Kolmogorov-Arnold Networks (KANs) for Explainable Molecular Property Prediction

Chistian Geils, Christopher Sutton, Sophya Garashchuk, Vitaly Rassolov

Department of Chemistry and Biochemistry, University of South Carolina

Problem and Motivation

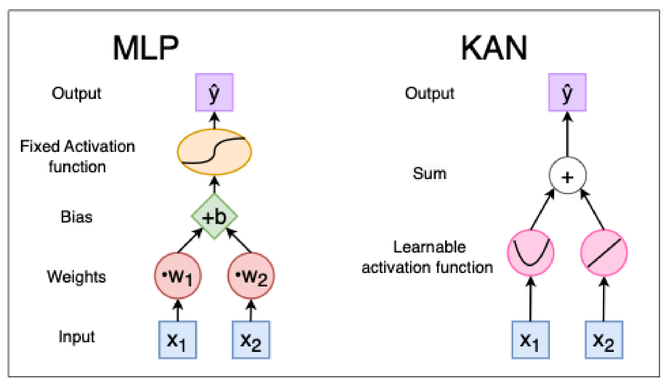
- Making in silico predictions about molecular properties is essential for computational drug and material design
- Different approaches are available, with tradeoffs between accuracy and speed
  - Ultimate solution: solve the Schrödinger equation[1]. Usually not feasible.
  - Density functional theory[2] is the gold-standard, but scales poorly: >= $O(n^3)$
  - **Machine learning**: highly efficient, learn from experimental/simulation data
- **Explainability:** experimentalists want to know why a model makes a prediction. Huge black-box neural nets are not only nearly impossible to explain, but also much more expensive to train and run inference on.

KAN Theory, Implementation, and Comparison to MLP

- **Kolmogorov-Arnold representation theorem**[3]: can represent an arbitrarily complex multivariate function as a sum of nested univariate functions

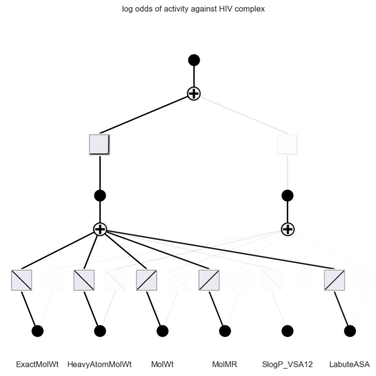
$$f(x_1 \dots x_d) = \sum_{q=1}^{2d+1} g_q \left(\sum_{p=1}^{d} \psi_{p,q}(x_p) \right)$$

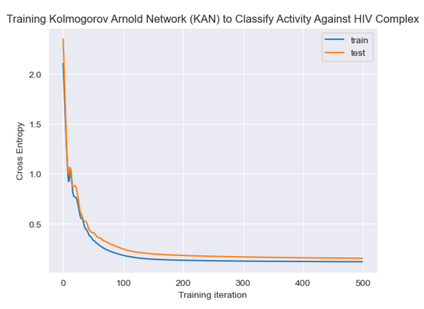
- **Multi-layer perceptron:** multiply / add inputs with scalar weights & biases, then pass through fixed activation function (ReLU, Sigmoid, Tanh, etc.)
- **Kolmogorov Arnold network**[4]**:** all learning happens in *shape of activation functions* (B-spline + SiLU) on edges, which replace weights and biases

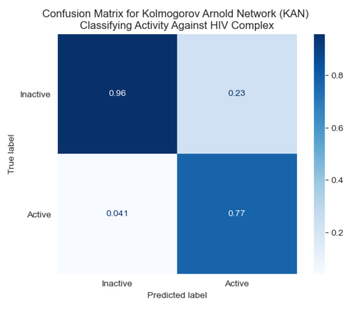


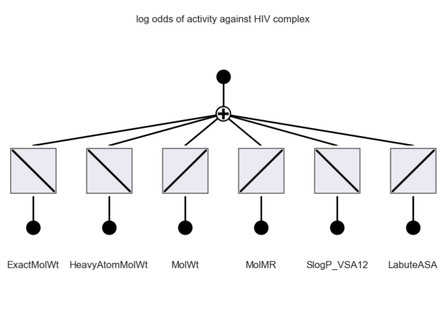
Task 1: Classifying Activity Against HIV Complex

- Motivation: accelerated *and* explainable virtual screening for designing compounds with activity against HIV
- 40,745 samples, 217 molecular descriptors (features)
- 3 models: one full, 2 with reduced feature set (4 features) selected based on importance in training set
- Impute median for missing values, max for infinity, min for –inf
- Decent precision (.61 test), terrible recall (.069 test)



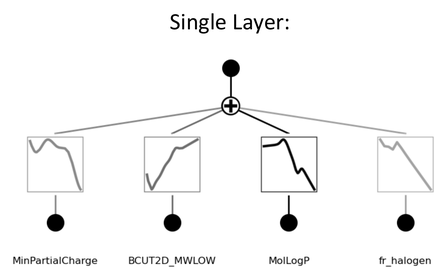


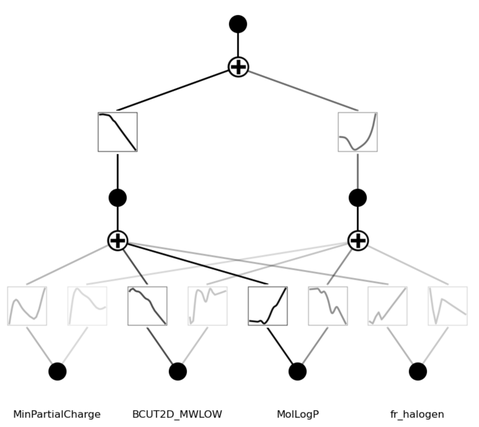


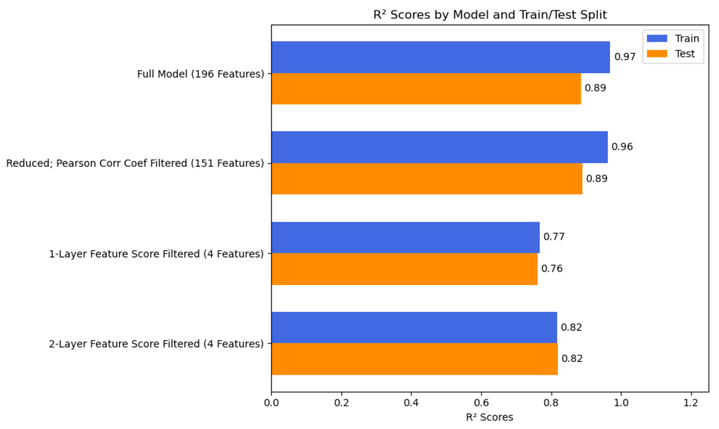


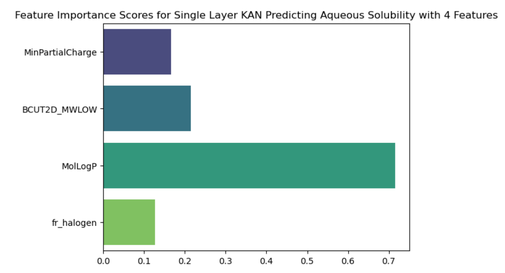
Task 2: Predicting Aqueous Solubility

- Motivation: determining aqueous solubility is essential for drug discovery – insoluble compounds more difficult to absorb[5]
- 1144 Samples w/ experimentally determined aqueous solubility
- Feature selection pipeline:
  1. Generate 217 molecular descriptors
  2. Filter to 196 based on low variance
  3. Filter to 151 based on Pearson corr. coef.
  4. Filter to 4 based on feature importance
- 80/20 train/test split based on Murcko scaffolds
- Train both single and 2-layer KANs









Conclusions and Future Work

- Somewhat finnicky to train – single extreme value broke my model, strange loss fluctuations for classification
- Based on task 1, KANs are likely not ideal for Quantitative Structure-Activity Relationship (QSAR) modeling; too many features needed in order to be interpretable
- Smaller models can be competitive with larger ones (see task 2)
- Susceptible to overfitting depending on number of basis functions in B-splines – could potentially be corrected with regularization
- Highly interpretable for small number of features (<= 5), and layers (<= 2), cumbersome with more
- KANs could be a powerful tool for unsupervised nonlinear chemical data analysis
- Compare to GNNs, more extensive architecture/hyperparameter search, implement cross-validation, repeat training over many trials to obtain confidence estimates for model performance metrics

References

1. Scherbela, M.; Reisenhofer, R.; Gerard, L.; Marquetand, P.; Grohs, P. Solving the Electronic Schrödinger Equation for Multiple Nuclear Geometries with Weight-Sharing Deep Neural Networks. Nat. Comput. Sci. 2022, 2 (5), 331–341.
2. Whitfield, J. D.; Schuch, N.; Verstraete, F. The Computational Complexity of Density Functional Theory. arXiv June 5, 2013
3. Schmidt-Hieber, J. The Kolmogorov–Arnold Representation Theorem Revisited. Neural Netw. 2021, 137, 119–126.
4. Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T. Y.; Tegmark, M. KAN: Kolmogorov-Arnold Networks. arXiv June 16, 2024..
5. Savjani, K. T., Gajjar, A. K., & Savjani, J. K. (2012). Drug solubility: importance and enhancement techniques. ISRN pharmaceutics, 2012, 195727.