi.org/10.1111/j.1440-1789.2011.01200.x.

14. Kim, W. S.; Kågedal, K.; Halliday, G. M. Alpha-synuclein biology in Lewy body diseases. *Alzheimer's Research & Therapy* (2014), 6 (5–8). https://doi.org/10.1186/s13195-014-0073-2.

15. Hershey, L. A.; Coleman-Jackson, R. Pharmacological Management of Dementia with Lewy Bodies. *Drugs & Aging* (2019), 36 (4), 309–319. https://doi.org/10.1007/s40266-018-00636-7.

16. Shults, C. W. Lewy bodies. *Proceedings of the National Academy of Sciences of the United States of America* (2006), 103 (6), 1661–1668. https://doi.org/10.1073/pnas.0509567103.

17. Ximenes, J. C. M.; Neves, K. R. T.; Leal, L. K. A. M.; Carmo, M. R. S. D.; De Castro Brito, G. A.; Da Graça Naffah-Mazzacoratti, M.; et al. Valproic acid neuroprotection in the 6-OHDA model of Parkinson's disease is possibly related to its Anti-Inflammatory and HDAC inhibitory

properties. *Journal of Neurodegenerative Diseases (Online)* (2015), 2015, 1–14. https://doi.org/10.1155/2015/313702.

18. Zhang, C.; Yuan, X.; Hu, Z.; Liu, S.; Li, H.; Wu, M.; et al. Valproic Acid Protects Primary Dopamine Neurons from MPP+-Induced Neurotoxicity: Involvement of GSK3βPhosphorylation by Akt and ERK through the Mitochondrial Intrinsic Apoptotic Pathway. *BioMed Research International* (2017), 2017, 1–12. https://doi.org/10.1155/2017/8124501.

19. *Parkinson's Disease – symptoms, diagnosis and treatment*. https://www.aans.org/en/Patients/Neurosurgical-Conditions-and-Treatments/Parkinsons-Disease.

20. Goldman, J. G.; Holden, S. Parkinson's Disease. In *Elsevier eBooks*; (2016); pp 242–248. https://doi.org/10.1016/b978-0-12-397045-9.00271-8.

21. Hindle, J. V. Ageing, neurodegeneration and Parkinson's disease. *Age And Ageing* (2010), 39 (2), 156–161. https://doi.org/10.1093/ageing/afp223.

22. Kim, T.; Song, S.; Park, Y.; Kang, S.; Seo, H. HDAC inhibition by valproic acid induces neuroprotection and improvement of PD-like behaviors in LRRK2 R1441G transgenic mice. *Experimental Neurobiology* (2019), 28 (4), 504–515. https://doi.org/10.5607/en.2019.28.4.504.

23. Shen, C.-H.; Chou, C.-H.; Liu, F.-C.; Lin, T.-Y.; Huang, W.-Y.; Wang, Y.-C.; Kao, C.-H. Association between Tuberculosis and Parkinson Disease. Medicine 2016, 95 (8), e2883. https://doi.org/10.1097/md.0000000000002883.

24. Treatment of Huntington's Chorea with Isoniazid. New England journal of medicine. 1977, 297 (15), 840–840. https://doi.org/10.1056/nejm197710132971517.

25. Badrinath, M.; John, S. Isoniazid Toxicity. PubMed. https://www.ncbi.nlm.nih.gov/books/NBK531488/.

26. Manyam, B. V.; Katz, L.; Hare, T. A.; Kaniefski, K.; Tremblay, R. D. Isoniazid-Induced Elevation of CSF GABA Levels and Effects on Chorea in Huntington's Disease. Annals of Neurology 1981, 10 (1), 35–37. https://doi.org/10.1002/ana.410100107.

27. Wouters, O. J.; McKee, M.; Luyten, J. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *JAMA* (2020), 323 (9), 844. https://doi.org/10.1001/jama.2020.1166.

28. *Parkinson's Disease: challenges, progress, and promise*. National Institute of Neurological Disorders and Stroke. https://www.ninds.nih.gov/current-research/focus-disorders/parkinsons-disease-research/parkinsons-disease-challenges-progress-and-promise#:~:text=Approximately%20500%2C000%20Americans%20are%20diagnosed,1%20million%20Americans%20have%20PD.

29. Silver, M.; Factor, S. A. Valproic Acid-Induced Parkinsonism: Levodopa Responsiveness with Dyskinesia. Parkinsonism & Related Disorders 2013, 19 (8), 758–760. https://doi.org/10.1016/j.parkreldis.2013.03.016.

30. Leng, Y.; Chuang, D.-M. . Endogenous -Synuclein Is Induced by Valproic Acid through Histone Deacetylase Inhibition and Participates in Neuroprotection against Glutamate-Induced Excitotoxicity. Journal of Neuroscience 2006, 26 (28), 7502–7512. https://doi.org/10.1523/jneurosci.0096-06.2006.

31. Rebas, E.; Rzajew, J.; Radzik, T.; Zylinska, L. Neuroprotective Polyphenols: A Modulatory Action on Neurotransmitter Pathways. Current Neuropharmacology 2020, 18. https://doi.org/10.2174/1570159x18666200106155127.

32. Xu, B.; Chen, J.; Liu, Y. Curcumin Interacts with α-Synuclein Condensates to Inhibit Amyloid Aggregation under Phase Separation. ACS omega 2022, 7 (34), 30281–30290. https://doi.org/10.1021/acsomega.2c03534.

33. Ruan, S.; Xie, J.; Wang, L.; Guo, L.; Li, Y.; Fan, W.; et al. Nicotine Alleviates MPTP-Induced Nigrostriatal Damage through Modulation of JNK and ERK Signaling Pathways in the Mice Model of Parkinson's Disease. Frontiers in pharmacology 2023, 14. https://doi.org/10.3389/fphar.2023.1088957.

34. Lipska, K.; Gumieniczek, A.; Filip, A. A. Anticonvulsant Valproic Acid and Other Short-Chain Fatty Acids as Novel Anticancer Therapeutics: Possibilities and Challenges. Acta Pharmaceutica 2020, 70 (3), 291–301. https://doi.org/10.2478/acph-2020-0021.

35. Al-Nusaif, M.; Lin, Y.; Li, T.; Cheng, C.; Le, W. Advances in NURR1-Regulated Neuroinflammation Associated with Parkinson's Disease. International Journal of Molecular Sciences 2022, 23 (24), 16184. https://doi.org/10.3390/ijms232416184.

Appendix

Drug Candidates Predicted with a Score > 0.852 by Best-Performing Deep Learning Model: MLP Model. Their use and mechanisms of action are also listed. See section 4.3 for an explanation of the color coding scheme.

1. CID: 4993, Drug Name: Pyrimethamine, antiparasitic, antimalarial, Folic acid antagonist, folic acid is needed for GABA synthesis
2. CID: 4994, Drug Name: Pyrithyldione, psychoactive, hypnotic, sedative, positive modulator of the $GABA_A$ ion channel.
3. CID: 1678, Drug Name: 3-nitropropionic acid, mitochondrial toxin, OXPHOS inhibitor, increased dopamine driven inhibition of GABA.
4. CID: 2723601, Drug Name: 6-Thioguanine, antineoplastic, antimetabolite
5. CID: 2578, Drug Name: Carmustine, DNA alkylating agent, antineoplastic
6. CID: 657298, Drug Name: Propylthiouracil, hyperthyroidism, thyroid hormone inhibitor, decreases extracellular GABA levels via kynurenic acid modulation
7. CID: 1349907, Drug Name: Methimazole, hyperthyroidism, thyroid hormone inhibitor, decreases extracellular GABA levels via kynurenic acid modulation
8. CID: 4498, Drug Name: Nipecotic acid, GABA reuptake inhibitor, antiseizure
9. CID: 1046, Drug Name: Pyrazinamide, anti TB, Reduction in pyridoxine metabolism decreases GABA synthesis.
10. CID: 5526, Drug Name: Tranexamic acid, antifibrinolytic, plasminogen inhibitor, GABA inhibition in cortical neurons
11. CID: 6047, Drug Name: Levodopa, dopamine replenisher, for PD
12. CID: 5665, Drug Name: Vigabatrin, GABA analog, GABA transaminase inhibitor
13. CID: 3366, Drug Name: 5-Fluorocytosine, antifungal, antimetabolite, DNA synthesis inhibitor
14. CID: 938, Drug Name: Nicotinic acid, lipid and lipoprotein modulator
15. CID: 3117, Drug Name: Disulfiram, dopamine beta hydroxylase inhibitor
16. CID: 3121, Drug Name: **Valproic acid**, antiseizure, voltage gated Na channel inhibitor, increases GABA levels by inhibiting GABA transaminase and succinate semialdehyde dehydrogenase
17. CID: 564, Drug Name: 6-aminohexanoic acid, antifibrinolytic, plasminogen activator inhibitor, Recombinant tissue plasminogen activator decreases GABA levels in rat hippocampus.
18. CID: 3767, Drug Name: Isoniazid, antimicrobial, mycolic acid inhibitor, Reduction in pyridoxine metabolism decreases GABA synthesis.
19. CID: 14520, Drug Name: Citioline, neuroprotective, increases synthesis of phosphatidylcholine
20. CID: 1983, Drug Name: Acetaminophen, antipyretic, prostaglandin inhibitor, prostaglandins depress GABA release

21. CID: 1986, Drug Name: Acetazolamide, CNS carbonic anhydrase inhibitor, anticonvulsant, GABAergic excitation is related to increased carbonic anhydrase activity.
22. CID: 2244, Drug Name: Aspirin, cyclo-oxygenase inhibitor, prostaglandin inhibitor, prostaglandins depress GABA release
23. CID: 38853, Drug Name: Methyldopa, central alpha-2 adrenergic agonist, decreases blood pressure, GABA release stimulant.
24. CID: 3019, Drug Name: Diazoxide, vasodialator, activates K channels, inhibits insulin release
25. CID: 5324, Drug Name: Sulfaguanidine, antiparasitic, dihydrofolic acid blocker, folic acid is needed to synthesize GABA
26. CID: 4173, Drug Name: Metronidazole, antibacterial, antiprotozoal, protein synthesis inhibitor, inhibits $GABA_A$ activation
27. CID: 4943, Drug Name: Propofol, antiemetic, anesthetic, GABA stimulant
28. CID: 31072, Drug Name: Carbimazole, hyperthyroidism, thyroid hormone synthesis inhibitor, decreases extracellular GABA levels via kynurenic acid modulation
29. CID: 445154, Drug Name: Resveratrol, anti-inflammatory, various MOA
30. CID: 2788, Drug Name: Clioquinol, antifungal, antibacterial, DNA replication enzyme inhibitor
31. CID: 39912, Drug Name: (S)-(+)-Ibuprofen, cyclo-oxygenase inhibitor, prostaglandin synthesis inhibitor, prostaglandins depress GABA release
32. CID: 4201, Drug Name: Minoxidil, vasodilator, K channel modulator
33. CID: 3690, Drug Name: Ifosfamide, antineoplastic, DNA synthesis inhibitor
34. CID: 4075, Drug Name: 5-Aminosalicylic acid, PPAR-gamma activator, Inflammatory bowel disease, PPAR-gamma facilitates GABAergic transmission.
35. CID: 2284, Drug Name: Baclofen, antispasmodic agent, muscle relaxant, GABA mimetic
36. CID: 1645, Drug Name: 3-aminobenzamide, PARP inhibitor, apoptosis inhibitor, PARPs' inhibit synapse formation in regenerating axons,
37. CID: 54670067, Drug Name: l-Ascorbic acid, GABA mimic
38. CID: 3446, Drug Name: Gabapentin, antiepileptic, GABA mimetic, decreases the activity of calcium channels