# Development of a suite of models for molecular property prediction

Allyson Stanley, Brenden Pelkie, Elysiana Batingan, Jinny Ryu, Samarth Agarwal, Vince Choi
*Department of Chemical Engineering, University of Washington, Seattle, WA, USA*

Wisecube

## ADMET Properties Determine Drug Viability

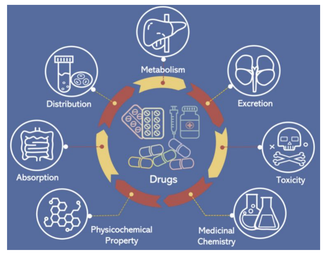



Figure 1: ADMET Properties describe drug interactions with the humand body

ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) properties play a crucial role in the success of a drug candidate. According to statistics, unsatisfying ADMET properties in drug molecules account about 60% of the failures in the drug development process[1].

Our sponsor, Wisecube, proposed that an early prediction for such properties would possibly reduce the amount of failures and possibly the cost of development.

## Project Goals

**We aimed to develop and deploy a suite of machine learning models to predict 25 ADMET properties.**

*Milestones*

1. Establish baseline models that utilize architectures from and match the performance of previously published models[2].
2. Improve model performance by evaluating additional featurization, tuning hyperparameters, and experimenting with various ML algorithms.
3. Evaluate the performance of novel ML methods in predicting ADMET properties compared to classic ML methods (e.g. ensemble methods).
4. Fine-tune hyperparameters, deploy and document models.

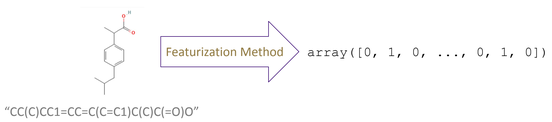
*ADMET Property list*

| | | | |
|---|---|---|---|
| - Ames<br>- BBB<br>- Caco-2<br>- Pgp Inhibitor<br>- Pgp Substrate | - F-20<br>- F-30<br>- Inhibitor and substrate status for : CYP1A2, 3A4, 2C19, 2C9 | - HIA<br>- hERG<br>- H-HT<br>- LogD<br>- LogS | - SAMPL<br>- SIDER<br>- Tox21<br>- ToxCast<br>- ClinTox |

## Methods in Molecular Machine Learning

### Featurization:

- Molecules commonly represented as "SMILES" strings: e.g. Ibuprofen = CC(C)CC1=CC=C(C=C1)C(C)C(=O)O
- ML models do not learn properties well directly from SMILEs string inputs
- **Featurization** encodes molecular structure and knowledge of chemistry into a form ML models can work with

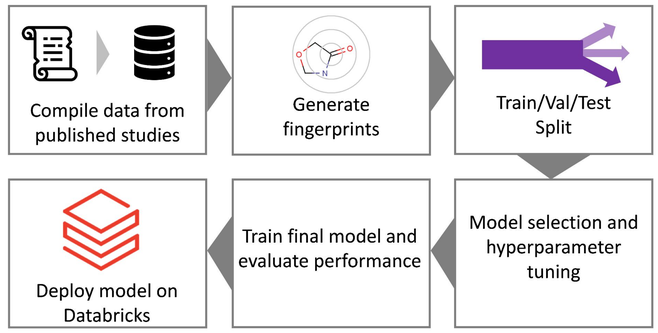



- Established featurization methods include:
  - Morgan (Extended Connectivity) Fingerprints[3]
  - MACCS Keys[4]
  - Physicochemical descriptors (i.e. RDKit Descriptors)[5]

### Chemistry - Informed Splits

- Molecular ML models were expected to generalize to new domains
- Splits were according to the molecular structure or diversity which better the model performance
- Here, we used Deepchem's implementation of a diversity splitter[6]

## Workflow for Established Methods Models




## Novel Machine Learning Method Experimentation

### Variational Autoencoder Features

- Generated VAE latent space features from pre-trained encoder[7], trained a neural network to predict properties.
- Performance was promising, but the pre-trained VAE model not trained for all species in our datasets.

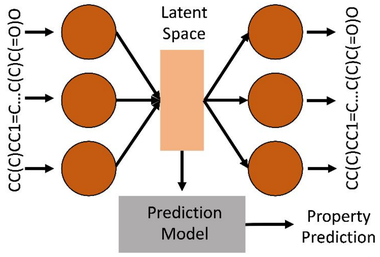



Figure 2: VAEs learn a latent space embedding that can be used as an input feature

### Lipophilicity Transfer Learning

Main Concept - Lipophilicity is a physical property that can be calculated cheaply[8].

*Experimentation*

1. Pre-train a model with calculated values to develop a diverse chemical space
2. Transfer the pre-trained model to predict desired property with limited data

Our initial testing did not show performance improvements when comparing with a classic machine learning model.

## Model Performance Examples

### Caco-2 Model

A Gradient Boosting regressor model predicting the drug permeability in the logarithmic scale.

| Target Properties | Benchmark Model | Current Model |
|---|---|---|
| Drug Permeability (log) | $RMSE_{Test}$: 0.774 | $RMSE_{Test}$: 0.422<br>R2 score: 0.660 |

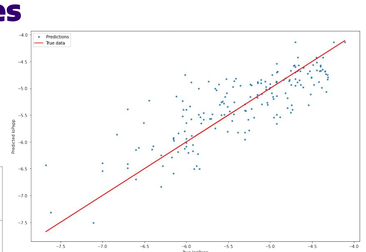



Figure 3: Parity plot for CACO-2 Model

### F-50 Model

A Random Forest classifier model predicting the drug bioavailability (cut-off at 50%)

| Target Properties | Benchmark Model | Current Model |
|---|---|---|
| Bioavailability (50%) | Accuracy: 0.67<br>AUC: 0.72 | Accuracy: 0.68<br>AUC: 0.74 |

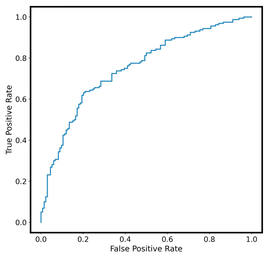



Figure 4: Parity plot for F-50 Model

### LogS Model

A Gradient Boosting regressor model predicting the drug solubility in the logarithmic scale.

| Target Properties | Benchmark Model | Current Model |
|---|---|---|
| Aqueous Solubility (log) | $RMSE_{Test}$: 0.712<br>R2 score: 0.979 | $RMSE_{Test}$: 0.440<br>R2 score: 0.930 |

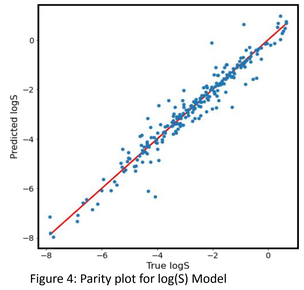
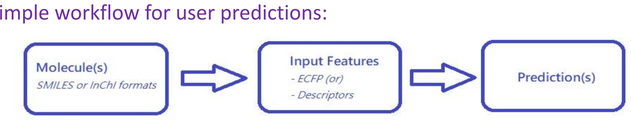



Figure 4: Parity plot for log(S) Model

*Other models had similar results, where they improved upon the benchmark performances.

## Project Outcomes

- 25 machine learning models for 25 different ADMET properties were developed, tested, deployed and documented.
- The current version mostly adopted the ensemble methods and the Support Vector Machine while one in particular adopted a neural network after a successful experimentation.
- Simple workflow for user predictions:




- The current version achieved better or competitive performances compared to benchmark or prior models.

## Potential Future Work

- We feel we have thoroughly explored the ML space of using published datasets and traditional molecular featurizations and ML models
- VAE features and transfer learning could hold promise with investment of additional time and resources
- Current state-of-the-art involves graph neural networks[9]
  - Select a graph architecture to use
  - Likely would require partnering with someone who has data, or investing in a more extensive data-gathering process to create significantly larger datasets.

## Acknowledgements

This project is sponsored by Wisecube and supported by University of Washington.
Special Thanks to Alex Thomas for mentoring all contributors and all instructors for supporting the project.

## References

1. Kesharwani, R.K. et al. (2020). DOI 10.1007/978-981-15-6815-2_4
2. Dong, J., Wang, NN., Yao, ZJ. *et al.* ADMETlab: a platform for systematic ADMET evaluation based on a comprehensively collected ADMET database. *J Cheminform* **10,** 29 (2018). https://doi.org/10.1186/s13321-018-0283-x
3. Rogers, David, and Mathew Hahn. "Extended-connectivity fingerprints." *Journal of chemical information and modeling* 50.5 (2010): 742-754.
4. Durant, Joseph L., et al. "Reoptimization of MDL keys for use in drug discovery." *Journal of chemical information and computer sciences* 42.6 (2002): 1273-1280.
5. Landrum, Greg. "Rdkit documentation." *Release* 1.1-79 (2013): 4.
6. Ramsundar, Bharath. *Molecular machine learning with DeepChem*. Diss. Stanford University, 2018.
7. Dollar, Orion, et al. "Attention-based generative models for de novo molecular design." *Chemical Science* 12.24 (2021): 8362-8372.
8. Chen, Yan-Kai, Steven Shave, and Manfred Auer. "MRlogP: Transfer Learning Enables Accurate logP Prediction Using Small Experimental Training Datasets." *Processes* 9.11 (2021): 2029.
9. Yang, Kevin, et al. "Analyzing learned molecular representations for property prediction." *Journal of chemical information and modeling* 59.8 (2019): 3370-3388.

Workflow figure fingerprint image: https://docs.chemaxon.com/display/docs/extended-connectivity-fingerprint-ecfp.md
Admet property wheel: https://admetmesh.scbdd.com