Learning Molecule Drug Function from Structure Representations with Deep Neural Networks or Random Forests

Jesse G. Meyer[1,2,3]*, Shengchao Liu[4,5], Ian J. Miller[3], Anthony Gitter[4,5,6], Joshua J. Coon[1,2,3,5,7]

[1] Department of Chemistry
[2] Department of Biomolecular Chemistry
[3] National Center for Quantitative Biology of Complex Systems
[4] Department of Computer Sciences
[5] Morgridge Institute for Research
[6] Department of Biostatistics and Medical Informatics
[7] DOE Great Lakes Bioenergy Research Center

University of Wisconsin—Madison, Madison, Wisconsin 53706, United States

*Correspondence to: jessegmeyer@gmail.com

Abstract
Empirical testing of chemicals for desirable properties, such as drug efficacy, toxicity, solubility, and thermal conductivity, costs many billions of dollars every year. Further, the choice of molecules for testing relies on expert knowledge and intuition in the relevant domain of application. The ability to predict the action of a molecule *in silico* would greatly increase the speed and decrease the cost of prioritizing molecules with desirable function for experimental testing. We explore the use of molecule structures to predict their drug classification without any explicit biological model (e.g. protein structure or cell system). Two approaches were compared: (1) a two-dimensional image representation of molecule structure with transfer learning from a pre-trained convolutional neural network (CNN), and (2) Morgan molecular fingerprints (MFP) with a random forest (RF). Both methods using only molecule information as input achieved significantly better accuracy than previous work that used transcriptomic measurements of molecular effects. Because these data suggest that much of a molecule's function is encoded by its chemical structure, we explore which general molecular properties are drug class-specific and how misclassification of structures might provide drug repurposing opportunities.

Introduction

Development of molecules with new or improved properties is needed in many industries, including energy, agriculture, and medicine. However, the number of possible molecules to explore, also referred to as chemical space, is exceedingly large[1,2]. Even when chemical space is limited to 'Lipinski's Rule of Five'[3], which applies to the subtask of drug development, there are still as many as 10^{60} possible chemical structures[4]. Regardless of the available chemical diversity, the pace of molecule discovery for drug development and other applications has steadily decreased.

A promising approach for discovering new molecules with desirable properties from the massive chemical space is machine learning. Machine learning using deep neural networks (DNNs), also known as deep learning, is a powerful and flexible approach that has created new opportunities in multiple fields, including chemistry[5–7]. A weakness of neural networks is the size of training data required. The greatest successes of deep learning have trained with tens of thousands to millions of examples. However, methods are constantly evolving and enabling application of deep learning to problems with less training data available. One such method is transfer learning, where neural networks trained on a large dataset for one problem are applied to another similar problem[8–10].

Recently there has been an increase in the amount of research on using machine learning for various tasks in chemistry[6,7]. Many papers describe methods for embedding molecules into a latent space and engineering molecules with desirable properties[11,12]. Reinforcement learning has been applied with paired DNNs to design molecules with desired properties, such as solubility or transcription factor inhibition[13]. A framework for benchmarking model predictions is available[14]. Recent work has even explored the concept of using 2D molecule images for predicting toxicity with DNNs[15] and molecule property prediction[16,17]. One study used a generative adversarial neural network architecture to generate molecules that induce specific transcriptomic states[18]. Most of these examples use SMILES strings[19–21] as initial molecular inputs for embedding, but there is a trend toward use of molecular graphs[22–24].

In this paper, we use machine learning and structure-based molecule representations to predict specific medical subheading (MeSH) 'Therapeutic Uses' classes[25]. Such a classifier would be useful in assigning potential utility to arbitrary chemicals. The same classification task was described previously by Aliper, *et al.*[26], who used a molecule's influence on cellular transcriptome measured by the LINCS project[27] as a proxy molecule representation. We compared two strategies, 2D images with convolutional neural networks (CNNs) and Morgan extended-connectivity fingerprints[28] with random forests (RF)[29]. For this drug classification task, our RF outperforms our CNN, and both our RF and CNN greatly outperform the previous gene expression-based models.

Methods

All code for the CNN and RF models and the exact datasets are available from https://github.com/jgmeyerucsd/drug-class.

Data

Molecules were selected from PubChem on October 2nd, 2018[30] according to their MeSH 'Therapeutic Uses' classification (Chemicals and Drugs Category > Chemical Actions and Uses > Pharmacologic Actions > Therapeutic Uses). All molecules from 12 classes were downloaded in a spreadsheet containing their compound identification number (CID). A total of 11,929 CIDs were converted to SMILES strings using the python package *pubchempy*. SMILES strings with length over 400, or membership to more than one of the twelve MeSH therapeutic classes were excluded, leaving 8372 SMILES. This final list was filtered to remove multiple versions of molecules that differ by only accompanying salts. The final filtered total was 6955 molecules. The distribution of molecules among classes is given in **Table 1.** SMILES strings were converted to three-color (RGB) images with size 500 x 500 pixels or 1024-bit Morgan fingerprints using the python package *RDKit*. Three-channel RGB images were also converted to two-channel black and white images for comparison. Molecule classes were then split into three subgroups for model training and prediction: 3-, 5-, or 12-class prediction tasks according to the groupings described previously by Aliper *et al.* (**Table 1**).

Images with Convolutional Neural Networks (IMG+CNN)

Molecule images with RGB channels were resized to 224x224 pixels and used for training and validation of a CNN with predetermined weights from resnext101_64[31] implemented using fastai and pytorch[32]. The loss function used was binary cross entropy and the output layer was logsoftmax. During training, all the layers were frozen except the output layer, which was retrained for various numbers of epochs with a learning rate of 0.01, or with cyclic learning rate decreasing from 0.01 for various numbers of epochs[33]. The CNN was trained without data augmentation unless otherwise noted.

Molecular Fingerprints with Random Forests (MFP+RF)

Random forests[29] are ensembles of decision trees, where each tree is learned on a subsample of data points and features (in this case bits in a molecular fingerprint). Benchmarking studies often include MFP+RF models because of their strong performance on a variety of computational chemistry tasks[14,34–39] The random forest model is implemented with scikit-learn[40]. The RF was configured to use 4000 trees, minimum of 1 sample per leaf, balanced class weight, and log2 max features,

hyperparameters that performed well in a previous high-throughput chemical screening study[34].

CNN Latent Representations

Latent image representations were used to compare the MFP and IMG representations of molecules. To generate latent representations of molecule images, the resnext101_64 CNN model was loaded and not trained on any molecule images. The final three layers, a batch normalization layer, linear layer, and logsoftmax, were removed. The same molecule images used for training and classification above were then passed through the truncated network to generate a 512-length vector of their latent representation in the final hidden layer.

Comparison of Drug-like Properties

CIDs were used to download Molecular weight, XLogP, HBondAcceptorCount, HBondDonorCount, and IsomericSMILES values using the pubchempy module. Compounds were then filtered to include non-redundant IsomericSMILES values to include only drugs with a single class label (see Supplemental Data File 1). XLogP values were not available for all queried compounds (see Supplemental Table 1). Violin plots were created using ggplot2 (see **Figure 4**). For each quantitative feature, a Welch's ANOVA and Games-Howell post hoc test (implemented in R's userfriendlyscience package) was used to compare differences between chemical features between drug-class groups. This test was selected because it does not assume a normal distribution, even variance, or equal sample sizes between groups[41]. Adjusted p-values from Games-Howell post hoc test are reported in Supplemental table 2). Drug-class level distribution and pairwise relations of chemical features were further visualized with Seaborn (see **Supplemental Figure 2**).

Model Comparison and Evaluation

The primary IMG+CNN and MFP+RF model comparison used default hyper parameters that have worked well previously[33]. Subsequent analyses assessed the sensitivity of the IMG+CNN model to the hyperparameters and training data augmentation. For model validation, five validation sets were constructed by stratified random sampling of 20% of the molecules in each class. The classification accuracy on the five validation sets was used as the primary model comparison metric. Principal component analysis (PCA) was used to compare the Morgan fingerprints and latent molecule representations from the CNN, and the results were visualized using plots of the first two principal components. Class-specific prediction accuracy of the models was compared using confusion matrices. We also compared our prediction accuracy with the accuracy previously reported by Aliper *et al.*[26]. However, it should be clearly noted that

we used different training and validation sets in this paper so the accuracies are not directly comparable.

Results

General Overview

Two structure-based representations were used for training and classification with two different model architectures: (1) IMG+CNN or (2) MFP+RF (**Figure 1**). Molecules were split into three subtask sets as described previously[26]. Each subtask set contained 3105, 5762, or 6955 molecules for the 3-, 5- and 12-class problems, respectively. **Table 2** gives a summary of the validation set accuracy for the models described here in comparison with results from Aliper *et al.* (note that our training and validation sets were different than that used by Aliper *et al.*) For all subtasks, MFP+RF performed best, achieving up to 88.6 % accuracy, even on the 12-class prediction task. The IMG+CNN models with default settings were worse, but still significantly outperformed previous results using a multilayer perceptron DNN or support vector machine (SVM) and gene expression inputs.

Additional training strategies were assessed for their impact on the CNN hyperparameter sensitivity on the 3-class subtask (**Table 3**). Cyclic learning rates were also tested, and improved neural network training[26]. Batch size was varied from 50-800 pictures per batch using either constant learning rate or cyclic learning rates. Validation set accuracy slightly depends on the batch size and number of epochs using both static learning rates and cyclic learning rates, with batch size of 400 or more generally performing worse than batches of 200 pictures or less. When training over 255 epochs, cyclic learning had no significant effect on model performance. However, with 1023 training epochs, a cyclic learning rate was better than static learning rate (two-sided t-test p-value = 0.035). The best results were achieved using cyclic learning rates with 10 cycles with a batch size of 200. Most interestingly, when the CNN was trained with grayscale images instead of color images, validation accuracy was significantly lower than when using color molecule images (**bottom row, Table 3,** Color: 0.817 ± 0.005, Greyscale: 0.801 ± 0.005, 2-tailed t-test p-value = 0.001).

A common strategy for improving the performance of CNNs is data augmentation by randomly altering training images[42]. We also tried training with pictures augmented by randomly flipping horizontally or vertically, randomly rotating up to 10 degrees, and randomly zooming by up to 10%. Despite the input data's relative simplicity, data augmentation was effective at improving validation set accuracy by 3.7-6.1%. On a single cross-validation fold (validation set #1), IMG+CNN plus data augmentation achieved 0.866, 0.8226, and 0.779 validation accuracy on the 3-class, 5-class, and 12-class problems, respectively (2047 epochs, initial learning rate 0.01 with cosine annealing).

Model and Representation Comparisons

To better understand the differences between model inputs, PCA was used to compare the multidimensional distributions of latent images and MFPs. This analysis revealed that MFP representations capture much greater variance than CNN-processed images. The first 10 principle components (PCs) from PCA of MFPs explain only 20.4% of the variance in the data (**Figure 2A**), but the first 10 PCs of the explained 54.0% of the variance in the CNN-transformed data (**Figure 2B**). This suggests that MFPs allow greater embedding of structural uniqueness than the transformed images.

Given the unequal stratification among examples within classes in the training and validation sets, the per-class performance of both models was compared. Confusion matrices of true class versus predicted class from IMG+CNN for the 3 subtasks reveal differences in per-class validation accuracy. In the 3-class prediction subtask, the prediction performance has similar high accuracy among the three groups, but 17% of cardiovascular drugs are predicted incorrectly as central nervous system drugs (**Figure 3A**). In the 5-class prediction subtask, which includes the 3-class drugs plus gastrointestinal and anti-infective drugs, prediction performance is slightly decreased relative to the 3-class performance. However, the model learns anti-infective drugs with high accuracy (89%). Gastrointestinal drugs are predicted poorly with only 57% accuracy (**Figure 3B**). In the 12-class prediction task, the accuracy for anti-infective molecules remains the highest at 89%, but smaller classes are predicted less accurately (**Figure 3C**). The smallest class 'Urological Agent', which contains only 26 molecules, was rarely predicted correctly in the validation set. The difference in class sizes likely contributes to this deficiency, and we did not directly control for this during training. The same analysis of per-class validation set accuracy for results from MFP+RF revealed more consistent accuracies across the classes (**Supplemental Figure 1**).

Understanding Molecule Properties and Misclassification

A compound's drug-like properties are conventionally believed to be a function of chemical features such as molecular weight, lipophilicity, and number of hydrogen bond donors and acceptors. For instance, Lipinski's rule of five famously states that a drug-like compounds should have no more than five hydrogen bonds donors, 10 hydrogen bond acceptors, a logP greater than 5, and molecular weight under 500[3]. These properties are directly or indirectly encoded in the graphical representation of a compound's chemical structure. In the compound dataset explored in this study, anti-infective and respiratory system drugs exhibit different distributions in this chemical feature space (**Figure 4**). Based on a Welch's ANOVA and Games-Howell post hoc test, these two classes of drugs appear different in their distributions of molecular weight, hydrogen bond donors, hydrogen bond acceptors, but not XLogP (Supplemental Table 2). However, it is important to consider that XLogP values for this dataset

(collected via the pubchempy API) are sparse (Supplemental Table 1), are computed[43] rather than measured, and appear to have a different distribution than the other chemical features (**Supplemental Figure 2**), so comparisons of XLogP values in this dataset may be skewed. Assuming that there is in fact no meaningful difference between the XLogP values for anti-infective and respiratory drugs, perhaps the source of the classifier's errors (**Figure 3 and Supplemental Figure 1**) is it's overweighting of features related to lipophilicity relative to molecular weight or hydrogen bond donors or acceptors. Conversely, while gastrointestinal drugs and antineoplastic drugs do not show statistically significant differences across this drug-like chemical feature space, the classifier is able to easily resolve them. This phenomenon perhaps suggests that there are important structural features of a drug that the classifier is able to learn that fall beyond the conventional framework of how chemists understand the drug-like properties of compounds.

Misclassification for Mechanism or Drug Repurposing Opportunities

Misclassification can be interpreted in at least two ways: (1) the model hasn't learned enough to accurately predict the true class, or (2) the model has learned something beyond our understanding about the drugs and classes. Although the latter is more interesting, the former is the safer and more likely interpretation. However, cases where the model is wrong might present opportunities for drug repurposing. In addition, in this work we hypothesize those incorrect predictions might inform us about drug mechanisms. For example, among the 6 molecules in the urological drug validation set, the IMG+CNN misclassified Trospium as a central nervous system (CNS) agent. This is not surprising, however, because Trospium is known mechanistically as a muscarinic antagonist[44], which is a common function of CNS drugs. In fact, the structure of Trospium is remarkably similar to another muscarinic antagonist used to treat Parkinson's disease, Benztropine[45] (**Figure 5**).

Conclusion

Here we report two drug classification models that use structure-based representations. Both models greatly exceed a previous benchmark on the same prediction task. Our models use molecular structures directly as inputs. In contrast, the previous work used alterations of the transcriptome as a proxy for the molecule. The results presented in this paper suggest that experimental measurement of a molecule's influence in biological systems may not be needed to accurately predict some types of chemical properties, such as drug classes. Further research is required to determine which types of chemical prediction tasks require information about the biological state changes included by chemicals instead of the chemical properties alone. The richness and stochasticity of biological systems cannot possibly be captured without some direct empirical measure of the system. We believe there must be room to improve effect-

based models that would outperform models based solely on molecule structure, especially on more complex prediction tasks.

A related point is that the images and MFPs can be generated directly from the chemical structure for almost *any* chemical. A major limitation of the Aliper *et al.* featurization, or any empirically-determined featurization, is that it requires new experiments for each new compound. Therefore, there must be a major return for that extra cost in a virtual screening or chemical prediction setting. Therefore, we propose that future studies on this task or any chemical prediction task that use empirically-determined featurizations must also use models that consider only chemical structure features as a baseline.

Although there are several chemistry problems where DNNs outperform other machine learning methods[38,46,47], here the MFP+RF performed best in all three classification tasks. However, from our results we cannot determine if this is due to the featurization, the type of model, or the model configuration. For example, images may be worse featurizations than MFPs, but CNNs may be harder to train than the RF, especially when the amount of training data is limited. Our results suggest that MFPs are a more dispersed representation than images, suggesting that use of MFPs may help the accuracy of our RF model (**Figure 2**). A widespread hyperparameter search could be used to assess whether CNNs can outperform RFs when both models are directly optimized for this task. Even though our CNN method was worse than MFP+RF for these hyperparameter settings, our results show that transfer learning and molecule images for CNNs provide a simple and effective method for predicting drug function.

Much can be learned about chemical function from the cases where we find misclassification of chemical structures. We show cases where this can be rationalized by chemical properties of the molecules and cases where these properties that we often use to define the character of a chemical cannot explain the classification performance. In the latter case, this may mean that our models have learned something about chemistry that may not be comprehensible by chemists. Still, the class-specific differences in molecular properties are interesting to compare. Further, when the models misclassify a structure, we can interpret this both as suggestive of a shared drug mechanism, and as an opportunity for drug repurposing. Drug repurposing is an especially important aspect of this work because application of an already-approved drug is much less costly that *de novo* approval of a new chemical. We followed the setting from Aliper *et al.* that excluded chemicals with multiple therapeutic uses, but modeling these drugs will be important to more fully understand the model's potential for drug repurposing.

Acknowledgements

This work was supported by grants from the NIH NIGMS (P41 GM108538 and R35 GM118110 to JJC). JGM was supported by an NIH T15 fellowship (T15 LM007359). SL was supported by the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

References

(1) Lipinski, C.; Hopkins, A. Navigating Chemical Space for Biology and Medicine. *Nature* **2004**, *432*, 855.

(2) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52* (11), 2864–2875. https://doi.org/10.1021/ci300415d.

(3) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings1PII of Original Article: S0169-409X(96)00423-1. The Article Was Originally Published in Advanced Drug Delivery Reviews 23 (1997) 3–25.1. *Advanced Drug Delivery Reviews* **2001**, *46* (1), 3–26. https://doi.org/10.1016/S0169-409X(00)00129-0.

(4) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-like Chemical Space Based on GDB-17 Data. *J. Comput. Aided Mol. Des.* **2013**, *27* (8), 675–679. https://doi.org/10.1007/s10822-013-9672-4.

(5) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436.

(6) Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep Learning for Computational Chemistry. *Journal of Computational Chemistry* **2017**, *38* (16), 1291–1307. https://doi.org/10.1002/jcc.24764.

(7) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discovery Today* **2018**, *23* (6), 1241–1250. https://doi.org/10.1016/j.drudis.2018.01.039.

(8) Brown, A. L.; Kane, M. J. Preschool Children Can Learn to Transfer: Learning to Learn and Learning from Example. *Cognitive Psychology* **1988**, *20* (4), 493–523. https://doi.org/10.1016/0010-0285(88)90014-X.

(9) Hoo-Chang, S.; Roth, H. R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.; Summers, R. M. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE transactions on medical imaging* **2016**, *35* (5), 1285.

(10) Ching, T.; Himmelstein, D. S.; Beaulieu-Jones, B. K.; Kalinin, A. A.; Do, B. T.; Way, G. P.; Ferrero, E.; Agapow, P.-M.; Zietz, M.; Hoffman, M. M.; et al. Opportunities and Obstacles for Deep Learning in Biology and Medicine. *Journal of The Royal Society Interface* **2018**, *15* (141), 20170387. https://doi.org/10.1098/rsif.2017.0387.

(11) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science* **2018**, *4* (2), 268–276. https://doi.org/10.1021/acscentsci.7b00572.

(12) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361* (6400), 360. https://doi.org/10.1126/science.aat2663.

(13) Popova, M.; Isayev, O.; Tropsha, A. Deep Reinforcement Learning for de Novo Drug Design. *Sci Adv* **2018**, *4* (7). https://doi.org/10.1126/sciadv.aap7885.

(14) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chemical Science* **2018**, *9* (2), 513–530. https://doi.org/10.1039/C7SC02664A.

(15) Fernandez, M.; Ban, F.; Woo, G.; Hsing, M.; Yamazaki, T.; LeBlanc, E.; Rennie, P. S.; Welch, W. J.; Cherkasov, A. Toxic Colors: The Use of Deep Learning for Predicting Toxicity of Compounds Merely from Their Graphic Images. *J. Chem. Inf. Model.* **2018**, *58* (8), 1533–1543. https://doi.org/10.1021/acs.jcim.8b00338.

(16) Goh, G. B.; Siegel, C.; Vishnu, A.; Hodas, N. O.; Baker, N. Chemception: A Deep Neural Network with Minimal Chemistry Knowledge Matches the Performance of Expert-Developed QSAR/QSPR Models. *arXiv:1706.06689 [cs, stat]* **2017**.
(17) Goh, G. B.; Siegel, C.; Vishnu, A.; Hodas, N.; Baker, N. How Much Chemistry Does a Deep Neural Network Need to Know to Make Accurate Predictions? In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*; IEEE: Lake Tahoe, NV, 2018; pp 1340–1349. https://doi.org/10.1109/WACV.2018.00151.
(18) Mendez-Lucio, O.; Baillif, B.; Clevert, D.-A.; Rouquié, D.; Wichard, J. De Novo Generation of Hit-like Molecules from Gene Expression Signatures Using Artificial Intelligence. https://doi.org/10.26434/chemrxiv.7294388.v1.
(19) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36. https://doi.org/10.1021/ci00057a005.
(20) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29* (2), 97–101. https://doi.org/10.1021/ci00062a008.
(21) Weininger, D. SMILES. 3. DEPICT. Graphical Depiction of Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1990**, *30* (3), 237–243. https://doi.org/10.1021/ci00067a005.
(22) Xu, Y.; Pei, J.; Lai, L. Deep Learning Based Regression and Multiclass Models for Acute Oral Toxicity Prediction with Automatic Chemical Feature Extraction. *Journal of Chemical Information and Modeling* **2017**, *57* (11), 2672–2685. https://doi.org/10.1021/acs.jcim.7b00244.
(23) Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. *arXiv:1802.04364 [cs, stat]* **2018**.
(24) Liu, K.; Sun, X.; Jia, L.; Ma, J.; Xing, H.; Wu, J.; Gao, H.; Sun, Y.; Boulnois, F.; Fan, J. Chemi-Net: A Graph Convolutional Network for Accurate Drug Property Prediction. *arXiv preprint arXiv:1803.06236* **2018**.
(25) Lowe, H. J. Understanding and Using the Medical Subject Headings (MeSH) Vocabulary to Perform Literature Searches. *JAMA: The Journal of the American Medical Association* **1994**, *271* (14), 1103. https://doi.org/10.1001/jama.1994.03510380059038.
(26) Aliper, A.; Plis, S.; Artemov, A.; Ulloa, A.; Mamoshina, P.; Zhavoronkov, A. Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data. *Mol. Pharmaceutics* **2016**, *13* (7), 2524–2530. https://doi.org/10.1021/acs.molpharmaceut.6b00248.
(27) Subramanian, A.; Narayan, R.; Corsello, S. M.; Peck, D. D.; Natoli, T. E.; Lu, X.; Gould, J.; Davis, J. F.; Tubelli, A. A.; Asiedu, J. K.; et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **2017**, *171* (6), 1437-1452.e17. https://doi.org/10.1016/j.cell.2017.10.049.
(28) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.
(29) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32.
(30) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; et al. PubChem Substance and Compound Databases. *Nucleic Acids Res* **2016**, *44* (Database issue), D1202–D1213. https://doi.org/10.1093/nar/gkv951.
(31) Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. *arXiv:1611.05431 [cs]* **2016**.
(32) Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic Differentiation in PyTorch. **2017**.
(33) Smith, L. N. Cyclical Learning Rates for Training Neural Networks. *arXiv:1506.01186 [cs]* **2015**.

(34) Liu, S.; Alnammi, M.; Ericksen, S. S.; Voter, A. F.; Ananiev, G. E.; Keck, J. L.; Hoffmann, F. M.; Wildman, S. A.; Gitter, A. Practical Model Selection for Prospective Virtual Screening. **2018**. https://doi.org/10.1101/337956.
(35) Liu, S.; Chandereng, T.; Liang, Y. N-Gram Graph, A Novel Molecule Representation. *arXiv:1806.09206 [cs, stat]* **2018**.
(36) Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; Pande, V. Low Data Drug Discovery with One-Shot Learning. *ACS Cent. Sci.* **2017**, *3* (4), 283–293. https://doi.org/10.1021/acscentsci.6b00367.
(37) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 1947–1958. https://doi.org/10.1021/ci034160g.
(38) Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is Multitask Deep Learning Practical for Pharma? *J. Chem. Inf. Model.* **2017**, *57* (8), 2068–2076. https://doi.org/10.1021/acs.jcim.7b00146.
(39) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving Beyond Fingerprints. *Journal of Computer-Aided Molecular Design* **2016**, *30* (8), 595–608. https://doi.org/10.1007/s10822-016-9938-8.
(40) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-Learn: Machine Learning in Python. *Journal of machine learning research* **2011**, *12* (Oct), 2825–2830.
(41) Ruxton, G. D.; Beauchamp, G. Time for Some a Priori Thinking about Post Hoc Testing. *Behavioral Ecology* **2008**, *19* (3), 690–693. https://doi.org/10.1093/beheco/arn020.
(42) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. Imagenet Classification with Deep Convolutional Neural Networks. In *Advances in neural information processing systems*; 2012; pp 1097–1105.
(43) Wang, R.; Fu, Y.; Lai, L. A New Atom-Additive Method for Calculating Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **1997**, *37* (3), 615–621. https://doi.org/10.1021/ci960169p.
(44) Biastre, K.; Burnakis, T. Trospium Chloride Treatment of Overactive Bladder. *Annals of Pharmacotherapy* **2009**, *43* (2), 283–295. https://doi.org/10.1345/aph.1L160.
(45) Gelenberg, A. J.; Van Putten, T.; Lavori, P. W.; Wojcik, J. D.; Falk, W. E.; Marder, S.; Galvin-Nadeau, M.; Spring, B.; Mohs, R. C.; Brotman, A. W. Anticholinergic Effects on Memory: Benztropine versus Amantadine. *J Clin Psychopharmacol* **1989**, *9* (3), 180–185.
(46) Korotcov, A.; Tkachenko, V.; Russo, D. P.; Ekins, S. Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Molecular Pharmaceutics* **2017**, *14* (12), 4462–4475. https://doi.org/10.1021/acs.molpharmaceut.7b00578.
(47) Lenselink, E. B.; ten Dijke, N.; Bongers, B.; Papadatos, G.; van Vlijmen, H. W. T.; Kowalczyk, W.; IJzerman, A. P.; van Westen, G. J. P. Beyond the Hype: Deep Neural Networks Outperform Established Methods Using a ChEMBL Bioactivity Benchmark Set. *Journal of Cheminformatics* **2017**, *9* (1), 45.

Table 1:

Table 1: Summary of data classes and task groupings.		
MeSH ‘Therapeutic Uses’	Number of Examples	Task Subgroups
Anti-Neoplastic	1177	3, 5, 12
Cardiovascular	788	3, 5, 12
Central Nervous System	1140	3, 5, 12
Anti-Infective	2398	5, 12
Gastrointestinal	259	5, 12
Anti-Inflammatory	373	12
Dermatological	116	12
Hematologic	267	12
Lipid Regulating	164	12
Reproductive Control	148	12
Respiratory System	101	12
Urological	26	12

Table 2:

Table 2: Model validation accuracy among various machine learning strategies.
[1] Pathway Activation with Support Vector Machine (SVM) from [19] (different training and validation splits)
[2] Pathway Activation with Deep Neural Network (DNN) from [19] (different training and validation splits)
[3] Molecule Images with Convolutional Neural Network (IMG+CNN, default hyperparameters) reported here
[4] Molecular Fingerprints with Random Forest (MFP+RF) reported here

Subtask	SVM[1]	DNN[2]	IMG+CNN[3]	MFP+RF[4]
3-class	0.530	0.701	0.817 ± 0.005	0.886 ± 0.001
5-class	0.417	0.596	0.792 ± 0.002	0.871 ± 0.008
12-class	0.366	0.546	0.751 ± 0.004	0.833 ± 0.005

Table 3:

Table 3. Comparison of validation set accuracy from different strategies for CNN training. Only the output layer was trained from the resnext101_64 model using data from the 3-class subtask.

Batch Size	255 epochs	255 epochs, cyclic learning rate	1023 epochs	1023 epochs, cyclic learning rate
50	0.810 ± 0.007	0.802 ± 0.009	0.812 ± 0.007	0.813 ± 0.005
100	0.805 ± 0.003	0.804 ± 0.005	0.811 ± 0.004	0.818 ± 0.006
200	0.809 ± 0.014	0.807 ± 0.010	0.807 ± 0.007	0.817 ± 0.005
400	0.796 ± 0.005	0.794 ± 0.005	0.811 ± 0.006	0.807 ± 0.007
800	0.785 ± 0.006	0.790 ± 0.006	0.809 ± 0.007	0.808 ± 0.006
Grey, 200	0.796 ± 0.008	0.797 ± 0.009	0.801 ± 0.002	0.801 ± 0.005

Figure Legends:

Figure 1: Structure-based drug classification pipelines. All chemical identifiers for 12 major medical subheadings (MeSH) therapeutic use classifications were downloaded from pubchem. The identifiers were then filtered for only those mapping to a single MeSH class and then converted to SMILES. SMILES representations were converted to either 2-dimensional images or molecular fingerprints. The 2D images were used to train a convolutional neural network (CNN). The fingerprints were used to train a random forest (RF). The models were used separately to predict classes of drugs using stratified 5-fold cross validation.

Figure 2: Principal component analysis of the input vectors to the random forest models: (A) Molecular fingerprints or (B) latent output from CNN's last hidden layer.

Figure 3: Confusion matrices from IMG + CNN classification performance over sets of drug molecules belonging to (A) 3, (B) 5, and (C) 12 therapeutic classes. Each matrix is the classification using the random 20% validation set. Each confusion matrix is a random representative of the five models from the 5-fold cross validation.

Figure 4: Class-level distribution of Lipinski's drug-like properties: (A) molecular weight, (B) XlogP, (C) hydrogen bond acceptor count, and (D) hydrogen bond donor count.

Figure 5: Example of misclassified drug that reveal mechanism and re-purposing opportunities. (A) Structure of Trospium, a urological drug known to act as a cholinergic muscarinic antagonist, which was classified by the model as a CNS agent. (B) Structure of Benztropine, a muscarinic antagonist used to treat Parkinson's disease.

Figure 1

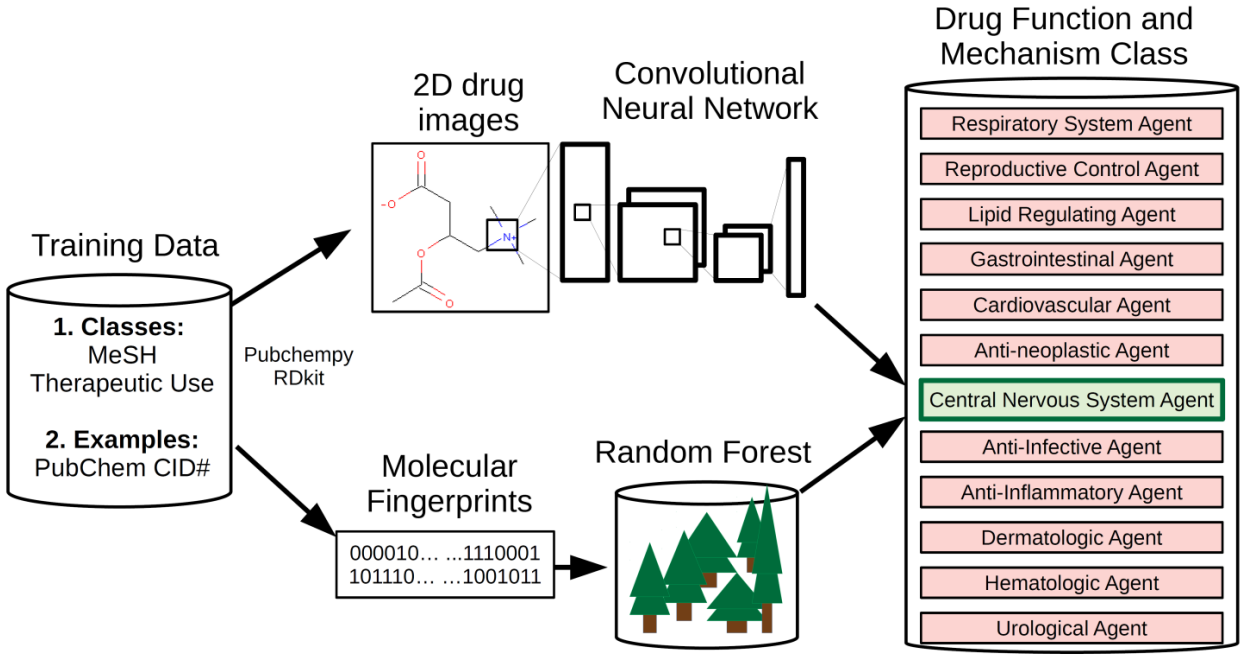

Drug Function and Mechanism Class
2D drug images
Convolutional Neural Network
Training Data
1. Classes:
MeSH
Therapeutic Use
2. Examples:
PubChem CID#
Pubchempy
RDkit
Molecular Fingerprints
000010… ...1110001
101110… ...1001011
Random Forest
Respiratory System Agent
Reproductive Control Agent
Lipid Regulating Agent
Gastrointestinal Agent
Cardiovascular Agent
Anti-neoplastic Agent
Central Nervous System Agent
Anti-Infective Agent
Anti-Inflammatory Agent
Dermatologic Agent
Hematologic Agent
Urological Agent

Figure 2

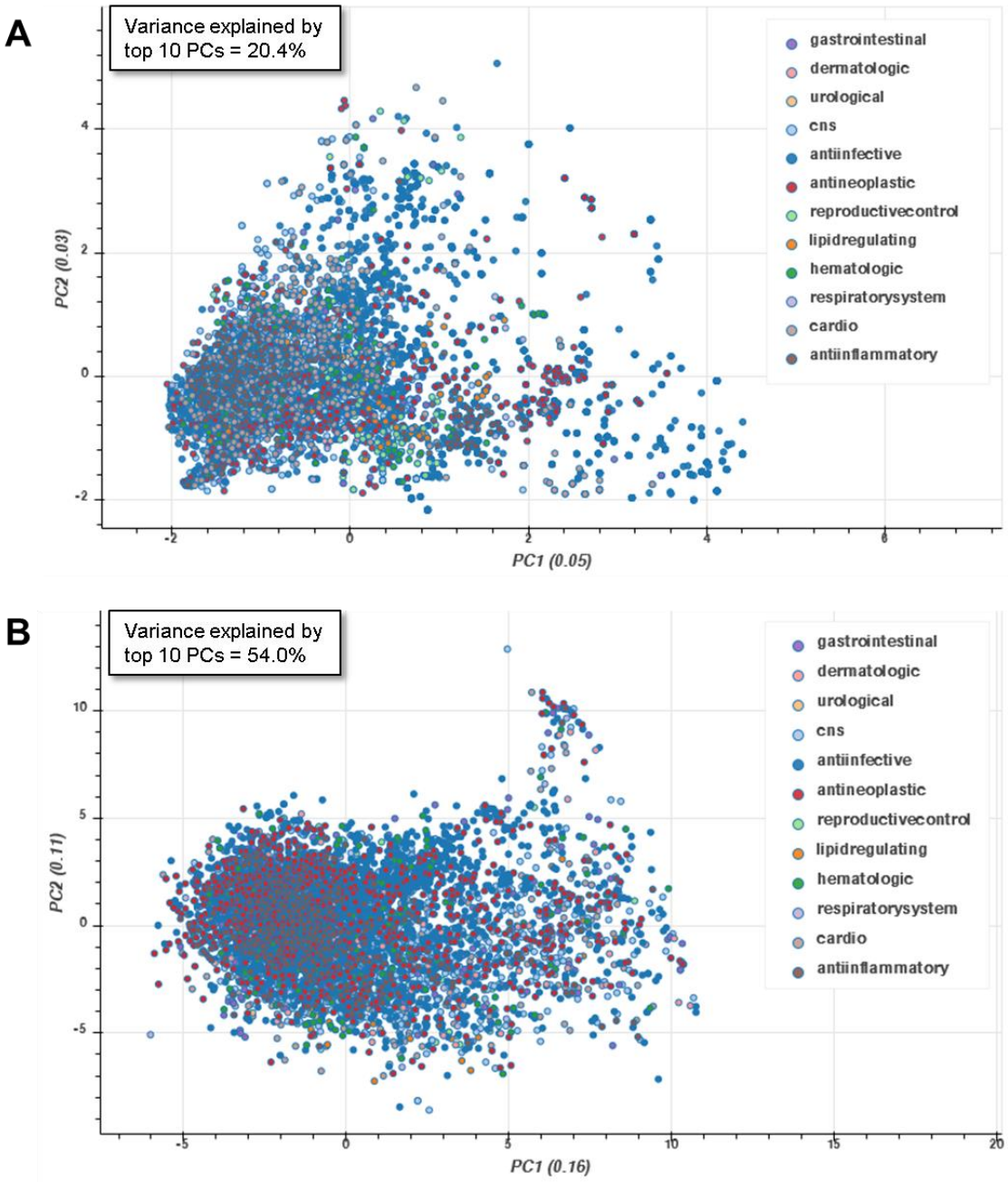

A
Variance explained by top 10 PCs = 20.4%
gastrointestinal
dermatologic
urological
cns
antiinfective
antineoplastic
reproductivecontrol
lipidregulating
hematologic
respiratorysystem
cardio
antiinflammatory
PC2 (0.03)
PC1 (0.05)
B
Variance explained by top 10 PCs = 54.0%
gastrointestinal
dermatologic
urological
cns
antiinfective
antineoplastic
reproductivecontrol
lipidregulating
hematologic
respiratorysystem
cardio
antiinflammatory
PC2 (0.11)
PC1 (0.16)

Figure 3

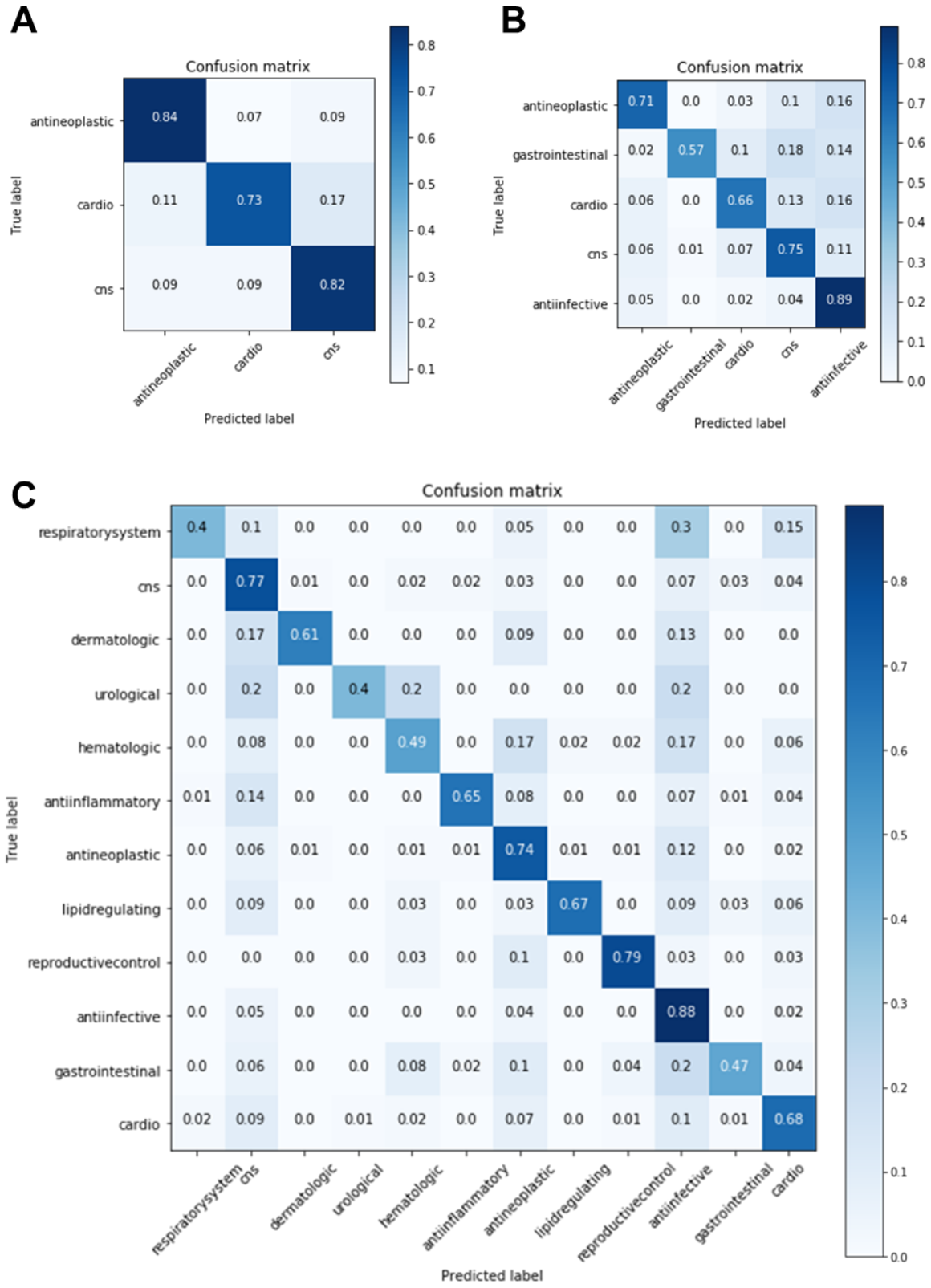
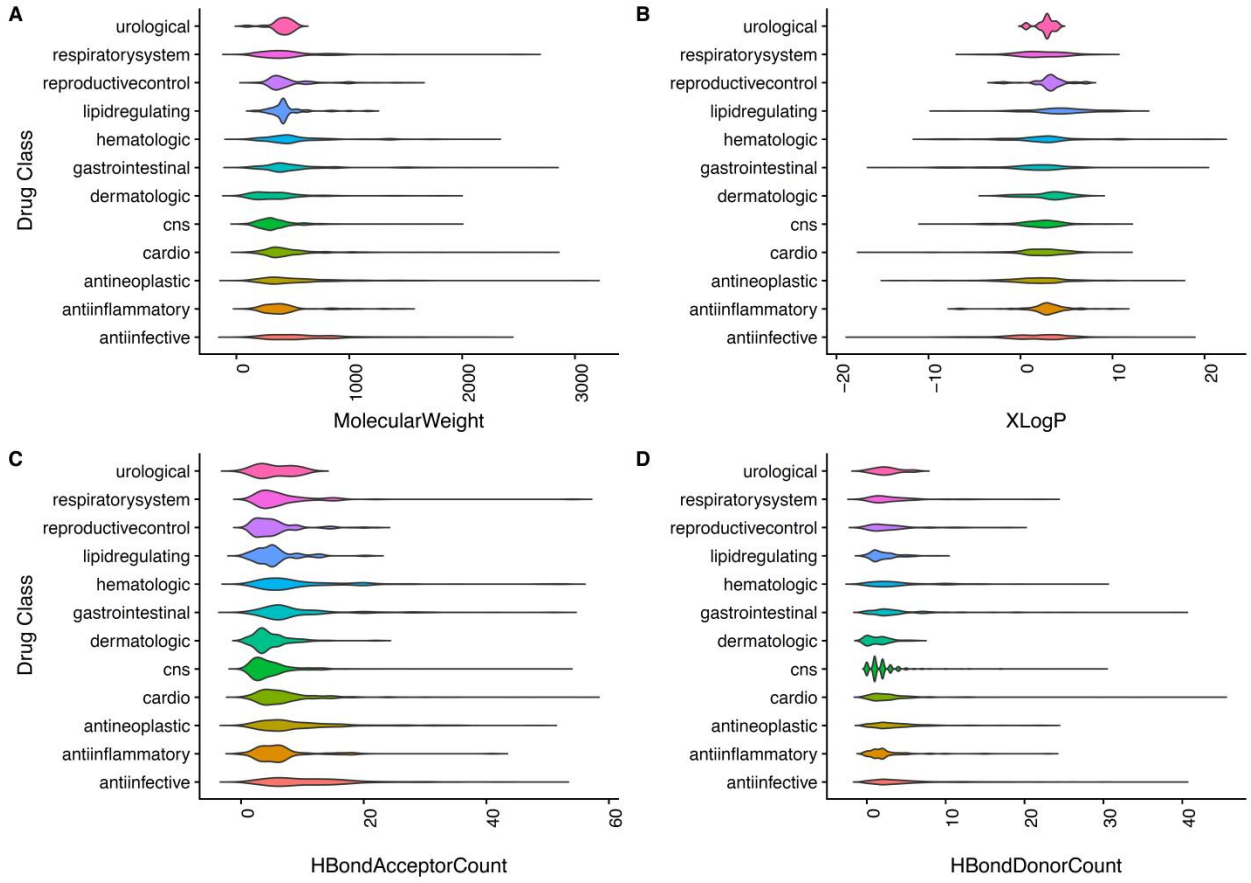

A
Confusion matrix
True label
Predicted label
antineoplastic
cardio
cns
0.84 0.07 0.09
0.11 0.73 0.17
0.09 0.09 0.82
B
Confusion matrix
True label
Predicted label
antineoplastic
gastrointestinal
cardio
cns
antiinfective
0.71 0.0 0.03 0.1 0.16
0.02 0.57 0.1 0.18 0.14
0.06 0.0 0.66 0.13 0.16
0.06 0.01 0.07 0.75 0.11
0.05 0.0 0.02 0.04 0.89
C
Confusion matrix
True label
Predicted label
respiratorysystem 0.4 0.1 0.0 0.0 0.0 0.0 0.05 0.0 0.0 0.3 0.0 0.15
cns 0.0 0.77 0.01 0.0 0.02 0.02 0.03 0.0 0.0 0.07 0.03 0.04
dermatologic 0.0 0.17 0.61 0.0 0.0 0.0 0.09 0.0 0.0 0.13 0.0 0.0
urological 0.0 0.2 0.0 0.4 0.2 0.0 0.0 0.0 0.0 0.2 0.0 0.0
hematologic 0.0 0.08 0.0 0.0 0.49 0.0 0.17 0.02 0.02 0.17 0.0 0.06
antiinflammatory 0.01 0.14 0.0 0.0 0.0 0.65 0.08 0.0 0.0 0.07 0.01 0.04
antineoplastic 0.0 0.06 0.01 0.0 0.01 0.01 0.74 0.01 0.01 0.12 0.0 0.02
lipidregulating 0.0 0.09 0.0 0.0 0.03 0.0 0.03 0.67 0.0 0.09 0.03 0.06
reproductivecontrol 0.0 0.0 0.0 0.0 0.03 0.0 0.1 0.0 0.79 0.03 0.0 0.03
antiinfective 0.0 0.05 0.0 0.0 0.0 0.0 0.04 0.0 0.0 0.88 0.0 0.02
gastrointestinal 0.0 0.06 0.0 0.0 0.08 0.02 0.1 0.0 0.04 0.2 0.47 0.04
cardio 0.02 0.09 0.0 0.01 0.02 0.0 0.07 0.0 0.01 0.1 0.01 0.68
respiratorysystem cns dermatologic urological hematologic antiinflammatory antineoplastic lipidregulating reproductivecontrol antiinfective gastrointestinal cardio

Figure 4


A
B
C
D
Drug Class
urological
respiratorysystem
reproductivecontrol
lipidregulating
hematologic
gastrointestinal
dermatologic
cns
cardio
antineoplastic
antiinflammatory
antiinfective
MolecularWeight
XLogP
HBondAcceptorCount
HBondDonorCount

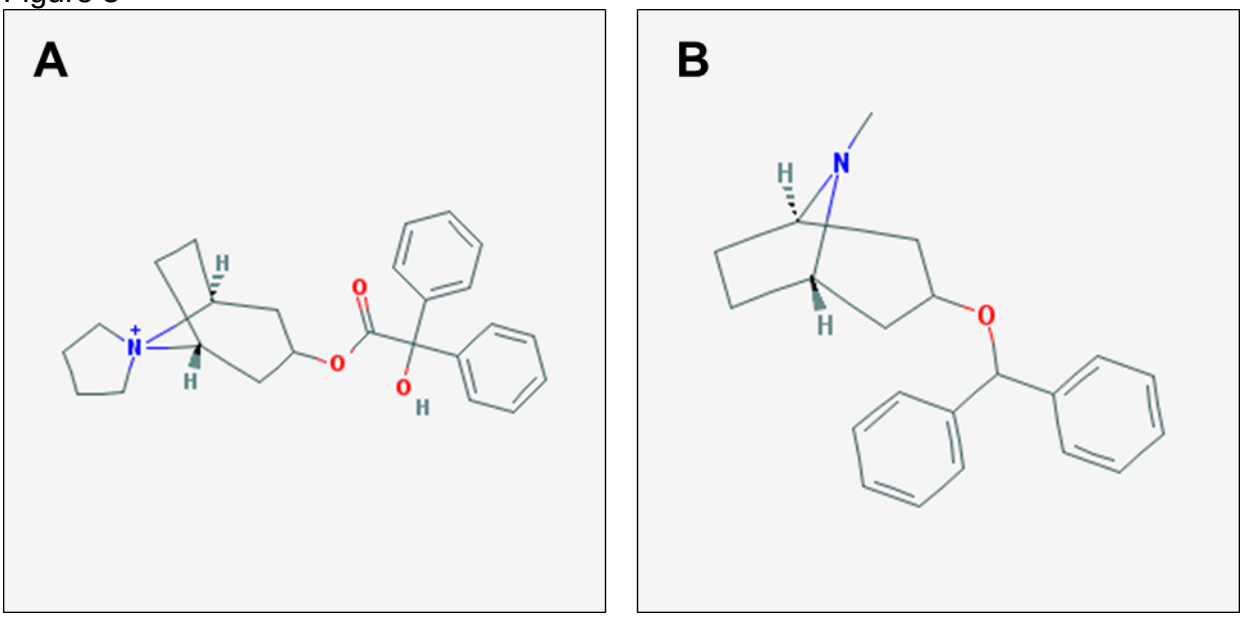



Figure 5