

NeuroCADR: Drug
Repositioning to Reveal Novel
Anti-Epileptic Drug Candidates
Through an Integrated
Computational Approach

Abstract:

Drug repositioning is an emerging approach for drug discovery involving the reassignment of existing drugs for novel purposes. An alternative to the traditional *de novo* process of drug development, repositioned drugs are faster, cheaper, and less failure prone than drugs developed from traditional methods. Recently, drug repositioning has been performed *in silico* - databases of drugs and chemical information are used to determine interactions between target proteins and drug molecules to identify potential drug candidates. A proposed algorithm is NeuroCADR, a novel approach for drug repositioning via spherical k-means and k-nearest neighbor algorithms (KNN). Data sourced from several databases consisting of interactions between genes, proteins, and drug molecules were compiled into separate binarified datasets. These were inputted into an KNN machine learning algorithm that learned associations between these. The proposed method displayed a high level of accuracy, outperforming nearly all *in silico* approaches. NeuroCADR was performed on epilepsy, a condition that is characterized by seizures, periods of time with bursts of uncontrolled electrical activity in brain cells. Existing drugs for epilepsy can be ineffective and expensive, revealing a need for new antiepileptic drugs. NeuroCADR identified novel drug candidates for epilepsy that can be further approved through clinical trials. The algorithm was incorporated into a user-friendly website for medical professionals to determine possible drug combinations to prescribe a patient based on a patient's prior medical history. This project examines NeuroCADR, a novel approach to computational drug repositioning capable of revealing potential drug candidates in neurological diseases such as epilepsy.

Intro/Background:

The traditional method of developing drugs is time-consuming, expensive, and has substantial risk of failure. This process has five main stages: preclinical trials or discovery, review of safety, clinical research, FDA review, and post-market FDA safety monitoring. Developing a new drug with this method costs \$12 billion and takes 6+ years to develop, with many failing at the early stages of development, forcing patients with certain conditions to have unaffordable drug treatments. For some diseases, patients may have virtually no drug treatments.

One emerging alternative to the traditional drug development process is drug repositioning. Drug repositioning involves the reassignment of existing drugs for novel therapeutic purposes. Drug repositioning has several advantages over the traditional method of drug discovery. They are billions of dollars cheaper than traditionally created drugs, in addition to being much faster to develop and send to patients. These types of drugs are also less failure prone than traditional drugs as they have already passed clinical trials and exist in the market.

There are multiple approaches to drug repositioning. One is a clinical approach where patients with a certain condition are given potential drug candidates to test the effectiveness of the drug in treating the condition. Drug candidates are selected by analysis of patient tissue or blood. Drug repositioning has also recently been performed via an in silico approach - databases of drugs and chemical information are used to identify potential drug candidates by forming associations between drug structures and protein and genes.

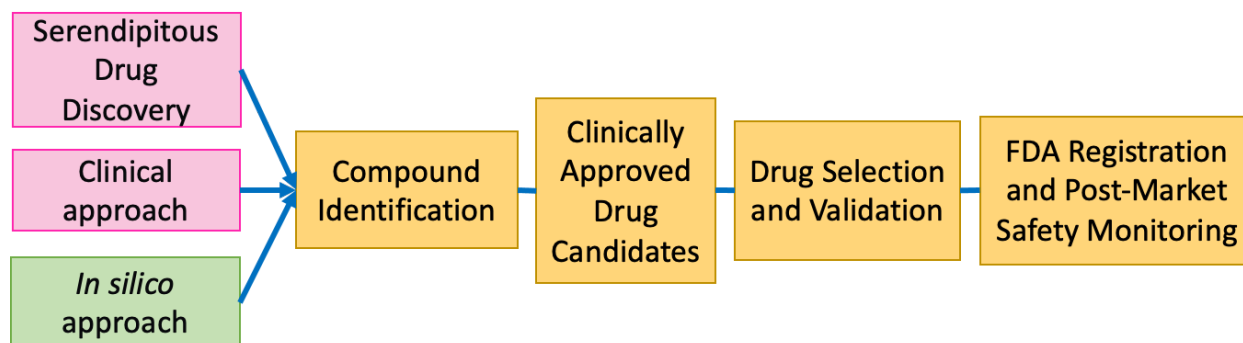


Figure 1: The drug repositioning process

One condition that could greatly benefit from drug repositioning is epilepsy. Epilepsy is a condition characterized by seizures, periods of time with bursts of uncontrolled electrical activity in brain cells. Not all types of seizures can be treated with anti-epileptic drugs (AEDs) and existing AEDs have many side effects, revealing a need for new AEDs. Epileptic drugs can also be expensive, and this cost can be reduced by the usage of previously existing drugs to treat epilepsy.

It was hypothesized that utilizing KNN (k-nearest neighbor) algorithms and spherical k-means, two types of machine learning algorithms, to perform drug repositioning would result in more potential repurposed drug candidates and more accurate drug predictions as the combination of supervised and unsupervised ML for classification and clustering would allow for more functionality

This algorithm eventually became NeuroCADR - a novel computational approach for drug repositioning. NeuroCADR can identify novel drug candidates for epilepsy that can be further approved through clinical trials and eventually used to treat patients. Furthermore, the integrated algorithm was incorporated in a user-friendly website that medical professionals can use to determine possible drug combinations to prescribe a patient based on a patient's prior medical history.

Data Compilation:

Data was sourced from several different well established datasets such as Drug Bank, DDInter, and HIPPIE. All datasets have been verified and used in numerous other scientific literature.

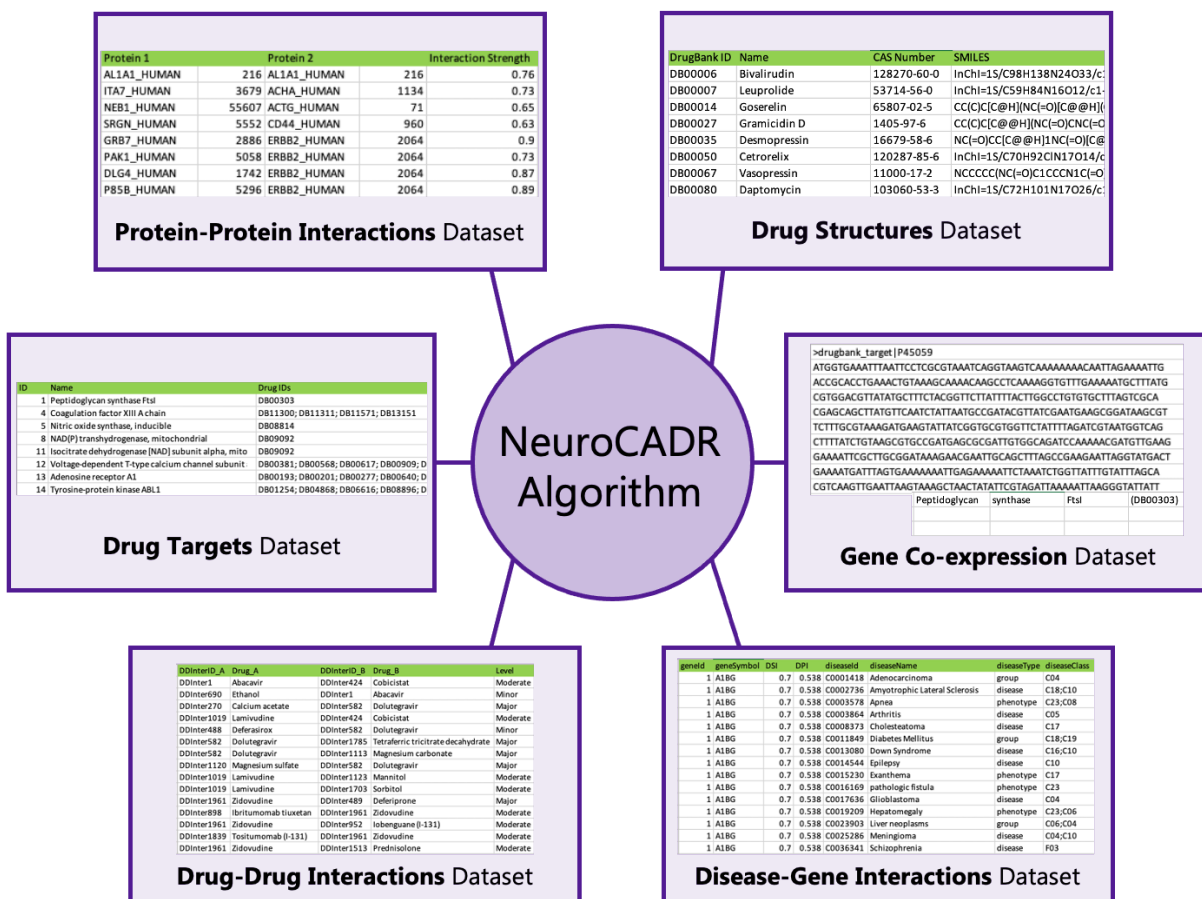


Figure 2: Dataset Combination

First, the Drug-Drug Interactions dataset (DDI) was compiled from the DDInter database. Data consisted of names of Drug A and B, along with their Drug IDs. An interaction severity level was then allotted to each drug pair using the keywords “Minor”, “Moderate”, or “Major” which were then translated into numbers 1, 2, and 3 respectively with 0 being no interaction for easier classification by the algorithm. This dataset consisted of 12,025 interactions.

DDInterID_A	Drug_A	DDInterID_B	Drug_B	Level
DDInter1	Abacavir	DDInter424	Cobicistat	Moderate
DDInter690	Ethanol	DDInter1	Abacavir	Minor
DDInter270	Calcium acetate	DDInter582	Dolutegravir	Major
DDInter1019	Lamivudine	DDInter424	Cobicistat	Moderate
DDInter488	Deferasirox	DDInter582	Dolutegravir	Minor
DDInter582	Dolutegravir	DDInter1785	Tetraferic tricitrate decahydrate	Major
DDInter582	Dolutegravir	DDInter1113	Magnesium carbonate	Major
DDInter1120	Magnesium sulfate	DDInter582	Dolutegravir	Major
DDInter1019	Lamivudine	DDInter1123	Mannitol	Moderate
DDInter1019	Lamivudine	DDInter1703	Sorbitol	Moderate
DDInter1961	Zidovudine	DDInter489	Deferiprone	Major
DDInter898	Ibritumomab tiuxetan	DDInter1961	Zidovudine	Moderate
DDInter1961	Zidovudine	DDInter952	Iobenguane (I-131)	Moderate
DDInter1839	Tositumomab (I-131)	DDInter1961	Zidovudine	Moderate
DDInter1961	Zidovudine	DDInter1513	Prednisolone	Moderate

Figure 3: Drug-Drug Interactions with strength of interaction

Drug Structures were compiled from DrugBank and were grouped with DrugBank IDs and CAS numbers. Structures were expressed in multiple forms, including the KEGG Compound ID and the SMILES files. This project principally used SMILES files as they are well defined and one of the most common chemical notation systems. 2716 drug structures were recorded.

DrugBank ID	Name	CAS Number	SMILES
DB00006	Bivalirudin	128270-60-0	InChI=1S/C98H138N24O33/c1-5-52(4)82(96(153)12
DB00007	Leuprolide	53714-56-0	InChI=1S/C59H84N16O12/c1-6-63-57(86)48-14-10-
DB00014	Goserelin	65807-02-5	CC(C)C[C@H](NC(=O)[C@@H](COC(C)(C)C)NC(=O)[C@
DB00027	Gramicidin D	1405-97-6	CC(C)C[C@H](NC(=O)CNC(=O)[C@@H](NC(=O)C(C)C)
DB00035	Desmopressin	16679-58-6	NC(=O)CC[C@H]1NC(=O)[C@H](CC2=CC=CC=C2)NC
DB00050	Cetrorelix	120287-85-6	InChI=1S/C70H92ClN17O14/c1-39(2)31-52(61(94)8-
DB00067	Vasopressin	11000-17-2	NCCCCC(NC(=O)C1CCCN1C(=O)C1CSCC(N)C(=O)NC(C
DB00080	Daptomycin	103060-53-3	InChI=1S/C72H101N17O26/c1-5-6-7-8-9-10-11-22-5

Figure 4: Drug Structures Dataset with structures expressed with the SMILES file

Drug Targets were taken from DrugBank. Data was organized by protein targets, which were linked to the gene name, gene and protein IDs, and drug IDs that target the specific protein. There are 1233 protein targets that were identified, resulting in 8000+ drugs.

ID	Name	Drug IDs
1	Peptidoglycan synthase FtsI	DB00303
4	Coagulation factor XIII A chain	DB11300; DB11311; DB11571; DB13151
5	Nitric oxide synthase, inducible	DB08814
8	NAD(P) transhydrogenase, mitochondrial	DB09092
11	Isocitrate dehydrogenase [NAD] subunit alpha, mitochondrial	DB09092
12	Voltage-dependent T-type calcium channel subunit alpha-1I	DB00381; DB00568; DB00617; DB00909;
13	Adenosine receptor A1	DB00193; DB00201; DB00277; DB00640;
14	Tyrosine-protein kinase ABL1	DB01254; DB04868; DB06616; DB08896;

Figure 5: Drug Targets Dataset showing drug target and corresponding drug IDs

Protein-Protein Interactions were taken from the HIPPIE database. Data included the names of two proteins along with their respective gene IDs. An interaction strength was displayed for each pair of proteins, with 0 being no interaction and 1 being very high interaction. The PPI dataset consisted of 82,872 interactions among various proteins in the human body.

Protein 1		Protein 2		Interaction Strength
AL1A1_HUMAN	216	AL1A1_HUMAN	216	0.76
ITA7_HUMAN	3679	ACHA_HUMAN	1134	0.73
NEB1_HUMAN	55607	ACTG_HUMAN	71	0.65
SRGN_HUMAN	5552	CD44_HUMAN	960	0.63
GRB7_HUMAN	2886	ERBB2_HUMAN	2064	0.9
PAK1_HUMAN	5058	ERBB2_HUMAN	2064	0.73
DLG4_HUMAN	1742	ERBB2_HUMAN	2064	0.87
P85B_HUMAN	5296	ERBB2_HUMAN	2064	0.89

Figure 6: Protein-Protein Interactions Dataset comparing the interaction strength between two proteins

Gene Co-expression sequences were taken from DrugBank. The data compiled measured the similarity of gene expression patterns between disease types in addition to the genetic sequences themselves. The GC dataset contained 95401 interactions with gene sequences.

>drugbank_target P45059			
ATGGTGAAATTAATTCCTCGCGTAAATCAGGTAAGTCAAAAAACAATTAGAAAATTG			
ACCGCACCTGAAACTGTAAAGCAAAACAAGCCTCAAAAGGTGTTTGAAAAATGCTTTATG			
CGTGGACGTTATATGCTTTCTACGGTCTTATTTTACTTGGCCTGTGTGCTTTAGTCGCA			
CGAGCAGCTTATGTTCAATCTATTAATGCCGATACGTTATCGAATGAAGCGGATAAGCGT			
TCTTTGCGTAAAGATGAAGTATTATCGGTGCGTGGTTCTATTTAGATCGTAATGGTCAG			
CTTTATCTGTAAGCGTGCCGATGAGCGCGATTGTGGCAGATCCAAAAACGATGTTGAAG			
GAAAATCGCTTGGCGATAAAGAACGAATTGCAGCTTTAGCCGAAGAATTAGGTATGACT			
GAAAATGATTAGTGAAAAAATTGAGAAAAATCTAAATCTGGTTATTGTATTAGCA			
CGTCAAGTTGAATTAAGTAAAGCTAACTATATTCGTAGATTAATAAAGGGTATTATT			

Peptidoglycan	synthase	FtsI	(DB00303)

Figure 7: Gene Co-expression dataset displaying gene sequences with respective polymers

Disease-Gene Interactions were compiled from DisGeNET. Data entries consisted of the gene ID, gene name, disease ID, disease name, and disease class. There are 50,000+ genes with 1,048,576 interactions. This dataset is not required for NeuroCADR to be run to identify drug candidates for epilepsy but is necessary for NeuroCADR to output results for other diseases.

genelid	geneSymbol	DSI	DPI	diseaseId	diseaseName	diseaseType	diseaseClass
1	A1BG	0.7	0.538	C0001418	Adenocarcinoma	group	C04
1	A1BG	0.7	0.538	C0002736	Amyotrophic Lateral Sclerosis	disease	C18;C10
1	A1BG	0.7	0.538	C0003578	Apnea	phenotype	C23;C08
1	A1BG	0.7	0.538	C0003864	Arthritis	disease	C05
1	A1BG	0.7	0.538	C0008373	Cholesteatoma	disease	C17
1	A1BG	0.7	0.538	C0011849	Diabetes Mellitus	group	C18;C19
1	A1BG	0.7	0.538	C0013080	Down Syndrome	disease	C16;C10
1	A1BG	0.7	0.538	C0014544	Epilepsy	disease	C10
1	A1BG	0.7	0.538	C0015230	Exanthema	phenotype	C17
1	A1BG	0.7	0.538	C0016169	pathologic fistula	phenotype	C23
1	A1BG	0.7	0.538	C0017636	Glioblastoma	disease	C04
1	A1BG	0.7	0.538	C0019209	Hepatomegaly	phenotype	C23;C06
1	A1BG	0.7	0.538	C0023903	Liver neoplasms	group	C06;C04
1	A1BG	0.7	0.538	C0025286	Meningioma	disease	C04;C10
1	A1BG	0.7	0.538	C0036341	Schizophrenia	disease	F03

Figure 8: Disease-Gene Interactions Dataset with several gene and disease attributes

Algorithm Construction and Training:

NeuroCADR utilized a KNN (k-nearest neighbors) algorithm and k-spherical means. KNN was chosen as it requires no training time, contrary to deep learning and other neural network types (CNN, RNN). KNN algorithms are relatively straightforward and only require tuning one parameter at a time (the value of k), which makes establishing associations between drugs and drug targets more streamlined. KNN is a supervised machine learning algorithm, requiring input data to analyze patterns and predict output data when given new unlabeled data. KNNs are also shown to have greater theoretical guarantees than other similar algorithms.

The primary assumption in a KNN algorithm is that similar data exists near each other. Therefore, associations can be made by finding the distance between two points with a given number of "neighbors", denoted as k . The selected k -value produces the least number of errors with the training data while still being able to make accurate predictions. In addition, KNNs use the concept of a k -fold cross validation. K -fold cross validation is a way to evaluate algorithms with new data by dividing it into k shuffled groups. In each of k iterations, one portion is set aside while the others are trained. The separated portion is then used as testing data.

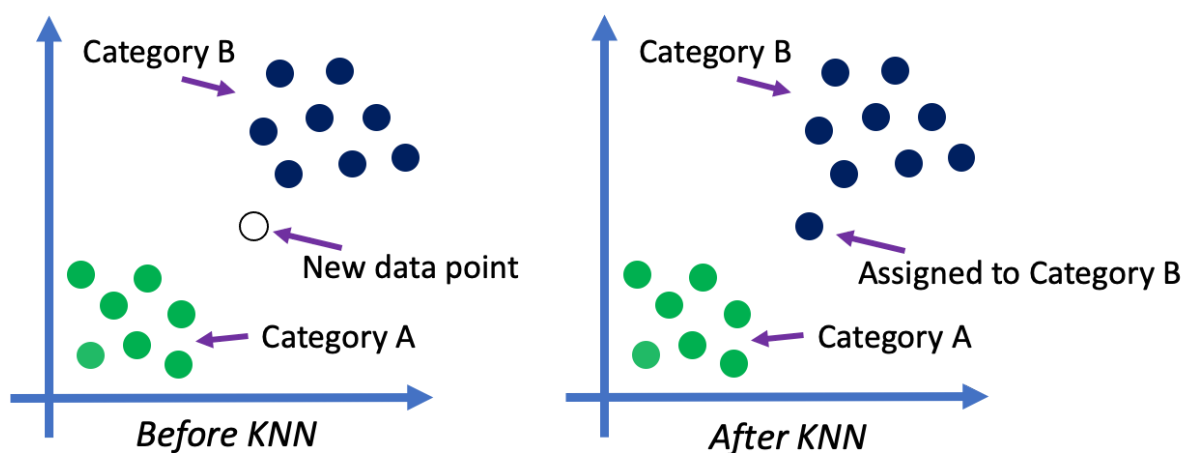


Figure 9: Conceptual diagram of KNN algorithm showing assignment of new data point

Another method employed in the NeuroCADR algorithm is spherical k-means. Spherical k-means is a method for clustering high-dimensional text data where associations are depicted as vectors using cosine dissimilarity and was used to initialize clusters between data. The centroid of each cluster is then used to initialize the matrices.



Figure 10: Five main steps used to create NeuroCADR.

Matrix Conversion:

First, datasets were converted to association matrices to allow for easier classification for the algorithm. A separate utility class parsed the data and mapped the drugs to their respective structures, diseases, and proteins.

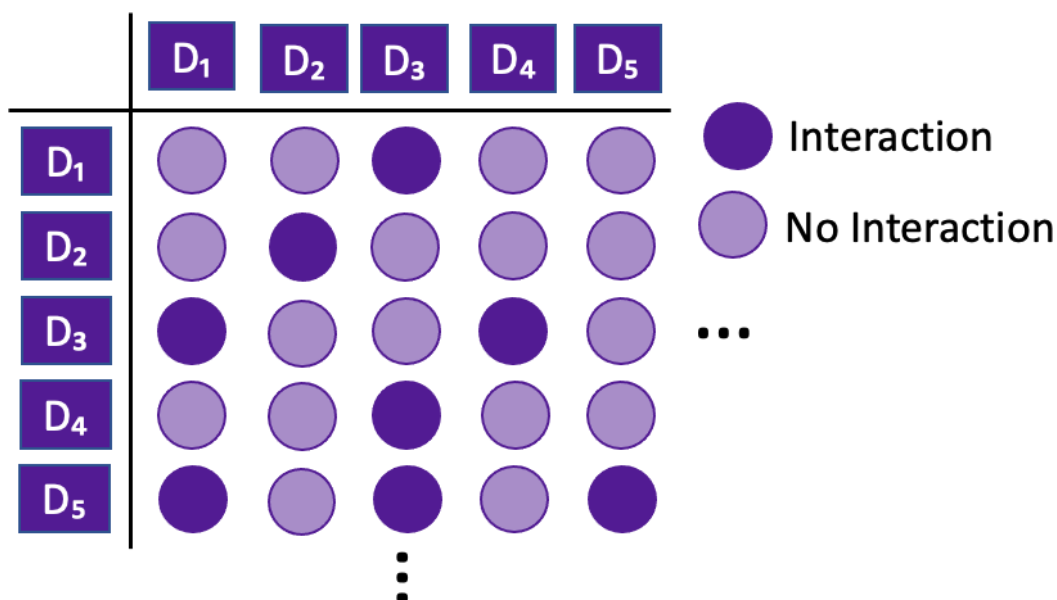
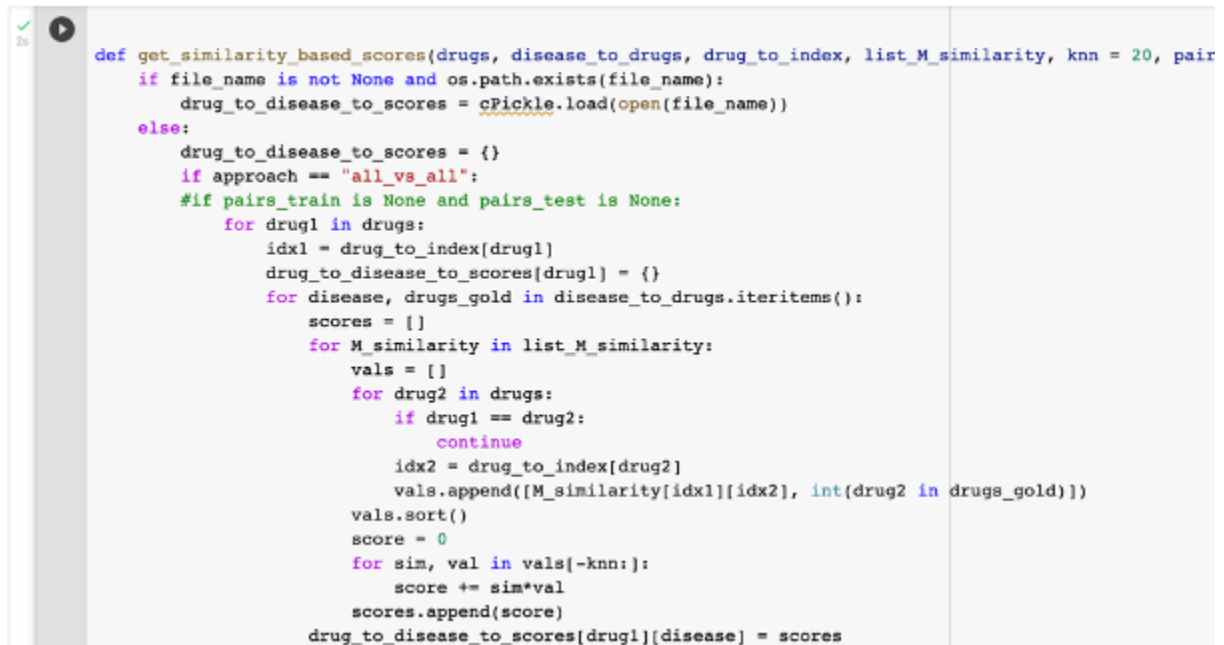


Figure 11: Graphical representation of Drug-Drug Interaction Matrix

A similarity function was run to score the similarity between each pair of drugs based on its attributes.



```
def get_similarity_based_scores(drugs, disease_to_drugs, drug_to_index, list_M_similarity, knn = 20, pairs_train=None):
    if file_name is not None and os.path.exists(file_name):
        drug_to_disease_to_scores = cPickle.load(open(file_name))
    else:
        drug_to_disease_to_scores = {}
        if approach == "all_vs_all":
            #if pairs_train is None and pairs_test is None:
            for drug1 in drugs:
                idx1 = drug_to_index[drug1]
                drug_to_disease_to_scores[drug1] = {}
                for disease, drugs_gold in disease_to_drugs.iteritems():
                    scores = []
                    for M_similarity in list_M_similarity:
                        vals = []
                        for drug2 in drugs:
                            if drug1 == drug2:
                                continue
                            idx2 = drug_to_index[drug2]
                            vals.append([M_similarity[idx1][idx2], int(drug2 in drugs_gold)])
                        vals.sort()
                        score = 0
                        for sim, val in vals[-knn:]:
                            score += sim*val
                        scores.append(score)
                    drug_to_disease_to_scores[drug1][disease] = scores
```

Figure 12: Similarity function comparing drugs (all vs. all)

Definition of Training Parameters:

Parameters for training were then defined. Features for training included the “drug”, “target”, and “label” The k-value for the KNN algorithm was chosen to be 20, representing the nearest number of drugs to check for to assign a repurposing score. The k-value was selected by running the algorithm with different values of k to find the optimal value. The value producing the least number of errors with the training data along while still being able to make accurate predictions was chosen. The number of folds for cross validation was set to 10, and the analysis was set to run 10 times.

```

# Get parameters
n_seed = 52345
random.seed(n_seed) # for reproducibility
features = ["drug", "target", "label"]
model_type = "logistic" # ML model
prediction_type = "disease" # predict drug-disease associations
output_file = "data/validation.dat" # file containing run parameters and co
n_proportion = 2 # proportion of negative instances compared to positives
n_subset = -1 # for faster results - subsampling data
knn = 20 # number of nearest drugs to check in the pharmacological space to
n_run = 10 # number of repetitions of cross-validation analysis
n_fold = 10 # number of folds in cross-validation
recalculate_similarity = True # whether the k-NN based repurposing score sh
# whether the drugs in the drug-disease pairs of the cross-validation folds
disjoint_cv = False

```

Figure 13: List of parameters to prepare for training

Matrix Factorization (MF) and KNN Method:

Matrix Factorization (MF) methods are commonly used in recommender systems. In the context of drug repositioning, MF methods were used to assign weights to drugs, which were then ranked to reveal the most plausible drugs. This type of method is called content-based filtering, where attributes of an item are trained to recommend items with similar properties. A matrix for algorithm validation was set aside as 10% of the testing data.

First, the association matrices were run through the method of Singular Value Decomposition (SVD), which facilitates the decomposition of large matrices into two or three smaller matrices. This MF method then separates the matrices into “user” and “item” matrices. The user matrices represent the drug targets and attributes of epilepsy, while the item matrices represent the available drugs and drug attributes.

An objective function was then chosen to minimize the difference between the user and item matrices to provide the most accurate drug recommendations. The method used to minimize

this function was Weighted Alternating Least Squares (WALS). WALS first scrambled each of the user and item matrices.

Next, it alternates between fixing the user matrix and solving for the item matrix and fixing the item matrix and solving for the user matrix. KNN was implemented here as matrix sizes of k were solved with each iteration. Each step decreases the difference, guaranteeing convergence.

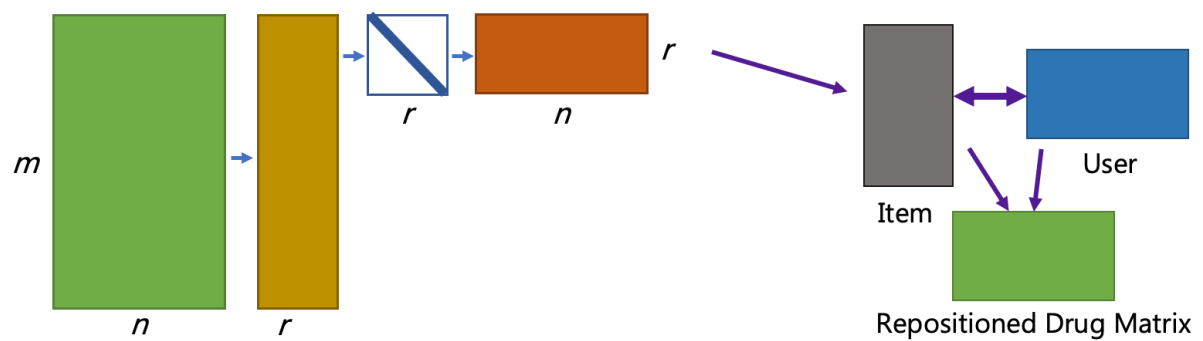


Figure 14: Decomposition of matrices via SVD, then solving for user and item matrices using WALS

After WALS, the RMSE (Root Mean Square Error) measure was calculated to determine the difference between the available drug matrix and the repositioned drug matrix.

```

class MF_KNNDrugRepo(nn.Block):
    def SVDMatrix(drug, target, label, **kwargs):
        super(MF, drug).__init__(**kwargs)
        self.P = nn.Embedding(input_dim=num_users, output_dim=num_factors) #P is u
        self.Q = nn.Embedding(input_dim=num_items, output_dim=num_factors) #Q is i
        self.user_bias = nn.Embedding(target, 1)
        self.item_bias = nn.Embedding(target, 1)

    def WALSDrugRepo(drug, user_id, item_id):
        P_u = drug.P(user_id)
        Q_i = self.Q(item_id)
        b_u = self.user_bias(user_id)
        b_i = self.item_bias(item_id)
        outputs = (P_u * Q_i).sum(axis=1) + np.squeeze(b_u) + np.squeeze(b_i)
        return outputs.flatten()

```

Figure 15: Code segment showing methods used to filter drug attribute matrices via SVD and WALIS

Finally, rank selection was employed. Data from the repositioned drug matrix was ranked to eliminate noise and include important features. The ranking of data allows for compression of data while preserving useful associations.

```

for i in range(4):
    R12_found = np.load('./tmp/R12_found_' + str(i+1) + '.npy')

    R12_2 = []
    R12_found_2 = []
    for i in range(n):
        for j in range(m):
            if M10[i, j] == 0:
                R12_2.append(R12[i, j])
                R12_found_2.append(R12_found[i, j])

    precision, recall, _ = metrics.precision_recall_curve(R12_2, R12_found_2)
    aps = metrics.average_precision_score(R12_2, R12_found_2)

    plt.plot(recall, precision, label="Model " + str(i+1) + ", APS= %0.2f" % aps)

```

Figure 16: Code segment showing ranking of data

With the compressed data, drugs were scored via the MF method with links containing the drug name, ID, label, and the score. The score was computed as the coefficient in the new matrix by drug targets to the molecular targets of epilepsy. scores were then ranked to reveal the most promising drug candidates.

```
index_new_links_R12 = np.argsort(new_links_R12.flatten())

new_drugs = []
new_labels = []
prob = []
new_drugnames = []
for i in range(1, 21):
    new_drugs.append(drugs[index_new_links_R12[-i] % m1])
    new_labels.append(labels[index_new_links_R12[-i] // m1])
    prob.append(new_links_R12[index_new_links_R12[-i] // m1, index_new_links_R12[-i]])
    new_drugnames.append(drugnames[index_new_links_R12[-i] % m1])
```

Figure 17: Code segment printing out the first 21 drugs with the highest scores computed by MF-method

Rank	Drug
1	metformin
2	nifedipine
3	pyrantel tartrate
4	quercetin
5	trifluoperazine

Figure 18: Table showing possible candidates for epilepsy

Metformin is a medication used to treat type 2 diabetes by controlling high blood sugar. Specifically, metformin restores the body's response to insulin in addition to decreasing the amount of sugar the liver makes and the stomach absorbs. Metformin also inhibits mitochondrial complex I activity, preventing production of ATP. These changes activate AMP-activated protein kinase (AMPK), an enzyme that plays an important role in the regulation of glucose metabolism (Drug Bank). However, the action mechanism of metformin is not fully known, so the reason as to why metformin can be an anti-epileptic drug is uncertain.

Nifedipine is a drug for management of hypertension, in addition to several subtypes of angina pectoris. A calcium channel blocker, nifedipine reduces blood pressure and increases the flow of oxygen to the heart. It was inferred that nifedipine was selected as an AED because it regulates movement of ions across respective channels. Similarly functioning anti-epileptic drugs, such as eslicarbazepine acetate, operate by regulating voltage-gated sodium channels (Drug Bank).

Pyrantel tartrate is commonly used for the removal of enterobiasis infections such as roundworm, pinworm, and hookworm in humans, cats, and dogs, among other animals. Pyrantel tartrate increases acetylcholine release and decreases cholinesterase levels in these parasites, leading to paralysis and eventual release from the walls of the host organism. It is unknown as to why pyrantel tartrate was selected as a potential anti-epileptic drug candidate as virtually no medical records exist connecting these two components (Drug Bank).

Quercetin is a natural flavonol used in food and natural supplement products. Often found in plants, quercetin is an antioxidant that is a specific quinone reductase 2 inhibitor. This catalyzes the metabolism of quinolines, toxic organic compounds. Quercetin may play a role in preventing seizures through combination with other neurotransmitter-targeted treatments.

Trifluoperazine is a phenothiazine used for treatment of anxiety and depression. Trifluoperazine's mechanism of action includes the blocking of mesolimbic dopaminergic D1 and D2 receptors, located in the brain, in addition to decreasing levels of the hypothalamic and hypophyseal hormones. There is no existing literature to connect trifluoperazine to epilepsy treatment other than a clinical approach to drug repositioning for epilepsy. Therefore, the reasoning behind trifluoperazine being selected as a potential AED is uncertain.

A threshold δ was selected to minimize false positives and maximize true positives. Testing data was run through the algorithm to determine this threshold.

```
delta = [0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8]
n, m = R12_found.shape
R12_delta = np.zeros((n,m))
T = [np.sum(R12)]*len(delta)
TP, FP = [], []

plt.rcParamsDefaults()
fig, ax = plt.subplots()
for l in range(len(delta)):
    for i in range(n):
        for j in range(m):
            R12_delta[i,j] = R12_found[i,j] > delta[l]

    R12_TP = np.multiply(R12_delta, R12)
    TP.append(np.sum(R12_TP))
    FP.append(np.sum(R12_delta) - np.sum(R12_TP) + T[0])

y_pos = np.arange(len(delta))

ax.barh(y_pos, FP, label = 'new links (FP)', color = 'gold')
ax.barh(y_pos, T, label = 'False Negative (FN)', color = 'red')
ax.barh(y_pos, TP, label = 'True positive (TP)', color = 'green')
```

Figure 19: Code segment analyzing effects of different thresholds ("delta")

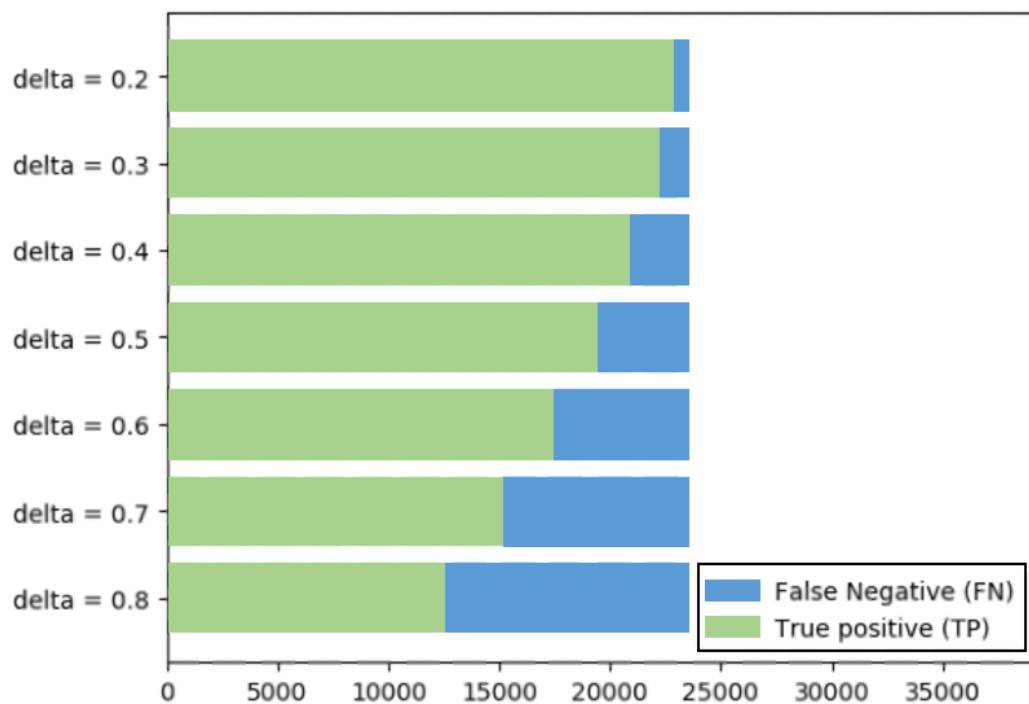


Figure 20: Graphical analysis of true positive and false negative rates with different threshold values (delta)

Each attribute of the drug was also compared separately to evaluate its effect on drug prediction.

```
n_fold = 10 # number of folds in cross-validation
features = ["structure", "target", "label"]

for feature in features:
    features_modified = [ feature ]
    print feature
    # Check prediction accuracy of ML classifier on the data set using the parameter
    ml.check_ml(data, n_run, knn, n_fold, n_proportion, n_subset, model_type, pres
```

Figure 21: Code segment analyzing each attribute (structure, target, label) separately

Algorithm Testing via PREDICT Dataset:

The well established PREDICT dataset for drug repositioning was used to assess the effectiveness of NeuroCADR. PREDICT, a method for predicting drug indications, contains a benchmark dataset with 1834 interactions between 526 FDA approved drugs and 314 diseases

PREDICT ranks additional drug-disease associations by comparing it to a gold-standard set of known drug-disease associations.

Results and Discussion:

NeuroCADR found more than 150 potential drug candidates based on the datasets inputted. Accuracy of the drug candidates was measured by structural similarity to drugs that were currently being used to treat epilepsy in addition to cross and external validation metrics. One of these metrics is the Area under the receiver operating characteristic curve (ROC), abbreviated AUC. This graph shows the performance of any classification model at all thresholds by comparing true positive and false positive rates.

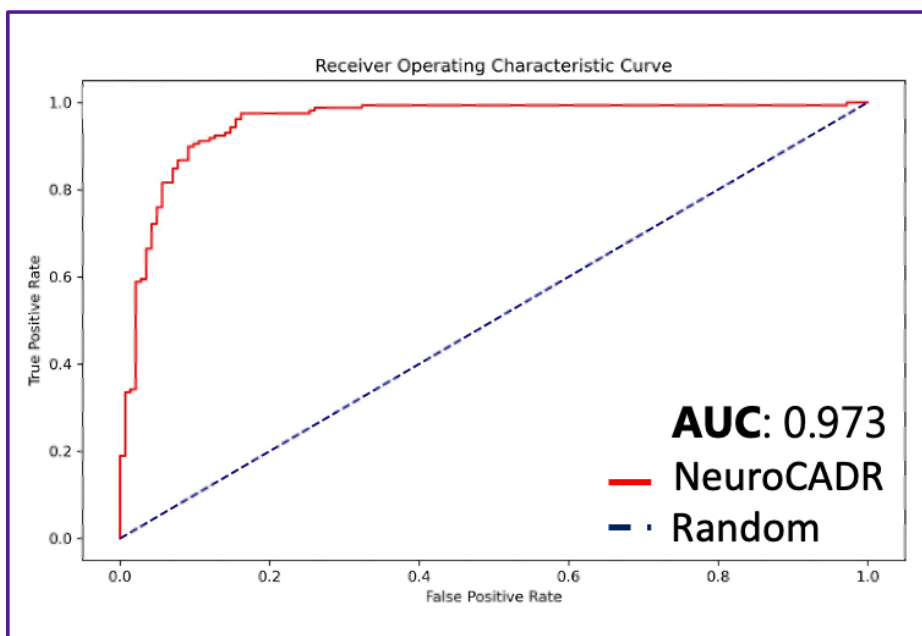


Figure 22: AUC Curve

The performance of NeuroCADR was compared to several other existing computational drug repositioning methods such as data/text mining, deep learning, network analysis, and logistic regression. NeuroCADR surpassed all existing approaches by the metrics mentioned

above. The proposed algorithm matched closest in performance to deep learning which was expected as KNN is a subset of deep learning and neural networks and therefore follows the same general principles.

```
"""
if model_type == "svm":
    clf = svm.SVC(kernel='linear', probability=True, C=1)
elif model_type == "logistic":
    clf = linear_model.LogisticRegression(penalty='l2', dual=False, tol=0.0001)
elif model_type == "knn":
    clf = neighbors.KNeighborsClassifier(n_neighbors=5) #weights='uniform', al
elif model_type == "tree":
    clf = tree.DecisionTreeClassifier(criterion='gini', random_state=n_seed) #
elif model_type == "rf":
    clf = ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', r
elif model_type == "gbc":
    clf = ensemble.GradientBoostingClassifier(n_estimators=100, loss='deviance
elif model_type == "custom":
    if fun is None:
        raise ValueError("Custom model requires fun argument to be defined!")
    clf = fun
else:
    raise ValueError("Unknown model type: %s!" % model_type)
return clf
```

Figure 23: Code segment showing data was analyzed with multiple model types based on the value of “model_type”

Method	AUC Value
NeuroCADR	0.973
Logistic Regression	0.880
Text Mining	0.785
Support Vector Machine	0.882
Deep Learning	0.948
Biological Network Analysis	0.830

Figure 24: Comparison of AUC between computational drug repositioning methods

The success of NeuroCADR was compared to that of a clinical approach to epilepsy, where hippocampal brain tissue of patients with epilepsy was analyzed with RNA sequencing.

NeuroCADR was able to identify a greater number of potential drug candidates. However, the above study tested the effectiveness of the most promising drug candidates on zebrafish and concluded positive results.

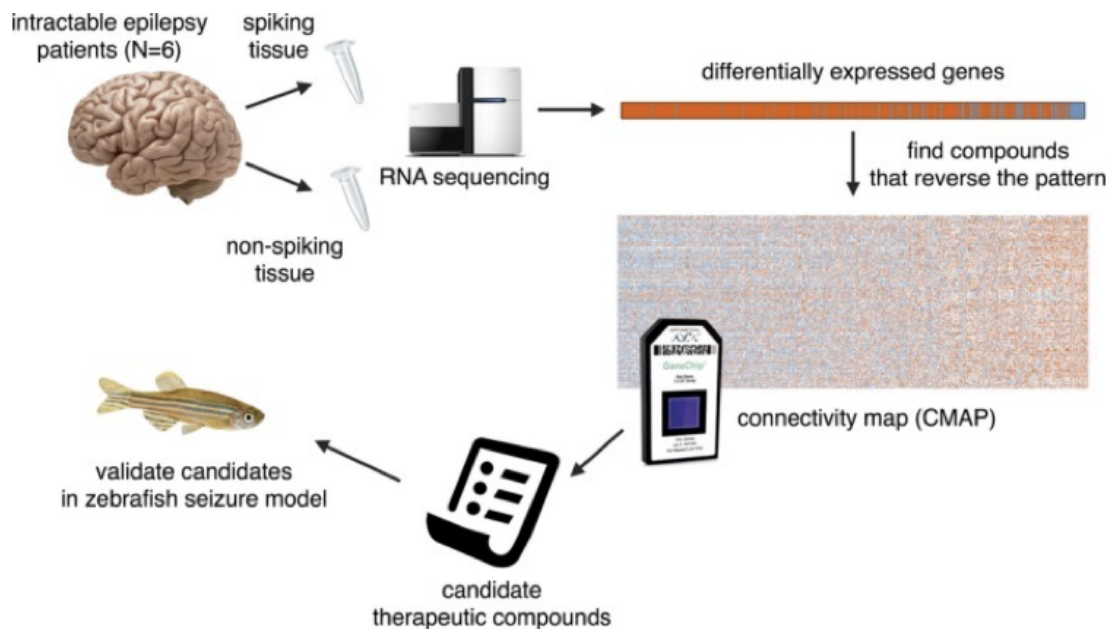


Figure 25: Study layout of clinical approach to epilepsy.

Source: Brueggeman, Leo et al. "Drug repositioning in epilepsy reveals novel antiseizure candidates." *Annals of clinical and translational neurology* vol. 6,2 295-309. 11 Dec. 2018, doi:10.1002/acn3.703

Potential Limitations:

One possible limitation of this algorithm is the databases used as they include all approved drugs only. This error can be mitigated by further training of the algorithm using drugs that are in later stages of clinical trials.

Another possible limitation of NeuroCADR is overfitting or underfitting of the data. Overfitting the data would have caused "false positives", drug candidates that are realistically not

suitable for treatment for epilepsy, while underfitting the data would have caused certain drugs that may be practical for treatment to not be recorded by the algorithm. The effects of overfitting and underfitting can be visualized by changing the threshold, δ , to either a higher number to eliminate false positives or to a lower number to include false negatives. However, the high AUC value suggests very few instances of this.

The value of k was determined experimentally. K represents the number of nearby drugs the algorithm checks to determine if they are viable candidates. Using different datasets may potentially return a different optimal value of k , therefore changing the value of the AUC.

NeuroCADR Website:

The machine learning algorithm was incorporated into a user-friendly website that can be used by doctors and other medical professionals to reveal potential drug candidates to prescribe patients based on their prior medical history.

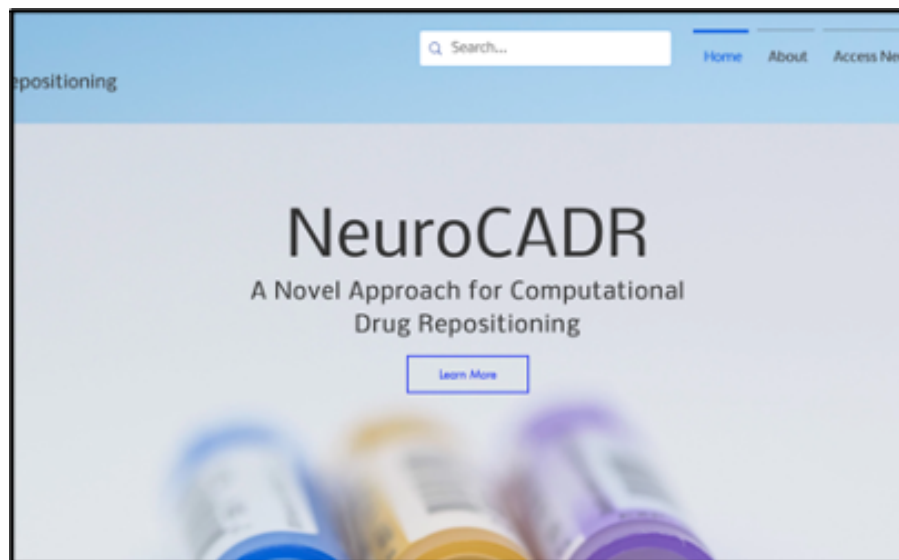



Figure 26: The NeuroCADR homepage, with menu options on the top bar.

The website has sections that describe the algorithm itself and the process that was used to create it, from data sources and compilation to algorithm construction and training. This was done to maintain maximal transparency and to provide a known baseline for additional data to be inputted in subsequent updates. Additionally, a section about drug repositioning and its advantages over the traditional method of drug repurposing was incorporated into the website in order to educate the general public about the need for NeuroCADR.

ABOUT

NeuroCADR presents a novel approach for drug repositioning utilizing biological networks and k-nearest neighbor algorithms (KNN). NeuroCADR can identify novel drug candidates for epilepsy that can be further approved through clinical trials.



Why NeuroCADR?

Drug repositioning is an emerging alternative to the traditional drug development process and involves the reassignment of existing drugs for novel purposes.

Repositioned drugs are cheaper, faster, and less failure prone than traditional drugs as they have already passed clinical trials.

Drug repositioning has recently been performed via an in silico approach - databases of drugs and chemical information are used to identify potential drug candidates

It was hypothesized that utilizing KNN (k-nearest neighbor) algorithms and spherical k-means to perform drug repositioning would result in a greater number of potential repurposed drug candidates in addition to more accurate drug predictions because the combination of supervised and unsupervised ML for classification and clustering will allow for more functionality.




Figure 27: Webpage explaining the evolution of NeuroCADR and drug repositioning.

An instruction manual is also included for anyone who wishes to use the software to run the algorithm for different diseases. Instructions include how to set up the software for optimal running, how to input a patient's prior medical history so NeuroCADR can check if potential drug candidates will affect drugs the patient is taking, and how to interpret the results outputted by the algorithm.

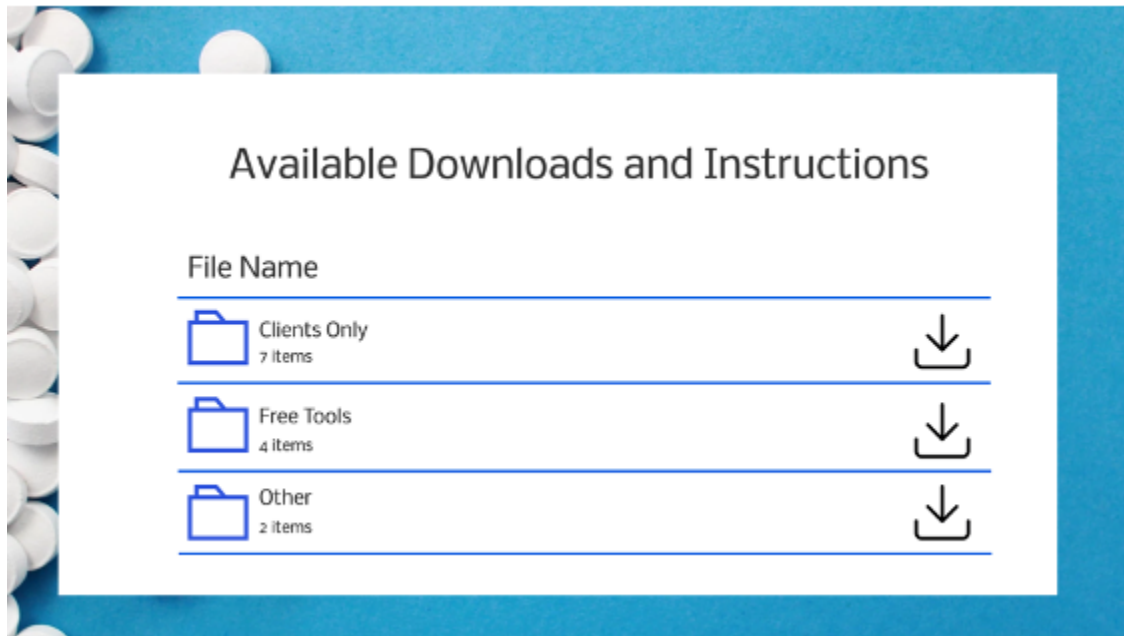


Figure 28: Instruction manuals containing files for users to download to use NeuroCADR.

A blog was established to catalog any updates made to NeuroCADR, such as further training of the algorithm or the addition of new data.

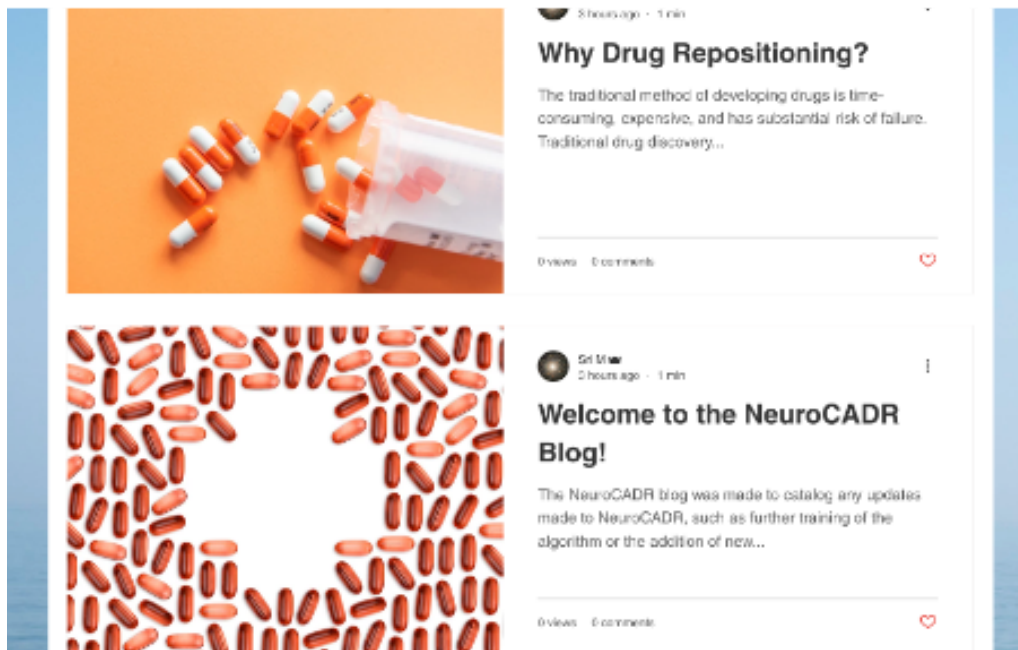


Figure 29: Blog posts on the website, allowing users to contribute questions and comments about drug repositioning and NeuroCADR.

Impact and Applications:

Development of Novel Pharmaceutical Treatments:

NeuroCADR can be used by pharmaceutical companies to develop novel therapeutic treatments for patients with conditions that have very few drug treatments. Pharmaceutical companies can save billions of dollars per drug, in addition to being able to send drugs to the market in at least half the time of a traditionally developed drug. People with conditions that are being treated with repurposed drugs can have the opportunity to get affordable treatment in which the effects are already clearly known.

Establishment of Drug Repurposing Candidates for Other Diseases:

NeuroCADR can be run to reveal novel drug candidates for other neurological diseases such as Alzheimer's disease and Parkinson's disease, two of the most severe neurodegenerative disorders that currently have no cure. With additional data, NeuroCADR can be expanded to return drug candidates for non-neurological conditions.

Orphan diseases, diseases that affect less than 200,000 people nationwide, would be greatly benefited by drug repositioning. Many orphan diseases currently do not have drugs developed for them because of the high financial cost needed to develop drugs via the traditional method, providing little financial incentive for pharmaceutical companies to develop them. Drug repositioning can provide novel treatments for patients with these diseases due to the reduced cost involved.

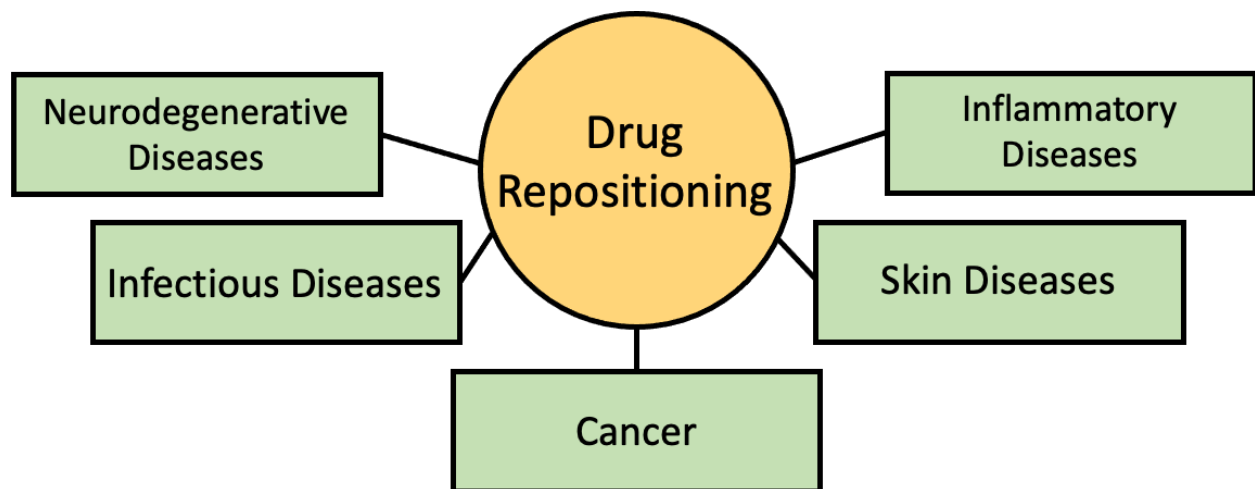
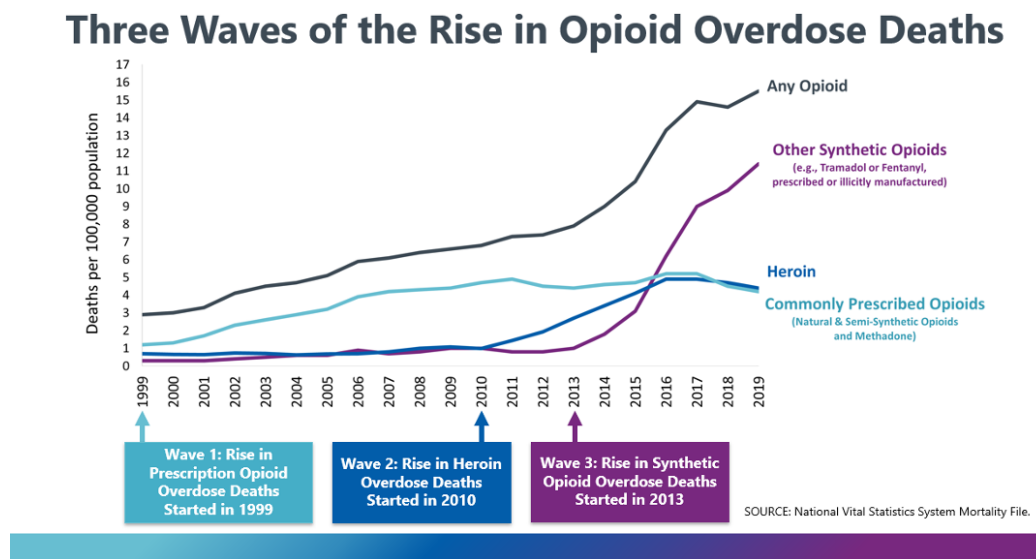


Figure 30: Disease groups that have been treated with drugs through drug repositioning

Combating the Opioid Epidemic:

Opioids are a class of drugs that include legally prescribed drugs such as oxycodone as well as illegal drugs such as heroin and fentanyl. Opioids are mostly prescribed for pain relief

and provide morphine-like effects, causing users to quickly become addicted to them. Every year, 10.1 million people misused prescription opioids. Of those people, over 70,000 died from a drug overdose. Opioid overdose deaths have quadrupled since 1999 and are only increasing each year. NeuroCADR can be used to identify alternatives to opioid painkillers that can be prescribed to prevent drug addiction and drug overdoses, saving lives.



Opioid overdose trends from 1999-2019.

Source: CDC, "Understanding the Opioid Overdose Epidemic"

Next Steps:

Expanding NeuroCADR to include different datasets of drugs, genes, and protein interactions to enable identification of potential drugs for diseases other than epilepsy. Further model training and data classification for NeuroCADR will improve accuracy of drug candidates that are returned by the algorithm in addition to revealing potential drug combinations that could be used to treat epilepsy and other diseases. Testing the NeuroCADR website with real life patient medical history data will allow NeuroCADR to make more accurate and safe predictions.

Publishing NeuroCADR for pharmaceutical companies can pave the way for future repurposed drug treatments.

Conclusion:

This project aimed to develop a novel computational approach for drug repositioning using a KNN algorithm and spherical k-means to reveal potential drug candidates for epilepsy. The hypothesis that this algorithm would be more accurate than existing in silico methods was supported. NeuroCADR reported a greater number of drug candidates for epilepsy than other methods such as logistic regression and support vector machines. NeuroCADR also performed better than clinical approaches to drug repositioning by reporting more drug candidates with greater accuracy.

The algorithm formed analyzed drugs with individual datasets containing associations between drug structures, proteins, and genes. The algorithm was initialized via spherical k-means to establish clusters, which were then organized into matrices and analyzed using KNN and matrix factorization methods. Drugs were then scored and ranked against a threshold to reveal the most plausible drug candidates.

The resulting algorithm was incorporated into a website to aid medical professionals in prescribing drugs to patients based on the patient's existing medical needs. NeuroCADR also contains information about drug repositioning to educate the general public, in addition to an instruction manual for anyone who wishes to use the software. A blog is included to contain any updates made and to spread awareness about epilepsy and other neurological diseases.

There are many applications of this project. NeuroCADR can help in the development of new pharmaceutical treatments for epilepsy by providing companies with information on the

most plausible drugs to repurpose. NeuroCADR can also provide insight for treatments of other diseases that are often overlooked by pharmaceutical companies due to the high cost involved. The opioid epidemic is also an issue that NeuroCADR can assist with by providing less addicting alternatives to opioids.

References:

This project was performed with the guidance of Professor Jaudelice de Oliveira of Drexel University and PhD student Dubem Ezech of Drexel University.

1. Brueggeman, L., Sturgeon, M. L., Martin, R. M., Grossbach, A. J., Nagahama, Y., Zhang, A., Howard, M. A., 3rd, Kawasaki, H., Wu, S., Cornell, R. A., Michaelson, J. J., & Bassuk, A. G. (2018). Drug repositioning in epilepsy reveals novel antiseizure candidates. *Annals of clinical and translational neurology*, 6(2), 295–309. <https://doi.org/10.1002/acn3.703>
2. Fahimian, G., Zahiri, J., Arab, S.S. et al. RepCOOL: computational drug repositioning via integrating heterogeneous biological networks. *J Transl Med* 18, 375 (2020). <https://doi.org/10.1186/s12967-020-02541-3>
3. Jarada, T. N., Rokne, J. G., & Alhajj, R. (2020). A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. *Journal of cheminformatics*, 12(1), 46. <https://doi.org/10.1186/s13321-020-00450-7>
4. Kobylarek, D., Iwanowski, P., Lewandowska, Z., Limphaibool, N., Szafranek, S., Labrzycka, A., & Kozubski, W. (2019). Advances in the Potential Biomarkers of Epilepsy. *Frontiers in neurology*, 10, 685. <https://doi.org/10.3389/fneur.2019.00685>
5. Park K. (2019). A review of computational drug repurposing. *Translational and clinical pharmacology*, 27(2), 59–63. <https://doi.org/10.12793/tcp.2019.27.2.59>
6. Rudrapal, M., Khairnar, S. J. , & Jadhav, A. G. (2020). Drug Repurposing (DR): An Emerging Approach in Drug Discovery. In (Ed.), *Drug Repurposing - Hypothesis, Molecular Aspects and Therapeutic Applications*. IntechOpen. <https://doi.org/10.5772/intechopen.93193>

7. Xiangxiang Zeng, Siyi Zhu, Xiangrong Liu, Yadi Zhou, Ruth Nussinov, Feixiong Cheng, deepDR: a network-based deep learning approach to in silico drug repositioning, *Bioinformatics*, Volume 35, Issue 24, 15 December 2019, Pages 5191–5198, <https://doi.org/10.1093/bioinformatics/btz418>