Draft version March 16, 2025
Typeset using LaTeX **twocolumn** style in AASTeX7

Identifying novel anxiolytic drugs for selective modulation of the α_2 subunit of $GABA_A$

Asmani Yamin[1, *] And Lian Wang[1, †]

In submission for CS 229 (Machine Learning)

[1] *Stanford University*

ABSTRACT

The goal of this paper is to predict molecules that selectively bind to the α_2 subunit of the $GABA_A$ receptor. $GABA_A$ receptors are ligand-gated, G-coupled, ion channels that regulate GABA, the primary inhibitory neurotransmitter in the central nervous system (CNS) (Olsen and Sieghart 2008). Low levels of GABA activity are associated with several neurological disorders, including anxiety, depression, and epilepsy. Most drugs that treat these disorders, such as benzodiazepines, bind to the $GABA_A$ receptor and enhance its sensitivity to GABA, thus increasing inhibitory activity. However, these drugs typically lack specificity in which subunit of the receptor they bind to, often binding to the α_1 subunit of GABA, which results in undesirable side effects like sedation and cognitive impairment (Zhu et al. 2022). By identifying molecules that selectively bind to the α_2 unit of the $GABA_A$ receptor, we contribute to enhancing drug specificity and reducing unwanted side effects from off-target binding.

1. INTRODUCTION

We begin by acquiring a large set of virtual molecules, which we then screen for BBB permeability using both known constraints on BBB permeability and classification using machine learning methods (Random Forest, XGBoost, and neural network classifiers). Then, molecules are further screened $GABA_A$ drug-likeness using a Message-Passing Neural Network (MPNN) that operates on graph representations of molecules.

1.1. *ML for Molecular Property Prediction*

Our machine learning models are concerned with molecular property prediction (specifically, predicting a molecule's BBB permeability and its binding affinity to the $GABA_A$ receptor). Traditional ML methods, especially decision tree ensemble learning methods, have been extensively applied to this problem (Muratov et al. 2020; Deng et al. 2023). These methods typically operate on representations of molecules as collections of physio-chemical descriptors (e.g. molecular weight). One major advantage of these models is their interpretability, as they can identify key molecular features that contribute to the target property. Thus they work well for properties that can largely be predicted by known descriptors. They are also typically more computationally efficient than neural network alternatives. However, one major limitation is the burden on having a manually engineered rich set of molecular descriptors.

Recent work has turned to neural networks to capture high-dimensional relationships. One class of such models are sequence-based models used in language processing tasks, such as recurrent neural networks (Shi et al. 2020) and transformer-based models (Fabian et al. 2020; Chithrananda et al. 2020). These models operate on molecules represented as SMILES strings (which encode the full chemical structure of the molecule using ASCII characters), allowing models to learn relevant features from the molecule's chemical structure, removing the reliance on manually engineered features. Another class of neural network models are graph neural networks (GNNs), which directly manipulate molecules as graphs. This representation allows the model to learn relevant molecule-level features (without the intermediate translation into SMILES strings) and incorporates richer atom- and bond-level features, as well as 3D spatial features. Many variants of GNNs, specifically in the framework of MPNNs (Gilmer et al. 2017), have been adopted for molecular property prediction (Heid et al. 2024; Swanson et al. 2024; Buterez et al. 2024).

1.2. *Developments in $GABA_A$ Drug Design*

Over the past decade, research into drug design targeting the $GABA_A$ receptor has advanced significantly, focusing on developing compounds with improved efficacy, selectivity, and safety profiles.

Benzodiazepines have long been utilized as $GABA_A$ receptor positive allosteric modulators (PAMs) for their anxiolytic and sedative properties. Recent efforts have aimed at designing subtype-selective PAMs to minimize side effects. For instance, zolpidem exhibits selectivity for α_1-containing $GABA_A$ receptors, offering hypnotic effects with reduced anxiolytic activity (McHugh et al. 2023). Similarly, adipiplon, an α_3-selective drug, has been developed to target specific receptor subtypes, potentially reducing undesirable effects associated with non-selective activation (Berezhnoy et al. 2008).

Email: ayamin@stanford.edu, lianwang@stanford.edu
* Department of Bioengineering
† Department of Statistics

2. BACKGROUND ON THE BLOOD-BRAIN BARRIER (BBB)

The blood–brain barrier (BBB) is a selectively permeable interface that separates the CNS from the blood, ensuring CNS homeostasis by regulating molecule, ion, and cell passage. Composed mainly of brain capillary endothelial cells with tight junctions, the BBB restricts paracellular transport and permits only small, lipophilic molecules to cross by passive diffusion, while efflux transporters such as P-glycoprotein actively remove many drugs.

Although this protective barrier shields the brain from toxins and pathogens, it also presents a significant challenge for CNS drug development—over 98% of small molecules and nearly all large biologics fail to achieve therapeutic brain concentrations (Wu et al. 2023). Consequently, strategies like prodrug design, nanoparticle-based delivery, and chemical modifications are being explored to enhance BBB permeability.

3. NUMERICAL BBB FILTERING CRITERIA AND COMPUTATIONAL OPTIMIZATION

3.1. *Data*

The Enamine REAL database is the largest enumerated database of synthetically feasible molecules (Enamine 2024). It comprises over 9.6 billion molecules which comply with Ro5 (Lipinski et al. 1997) and Veber (Veber et al. 2002) criteria for drug-likeness: MW$\leq$ 500, SlogP$\leq$5, HBA$\leq$10, HBD$\leq$5, RotBonds$\leq$10, and TPSA$\leq$140. 7.5 billion molecules were processed from this dataset. The B3DB database (Meng et al. 2021) was cleaned and utilized for training relevant BBB permeability models; 7,807 molecules were utilized with a class distribution of 4,956 BBB+ and 2,851 BBB-.

3.2. *Stage 1 Constraints*

To identify molecules with a high likelihood of crossing the blood-brain barrier (BBB), we applied physicochemical constraints based on established research (Pajouhesh and Lenz 2005). The selected criteria—molecular weight (MW), lipophilicity (sLogP), hydrogen bonding (HBA and HBD), topological polar surface area (TPSA), rotatable bonds, and fraction of sp^3 carbons (Fsp^3)—are known to correlate with passive diffusion and transporter-mediated uptake across the BBB.

MW was limited to 180–400 Da, since compounds above 400 Da tend to diffuse poorly due to size constraints (Wager et al. 2010). sLogP was restricted to 1.8–3.6 as values above 3.6 can lead to high nonspecific binding and increased efflux by P-glycoprotein, while lower values reduce passive permeability (Mahar Doan et al. 2002). TPSA was confined to 20–60 $Å^2$, as molecules with TPSA above 90 $Å^2$ exhibit markedly reduced penetration (Cornelissen et al. 2023). Hydrogen bond acceptors were capped at 6 and donors at 2 to minimize excessive hydrogen bonding that hinders membrane diffusion. Constraints on rotatable bonds ($\leq$6) and an Fsp^3 range of 0.3–0.6 were imposed to balance flexibility and rigidity, and protein–protein interaction modulators were excluded, as these typically target large extracellular surfaces rather than penetrate the CNS.

The filtering pipeline leverages CuPy to parallelize these constraints via boolean masks on the GPU, while LZ4 compression and chunk-based streaming. This strategy reduced the SMILES-based dataset from 7.5 billion to around 800 million.

3.3. *Stage 2 Constraints*

A second-stage filter refines the molecular dataset after BBB permeability screening, incorporating additional drug-likeness criteria, statistical downsampling, and structural features (Landrum 2006). Instead of outright rejection, a small fraction of deficient molecules is retained to preserve diversity for optimization.

Unwanted Functional Groups (5% Retained): Reactive moieties (epoxides, aldehydes, Michael acceptors), excessive polarity (phosphates, sulfates), and charged groups (quaternary ammoniums) are mostly filtered out (Bruns and Watson 2012), but 5% are kept for edge-case modifications.
Highly Conjugated Systems (10% Retained): Molecules with $>$ 3 aromatic rings are downsampled to 10% to balance strong binding against solubility/metabolic instability (Ritchie and Macdonald 2009).
High Flexibility, Low MW (3% Retained): MW $<$350 Da and $>$5 rotatable bonds are mostly removed due to permeability concerns, but 3% are retained for unique CNS-active structures (Dehnbostel et al. 2024).
Poor TPSA/sLogP Ratio (2% Retained): Molecules with TPSA/sLogP $>$20 (indicative of permeability/lipophilicity issues) are mostly removed, retaining only 2% (Dehnbostel et al. 2024).
Low QED (2% Retained): Only 2% of compounds with QED $<$0.45 are kept, ensuring some unconventional but lead-like scaffolds remain (Ahmad et al. 2023).

This CUDA-accelerated filtering reduced the dataset from 800M to 50M molecules while maintaining chemical diversity for further analysis.

4. FEATURE-BASED BBB LEARNING

Since feature-based learning on massive molecular datasets is computationally prohibitive on individual machines, we implemented a three-stage cascade model that progressively refines the dataset by incrementally increasing feature complexity at each stage. This hierarchical strategy efficiently filters out improbable CNS-active molecules early in the pipeline, significantly reducing computational overhead for subsequent stages. By training each cascade stage with distinct subsets of features, we minimize cumulative false-positive and false-negative errors, a common limitation observed in stacked modeling approaches. Notably, the final neural network classifier demonstrated a higher false-positive

rate but lower false-negative rate compared to the preceding second-stage XGBoost model. This behavior is beneficial as earlier cascade stages effectively eliminate most irrelevant candidates, allowing the NN to focus computational resources on a smaller, highly relevant subset. Consequently, the neural network's increased sensitivity ensures fewer genuinely BBB-permeable molecules are incorrectly rejected, enhancing the model's predictive reliability. Such cascade-based approaches have been previously shown to improve accuracy and computational efficiency in large-scale classification tasks, e.g., computer vision (Viola and Jones 2001). Detailed hyperparameter justification and model figures are provided in Appendix A and Appendix B.

4.1. *Stage 1: Fast Screening with Random Forest*

In Stage 1, each molecule is rapidly screened using an RF classifier based on a 23-dimensional feature vector. For a given molecule, we extract features $\mathbf{x} = \begin{bmatrix} \text{MW} & \dots \end{bmatrix}^\top \in \mathbb{R}^{23}$. Each decision tree in the forest computes a binary decision $h_t(\mathbf{x})$ and the final RF probability is given by the ensemble average

$$p(\mathbf{x}) = \frac{1}{T} \sum_{t=1}^{T} h_t(\mathbf{x}), \tag{1}$$

with $T = 200$ trees. A molecule is classified as BBB-penetrable if $p(\mathbf{x}) \geq \tau_1$, where the threshold τ_1 maximizes the F1 score on a validation set. An AUROC of 0.952 and AP of 0.963 were achieved on a threshold of 0.399.

4.2. *Stage 2: Medium Screening with XGBoost*

Stage 2 refines the candidate list by incorporating richer chemical information. The feature vector here is a concatenation of: 13 classical descriptors $\mathbf{d} \in \mathbb{R}^{13}$, 9 Stage 1-specific features $\mathbf{s} \in \mathbb{R}^9$, Fingerprint features including MACCS keys $\mathbf{m} \in \mathbb{R}^{167}$, ECFP4 $\mathbf{e} \in \mathbb{R}^{1024}$, and substructure fingerprints $\mathbf{f} \in \mathbb{R}^{308}$. Thus, the full feature vector is:

$$\mathbf{x} = \begin{bmatrix} \mathbf{d} & \mathbf{s} & \mathbf{m} & \mathbf{e} & \mathbf{f} \end{bmatrix} \in \mathbb{R}^{13+9+167+1024+308} = \mathbb{R}^{1521}.$$

After applying a variance threshold and a feature importance filter (via a preliminary XGBoost model), the final classifier is trained using gradient boosting. Its additive model is expressed as

$$f(\mathbf{x}) = \sum_{k=1}^{K} f_k(\mathbf{x}), \tag{2}$$

where each f_k is a regression tree. The training objective minimizes

$$\mathcal{L}^{(K)} = \sum_{i=1}^{N} \ell\left(y_i, f(\mathbf{x}_i)\right) + \sum_{k=1}^{K} \Omega\left(f_k\right), \tag{3}$$

with a logistic loss ℓ and regularization term Ω controlling model complexity. A molecule is advanced if the predicted probability

$$p(\mathbf{x}) = \sigma\left(f(\mathbf{x})\right) \geq \tau_2, \tag{4}$$

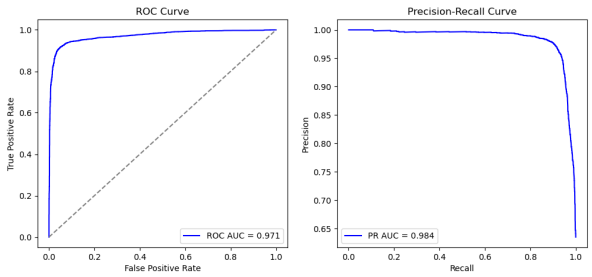


Figure 1. ROC and PR curve for Stage 3 NN. AUROC of 0.971 and PR AUC of 0.984 achieved on a threshold of 0.108.

with $\sigma(z) = \frac{1}{1+e^{-z}}$ and threshold τ_2 tuned to meet a target pass ratio. An AUROC of 0.955 and AP of 0.961 were achieved on a threshold of 0.551.

4.3. *Stage 3: Precise Screening with Neural Network*

In Stage 3, a deep neural network (NN) is employed on a comprehensive feature set that includes advanced fingerprints (ECFP6, MACCS, Klekota-Roth, and EState) and a full set of descriptors. Before training, the features were scaled and normalized. The NN architecture comprises: two fully connected layers with ReLU activation, interleaved with dropout layers ($p = 0.3$) to prevent overfitting; followed by an output linear layer with sigmoid transformation. It can be written as:

$$\mathbf{z}^{(1)} = \text{ReLU}\left(W^{(0)}\tilde{\mathbf{x}} + \mathbf{b}^{(0)}\right) \tag{5}$$

$$\mathbf{z}^{(2)} = \text{ReLU}\left(W^{(1)}\mathbf{z}^{(1)} + \mathbf{b}^{(1)}\right) \tag{6}$$

$$\hat{y} = \sigma\left(W^{(2)}\mathbf{z}^{(2)} + b^{(2)}\right) \tag{7}$$

where $W^{(0)} \in \mathbb{R}^{292\times m}$, $W^{(1)} \in \mathbb{R}^{188\times 292}$, $W^{(2)} \in \mathbb{R}^{1\times 188}$, and σ is the sigmoid function.

The NN is trained with binary cross-entropy loss and optimized via Adam (Kingma and Ba 2017). A molecule is classified as BBB-penetrable if $\hat{y} \geq \tau_3$, where τ_3 is initially set from the training set (0.108) but later dynamically adapted based on score distributions.

4.4. *Dynamic Threshold Adaptation*

To maintain target pass ratios across the cascade, score distributions are logged at each stage. Suppose S is the set of prediction scores for a stage; the threshold τ is dynamically computed as the $(1-\rho)$-th percentile:$\tau = \text{Percentile}(S, 100 \times (1-\rho))$, where ρ is the target pass ratio (0.5 for Stage 1, 0.2 for Stage 2, and 0.1 for Stage 3). This ensures the pipeline maintains consistent reduction of the candidate pool while minimizing false negatives and false positives. From 50 million molecules, 1 million were selected.

4.5. *BBB Feature Importance*

Feature importance analysis of stage 3 (NN) training was expected. All top features corresponded to a variation of

presence of electronegative atoms in the context of the H-bond donors/acceptors and polarizable surface area, including those from custom fingerprints. See Fig. 6d.

5. FILTERING FOR $GABA_A$ BINDING AFFINITY

We now further filter our BBB-permeable molecules by training a GNN to classify molecules that bind with (any subunit of) the $GABA_A$ receptor.

5.1. *Data*

To create the dataset, we scraped the IUPHAR/BPS database (Harding et al. 2024) to identify 45 known allosteric modulators of $GABA_A$ as positive examples, and we randomly sampled 1,000 diverse molecules not recorded to interact with $GABA_A$ from the B3DB database as negative examples. Train, validaton, and test sets were split 60-20-20 on the 1,045 examples.

A molecule can be represented as a graph whose nodes are atoms and edges are bonds. A graph can be represented by three matrices: a node feature matrix $V \in \mathbb{R}^{n_v \times d_v}$ where n_v is the number of nodes in the graph and d_v is the feature dimension of a node; an edge feature matrix $E \in \mathbb{R}^{n_e \times d_e}$ where n_e is the number of edges in the graph and d_e is the edge feature dimension; and an adjacency matrix $A \in \mathbb{R}^{2 \times n_e}$ where each row records the two nodes connecting each edge. We populate the node and edge feature vectors of a given molecule by extracting features of individual atoms and bonds in the molecule from its SMILES string. Note that features are interpretable at the atom- and bond-level, rather than the molecular level.

5.2. *Model Architecture*

The core of our model is two graph convolutional layers. Both layers follow the general format of an MPNN, where node representations are transformed by taking in "messages" from neighboring nodes and/or edges. Mathematically, we have

$$\mathbf{x}_v^{(l+1)} = \sum_{u \in \mathcal{N}(v)} f(\mathbf{x}_u^{(l)}, \mathbf{e}_{(u,v)}), \tag{8}$$

where $\mathbf{x}_v^{(l+1)}$ is the updated representation of node v, $\mathcal{N}(v)$ is the set of v's neighboring nodes, $\mathbf{x}_u^{(l)}$ is the current representation of node u, $\mathbf{e}_{(u,v)} \in \mathbb{R}^{d_e}$ is the edge feature vector between u and v, and $f(\cdot)$ is a transformation or "message" function.

The node feature matrix is first passed through a Graph Convolutional Network (GCN) layer (Kipf and Welling 2017), which is defined as

$$\mathbf{x}_v^{(l+1)} = \sum_{u \in \mathcal{N}(v) \cup \{v\}} \frac{1}{\sqrt{\deg(u) \cdot \deg(v)}} \cdot \left(W^\top \mathbf{x}_u^{(l)}\right) + \mathbf{b}, \tag{9}$$

where $\mathbf{x}_v^{(l)} \in \mathbb{R}^{d_v}$ and $\mathbf{x}_v^{(l+1)} \in \mathbb{R}^{d_h}$.

The updated node feature vector is then passed through a ReLU activation, then a custom layer that resembles a GCN but makes use of edge features in addition to node features, defined as

$$\mathbf{h}_e = W_e \mathbf{e}_{(u,v)} + \mathbf{b}_e \tag{10}$$

$$\mathbf{x}_v^{(l+1)} = \sum_{u \in \mathcal{N}(v)} W \left(\text{concat}(\mathbf{x}_u^{(l)}, \mathbf{h}_e)\right) + \mathbf{b}. \tag{11}$$

where $\mathbf{h}_e \in \mathbb{R}^{d_h}$, $\mathbf{x}_v^{(l)} \in \mathbb{R}^{d_h}$ (outputs of GCN layer), and $\mathbf{x}_v^{(l+1)} \in \mathbb{R}^{d_h}$.

The node feature vector is again passed through ReLU activation, then a global mean pooling layer, which aggregates node-level features into a single graph-level representation, defined as

$$\mathbf{h}_G = \frac{1}{|V|} \sum_{v \in V} \mathbf{x}_v. \tag{12}$$

Lastly, the pooled representation is linearly transformed to produce a scalar logit, which, after sigmoid transformation, represents the probability of the positive label. The model was implemented with PyTorch (Paszke et al. 2019) and PyTorch Geometric (Fey and Lenssen 2019), and metrics with scikit-learn (Pedregosa et al. 2011).

5.3. *Training and Results*

We train our model using stochastic gradient descent with a learning rate of 0.001 and batch size of 8 for 150 epochs. Training hyperparameters were selected to minimize validation AUROC. Due to the imbalanced dataset, we use weighted binary cross-entropy loss where positive examples are weighted by n_0/n_1, the ratio of negative to positive examples.

We report AUROC—the area under the ROC curve plotting TPR against FPR for different decision thresholds—which measures how well the model separates classes. Our model achieved a validation AUROC of 99% and a test AUROC of 95%, as shown in Fig. 2 (left), where the curve obtains high TPR at low FPR.

We also examined precision and recall; since our model is used as a preliminary filter for molecules that may interact with the $GABA_A$ receptor, we prioritize high recall over precision. As Fig. 2 (right) shows, precision decreases as recall increases, suggesting that the model produces more false positives as recall increases.

The high AUROC, along with the consistent decrease of validation loss with training loss (Fig. 7), suggests that our model generalized well. Regularization via weighted loss may have helped mitigate overfitting due to class imbalance in the training data.

We performed agglomerative clustering (Zepeda-Mendoza and Resendis-Antonio 2013) on the latent space embeddings of molecules to explore how well the embeddings captured underlying structure in the data. For a range of cluster numbers k, we calculated silhouette score (Shahapure and Nicholas 2020) to measure quality of cluster separation and Adjusted Rand Index (Halkidi et al. 2002) to measure how well clusters recovered true labels (Fig. 8). Low values

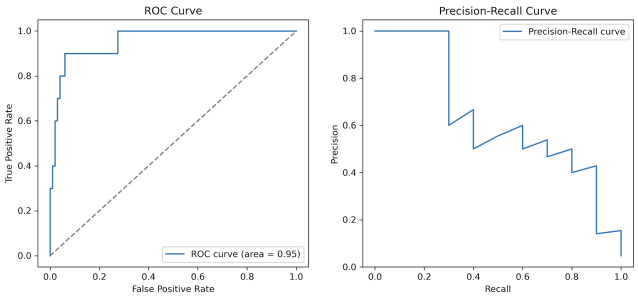


Figure 2. ROC curve and precision-recall curve on test set

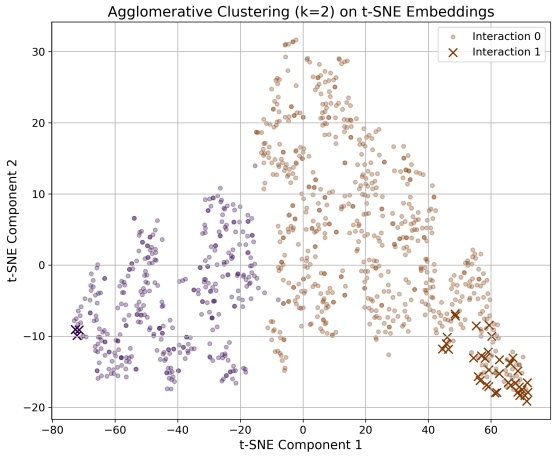


Figure 3. Embeddings in 2D t-SNE space. Colors correspond to agglomerative clusters with $k = 2$, and shapes (cross or dot) correspond to positive and negative labels.

of these metrics suggest clustering is suboptimal for all ks tested.

To visualize the embeddings, we employ t-SNE (van der Maaten and Hinton 2008), which projects high-dimensional data into 2D (Fig. 3). While the learned clusters do not align with true binary labels, we observe an interesting pattern: The positive labels form two clusters separated along t-SNE component 1. This suggests that there may be underlying substructure within the positive examples that is not captured by class labels, and binary classification may not be the optimal task to learn the data. Clustering and t-SNE were performed using scikit-learn.

5.4. *Filtering of BBB Penetrable Molecules*

Filtering 1 million BBB penetrant candidates for $GABA_A$ receptor binding affinity resulted in 63,169 positive samples. The top 2 probable drugs are visualized in Fig. 4 and the top 5 scaffolds in Fig. 9. In the post-inference analysis, we learned that predicted PAMs of the α_2 subunit of $GABA_A$ tended to be polycyclic and Nitrogen-dominant. The highest probable candidates contained halogens (Br, Cl) which could serve many functions, such as improved metabolic resistance,

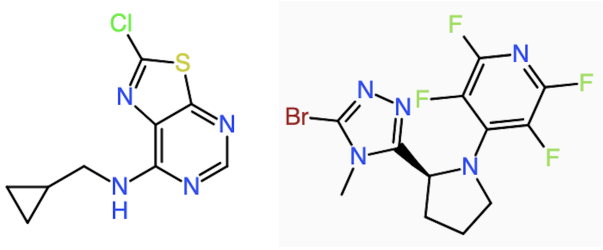


Figure 4. Top 2 $GABA_A$ α_2 PAM drug candidates

improved lipophilicity, as a salt counter-ion, or as a reactive handle to anchor the drug's geometry in the allosteric site.

6. CONCLUSION

We have implemented a procedure to screen a large set of molecules for BBB permeability and binding affinity with the $GABA_A$ receptor, using both ensemble learning methods and neural networks. For prediction of $GABA_A$ receptor interaction, one main limitation is the size and class imbalance of the dataset. We will thus explore transfer learning methods by pretraining models on a larger molecular database to learn important features on a range of property prediction tasks. We will also explore other message functions that better exploit both node and edge features. Moreover, we will incorporate molecule-level features selected by the BBB permeability models and 3D bond features such as bond angle and length.

Beyond improving current models, we will also implement the second half of our pipeline, where we will train a neural network to learn α_2 subunit binding selectivity. We will input the screened molecules into a variational autoencoder to generate a larger set of novel drug candidates. We will then perform docking studies to assess binding affinity with both the α_1 and α_2 subunits. The docking scores will feed into a directional loss function that positively weighs α_2 binding affinity and negatively weighs α_1 binding affinity, which will be used to train the neural network to learn binding specificity.

7. CONTRIBUTIONS

A.Y. implemented initial constraint-based BBB permeability filtering. A.Y. selected and implemented BBB permeability classifiers. L.W. performed exploratory model selection (not reported here). L.W. designed and implemented the MPNN for $GABA_A$ binding classification. A.Y. designed the overall pipeline for drug discovery. A.Y. and L.W. contributed equally to writing and data collection.

APPENDIX

A. HYPERPARAMETER RATIONALE

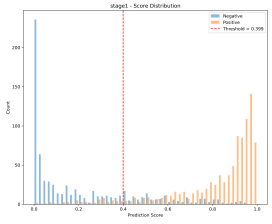
This section details the rationale behind hyperparameter selection in our three-stage cascade model for BBB permeability prediction. Choices were guided by computational efficiency, model performance, and empirical optimization.

Stage 1 (Random Forest): The Stage 1 model employed a Random Forest (RF) classifier consisting of 200 decision trees, chosen to balance predictive accuracy with computational efficiency, as the marginal gain in accuracy diminishes significantly beyond this number. A maximum depth of 12 provided sufficient complexity for capturing molecular feature interactions while preventing overfitting, given the moderate dimensionality (23 features). Requiring a minimum of 5 samples for splits and at least 2 samples per leaf enhanced model stability by reducing the likelihood of overly specific partitions. Class weighting was balanced to mitigate class imbalances commonly encountered in molecular datasets. The threshold (0.763) was empirically tuned to achieve approximately a 50% pass ratio, maximizing the F1 score to maintain a balance between precision and recall in initial screening stages.

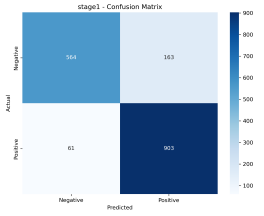
Stage 2 (XGBoost): The XGBoost model in Stage 2 utilized an enriched feature set, including MACCS fingerprints, ECFP4 fingerprints, and substructure descriptors, substantially increasing feature dimensionality. To manage this complexity, we first applied a Variance Threshold filter (0.01) to remove low-variance features, enhancing computational efficiency without losing meaningful information. Feature selection via an initial XGBoost model (100 trees, maximum depth 6, learning rate 0.1) retained only features above median importance, further reducing noise and redundancy. The final XGBoost model parameters were finely tuned: a maximum depth of 8 and a conservative learning rate (0.01) promoted stable convergence and improved generalization. The minimum child weight (2) and L1 (0.1) and L2 (1.0) regularization parameters prevented overfitting by constraining model complexity. The scale_pos_weight parameter dynamically balanced class distributions. Early stopping was employed after 50 iterations without improvement to prevent unnecessary computation and mitigate overfitting risks. The classification threshold (0.956) was selected to pass approximately 20% of molecules from the previous stage, despite the train threshold being around 0.55, effectively increasing precision while retaining critical molecular diversity for further analysis.

Stage 3 (Neural Network): The final neural network stage was designed for maximum precision, employing a deep, fully connected architecture with two hidden layers (292 and 188 neurons, respectively). These dimensions were empirically determined through iterative testing to balance representational capacity and computational tractability. Dropout (rate 0.3) after each hidden layer significantly mitigated overfitting by preventing co-adaptation of neurons, crucial for handling the large and complex feature set encompassing advanced fingerprints and descriptor-based features (MACCS, ECFP6, 2D Klekota-Roth, and 2D EState fingerprints). Features were standardized with a scaler to ensure equal weighting during training, improving numerical stability and convergence behavior. The Adam optimizer, renowned for handling sparse and high-dimensional data, was employed with an exceptionally low learning rate (2.3×10^{-5}) and minimal weight decay (1.5×10^{-8}), again determined empirically, which is essential for achieving stable convergence and avoiding aggressive parameter updates that could destabilize training (Kingma and Ba 2017). Training proceeded over 139 epochs with early stopping (patience=10 epochs) to efficiently terminate training upon convergence plateauing. A small batch size (8) was deliberately chosen to maintain stable gradients given the complexity of data and constraints of GPU memory. The initial classification threshold (0.108) was derived from extensive validation studies and dynamically adjusted in practice to accommodate shifts in score distributions, ensuring the model retained sensitivity to genuine BBB-permeable molecules.

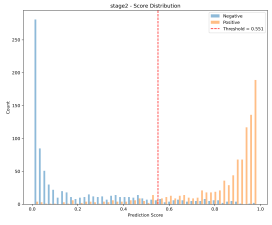
B. MACHINE LEARNING MODEL FIGURES



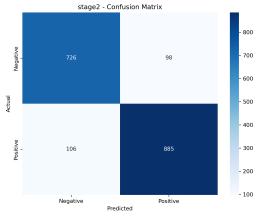
(a) Score Distribution of Stage 1 RF



(b) Confusion Matrix of Stage 1 RF

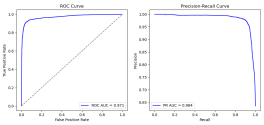


(c) Score Distribution of Stage 2 XGBoost

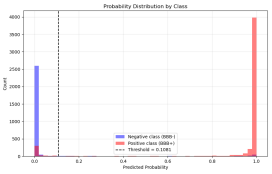


(d) Confusion Matrix of Stage 2 XGBoost

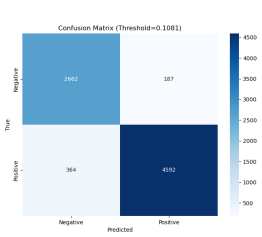
Figure 5. BBB Stage 1 and 2 Model Statistics



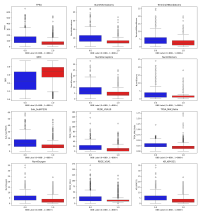
(a) PR Curve



(b) Probability Distribution



(c) Confusion Matrix



(d) Feature Importance

Figure 6. BBB Stage 3 NN Model Statistics

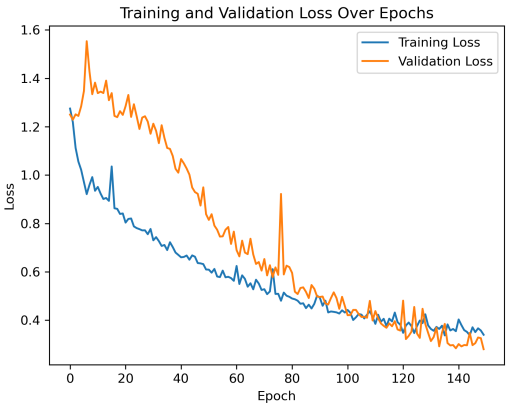


Figure 7. Training and Validation Loss for MPNN

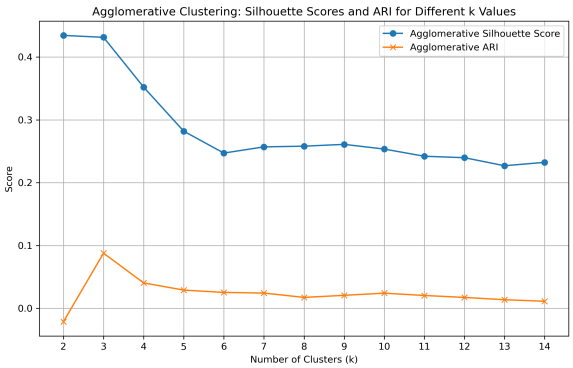


Figure 8. Silhouette and ARI Scores for Agglomerative Clustering of MPNN Embeddings

C. MOLECULE VISUALIZATIONS

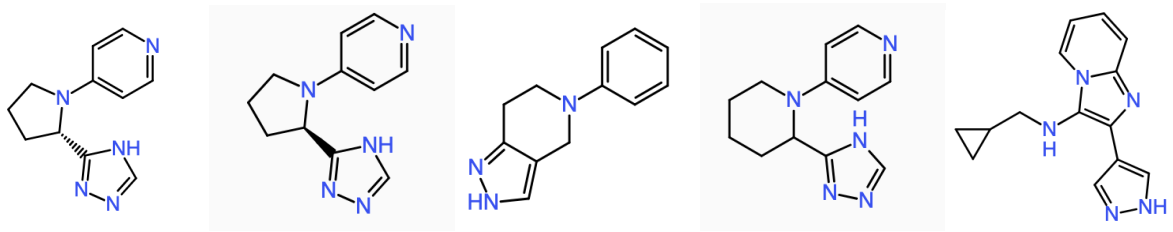


Figure 9. Top 5 Drug Scaffolds

REFERENCES

Ahmad, I., Kuznetsov, A. E., Pirzada, A. S., Alsharif, K. F., Daglia, M., and Khan, H. (2023). Computational pharmacology and computational chemistry of 4-hydroxyisoleucine: Physicochemical, pharmacokinetic, and DFT-based approaches. *Frontiers in Chemistry*, 11:1145974.

Berezhnoy, D., Gravielle, M. C., Downing, S., Kostakis, E., Basile, A. S., Skolnick, P., Gibbs, T. T., and Farb, D. H. (2008). Pharmacological properties of DOV 315,090, an ocinaplon metabolite. *BMC Pharmacology*, 8:11.

Bruns, R. F. and Watson, I. A. (2012). Rules for identifying potentially reactive or promiscuous compounds. *Journal of Medicinal Chemistry*, 55(22):9763–9772. PMID: 23061697.

Buterez, D., Janet, J. P., Kiddle, S. J., Oglic, D., and Lió, P. (2024). Transfer learning with graph neural networks for improved molecular property prediction in the multi-fidelity setting. *Nature Communications*, 15(1):1517. Publisher: Nature Publishing Group.

Chithrananda, S., Grand, G., and Ramsundar, B. (2020). ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. arXiv:2010.09885 [cs].

Cornelissen, F. M. G., Markert, G., Deutsch, G., Antonara, M., Faaij, N., Bartelink, I., Noske, D., Vandertop, W. P., Bender, A., and Westerman, B. A. (2023). Explaining blood-brain barrier permeability of small molecules by integrated analysis of different transport mechanisms. *Journal of Medicinal Chemistry*, 66(11):7253–7267.

Dehnbostel, F. O., Dixit, V. A., Preissner, R., and Banerjee, P. (2024). Non-animal models for blood–brain barrier permeability evaluation of drug-like compounds. *Scientific Reports*, 14(1):8908.

Deng, J., Yang, Z., Wang, H., Ojima, I., Samaras, D., and Wang, F. (2023). A systematic study of key elements underlying molecular property prediction. *Nature Communications*, 14(1):6395. Publisher: Nature Publishing Group.

Enamine (2024). Enamine REAL Database.

Fabian, B., Edlich, T., Gaspar, H., Segler, M., Meyers, J., Fiscato, M., and Ahmed, M. (2020). Molecular representation learning with language models and domain-relevant auxiliary tasks. arXiv:2011.13230 [cs].

Fey, M. and Lenssen, J. E. (2019). Fast graph representation learning with PyTorch Geometric.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural Message Passing for Quantum Chemistry. arXiv:1704.01212 [cs].

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2002). Cluster validity methods: Part i. *SIGMOD Record*, 31.

Harding, S., Armstrong, J., Faccenda, E., Southan, C., Alexander, S., Davenport, A., Spedding, M., and Davies, J. (2024). The IUPHAR/BPS guide to PHARMACOLOGY in 2024. *Nucleic Acids Research*.

Heid, E., Greenman, K. P., Chung, Y., Li, S.-C., Graff, D. E., Vermeire, F. H., Wu, H., Green, W. H., and McGill, C. J. (2024). Chemprop: A Machine Learning Package for Chemical Property Prediction. *Journal of Chemical Information and Modeling*, 64(1):9–17. Publisher: American Chemical Society.

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization.

Kipf, T. N. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks.

Landrum, G. (2006). Rdkit: Open-source cheminformatics. *GitHub repository*.

Lipinski, C., Lombardo, F., Dominy, B., and Feeney, P. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23:3–25.

Mahar Doan, K. M., Humphreys, J. E., Webster, L. O., Wring, S. A., Shampine, L. J., Serabjit-Singh, C. J., Adkison, K. K., and Polli, J. W. (2002). Passive permeability and P-glycoprotein-mediated efflux differentiate central nervous system (CNS) and non-CNS marketed drugs. *Journal of Pharmacology and Experimental Therapeutics*, 303(3):1029–1037.

McHugh, R. K., Votaw, V. R., Trapani, E. W., and McCarthy, M. D. (2023). Prevalence and correlates of the misuse of z-drugs and benzodiazepines in the national survey on drug use and health. *Frontiers in Psychiatry*, 14:1129447.

Meng, F., Xi, Y., Huang, J., and Ayers, P. W. (2021). A curated diverse molecular database of blood-brain barrier permeability with chemical descriptors. *Scientific Data*, 8(1):289.

Muratov, E. N., Bajorath, J., Sheridan, R. P., Tetko, I. V., Filimonov, D., Poroikov, V., Oprea, T. I., Baskin, I. I., Varnek, A., Roitberg, A., Isayev, O., Curtalolo, S., Fourches, D., Cohen, Y., Aspuru-Guzik, A., Winkler, D. A., Agrafiotis, D., Cherkasov, A., and Tropsha, A. (2020). QSAR without borders. *Chem. Soc. Rev.*, 49:3525–3564.

Olsen, R. W. and Sieghart, W. (2008). International union of pharmacology. lxx. subtypes of GABA-A receptors: classification on the basis of subunit composition, pharmacology, and function. update. *Pharmacological Reviews*, 60(3):243–260. Epub 2008 Sep 12.

Pajouhesh, H. and Lenz, G. R. (2005). Medicinal chemical properties of successful central nervous system drugs. *NeuroRx*, 2(4):541–553.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ritchie, T. J. and Macdonald, S. J. F. (2009). The impact of aromatic ring count on compound developability–are too many aromatic rings a liability in drug design? *Drug Discovery Today*, 14(21-22):1011–1020.

Shahapure, K. R. and Nicholas, C. (2020). Cluster quality analysis using silhouette score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 747–748.

Shi, T., Huang, S., Chen, L., Heng, Y., Kuang, Z., Xu, L., and Mei, H. (2020). A molecular generative model of ADAM10 inhibitors by using GRU-based deep neural network and transfer learning. *Chemometrics and Intelligent Laboratory Systems*, 205:104122.

Swanson, K., Liu, G., Catacutan, D. B., Arnold, A., Zou, J., and Stokes, J. M. (2024). Generative AI for designing and validating easily synthesizable and structurally novel antibiotics. *Nature Machine Intelligence*, 6(3):338–353. Publisher: Nature Publishing Group.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Veber, D. F., Johnson, S. R., Cheng, H.-Y., Smith, B. R., Ward, K. W., and Kopple, K. D. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry*, 5(12).

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9.

Wager, T. T., Hou, X., Verhoest, P. R., and Villalobos, A. (2010). Moving beyond rules: the development of a central nervous system multiparameter optimization (CNS MPO) approach to enable alignment of druglike properties. *ACS Chemical Neuroscience*, 1(6):435–449.

Wu, D., Chen, Q., Chen, X., Han, F., Chen, Z., and Wang, Y. (2023). The blood–brain barrier: Structure, regulation and drug delivery. *Signal Transduction and Targeted Therapy*, 8(1):217.

Zepeda-Mendoza, M. L. and Resendis-Antonio, O. (2013). *Hierarchical Agglomerative Clustering*, pages 886–887. Springer New York, New York, NY.

Zhu, S., Sridhar, A., Teng, J., Howard, R. J., Lindahl, E., and Hibbs, R. E. (2022). Structural and dynamic mechanisms of gabaa receptor modulators with opposing activities. *Nature Communications*, 13(1):4582.