An Assessment of Automated Quantitative Structure-Activity Relationship Modeling on Drug Discovery for Novel Treatment of Blood Disorders

Parnika Agrawal[1], Di Wu[1], David Williams[2]

1. Department of Biostatistics, University of North Carolina at Chapel Hill, Gillings School of Global Public Health
2. Department of Pathology and Lab Medicine, University of North Carolina at Chapel Hill School of Medicine

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

INTRODUCTION

Sickle cell disease

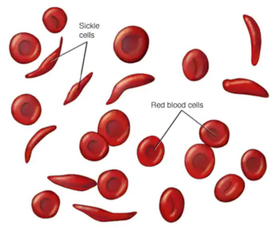


Figure 1. Healthy red blood cells compared to sickle-cells.
Source: Mayo Clinic[1]

- Inherited, lack of healthy red blood cells
- Symptoms: anemia, pain, swelling of hands and feet, and frequent infections
- >100,000 Americans have SCA
- Diagnoses and treatments → 30 year LE increase

A Novel Therapeutic Target

- HbF shown to alleviate SCA symptoms in infants
- Recruitment of the NuRD complex silences HbF production
- ETO2 recruits NuRD via MYND domain binding
- New target: disruption of a target protein domain (MYND Domain) to stop NuRD recruitment and HbF silencing

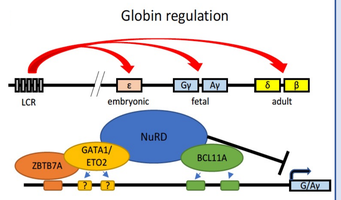


Figure 2. *Visualization of interaction between NuRD complex and ETO2 protein, as well as NuRD recruitment to globin regulation gene for silencing of HbF.*

Drug Discovery

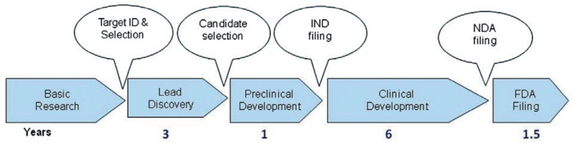
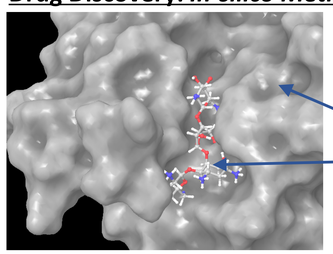


Figure 3. *From Hughes et. al. 2011: Traditional drug discovery pathway*

Drug Discovery: *In-silico Methods*



← Figure 4. *Model of in-silico docking of compounds*

Figure 5. *From Liu et. al. 2007;* 3D structure of the MYND domain (represented in orange), with its native peptide binder represented in green. →

METHODS

Library Selection
- NCI: large collection of free drug-like compounds suited for similar targets
- Enamine: more representative of large chemical spaces

Ligand Preparation
- Use of Schrödinger LigPrep Software
- 2D → 3D conversion
- Ensures chemical correctness

Docking
- Schrödinger Glide: HTVS → SP → XP
- Goes from least to most computationally expensive, filters at each stage
- Outputs docking scores for top % of initial library

Threshold Determination
- Schrödinger recommends selection a threshold for "hits" based on preliminary screening
- Threshold will be used moving forward as cutoff for binding

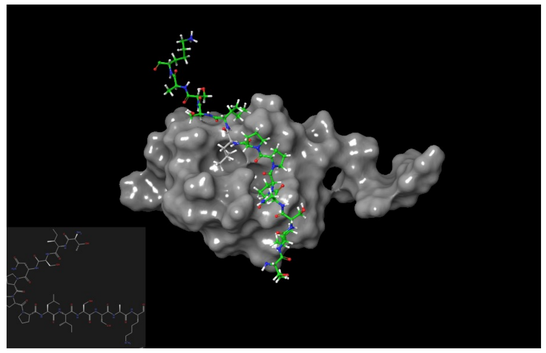


Figure 6. *Prepared protein (MYND Domain) with docked ligand for grid generation*

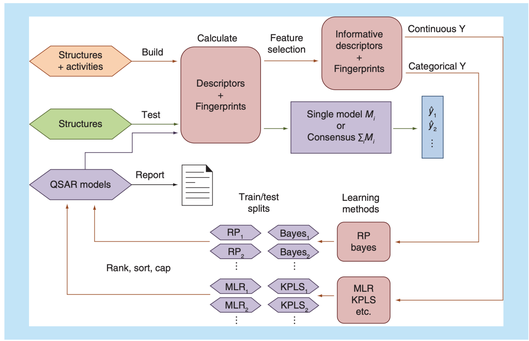


Figure 7. *From Dixon et. al. 2016 Sample schematic of QSAR workflow*

RESULTS

Preliminary Docking Study

- Docking score values range from -5.2931 to -13.569
- Threshold for "hits" ≤ -10
- 4.1% of *returned* compounds are hits
  - < 0.01% of compounds screened were hits

Data Splitting Ratios

Table 1. Summary of model comparison metrics across all 3 data splitting ratios.

	*Q^2	*RMSE	ROC-AUC	PR-AUC
Model A	0.87106	0.59421	0.92435	0.38665
Model B	0.87029	0.58852	0.92219	0.40152
Model C	0.85689	0.61476	0.91984	0.39659

→ Model A (64:16:20) was selected for remaining analysis.

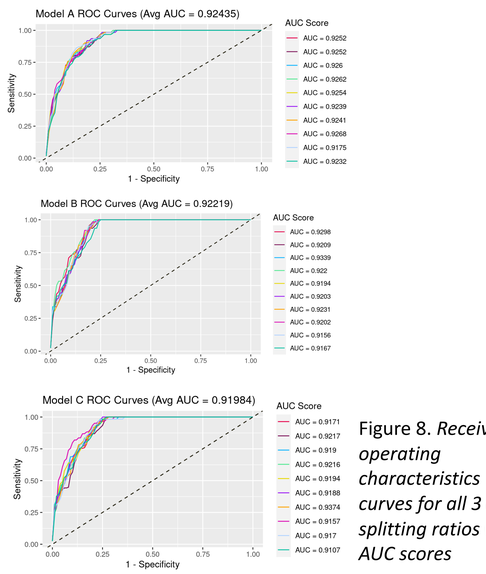


Figure 8. *Receiver operating characteristics curves for all 3 data splitting ratios with AUC scores*

RESULTS

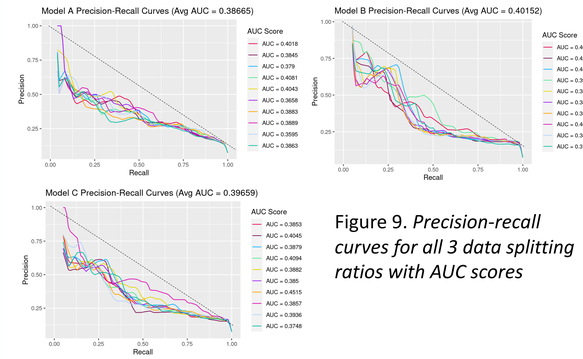


Figure 9. *Precision-recall curves for all 3 data splitting ratios with AUC scores*

Table 2. One-way ANOVA results of Q^2 across models created to test inclusion dependency

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
Model	4	0.003238	0.0008096	12.79	2.52e-05
Residuals	20	0.001266	0.0000633		

Table 3. Measures of accuracy of sample prediction

RMSE		
Overall	**"Hits"**	**"Misses"**
0.59614	1.4309	0.52398
Percent Hits Captured (%)		
64.8148		

CONCLUSIONS

- AutoQSAR modeling for this protein system was unsuccessful
  - Starting structure was from NMR data, not crystallography → uncertainty in foundational information regarding the binding pocket
  - No verified drug-like binders → training set data used to build the models may have been flawed
  - High variability of models depending on specific inclusion/exclusion of binders → overfitting and descriptor patterns found across ligands was not very robust

REFERENCES

1. Sickle cell anemia - Symptoms and causes - Mayo Clinic. Accessed January 27, 2022. https://www.mayoclinic.org/diseases-conditions/sickle-cell-anemia/symptoms-causes/syc-20355876
2. Hughes JP, Rees S, Kalindjian SB, Philpott KL. Principles of early drug discovery. *Br J Pharmacol*. 2011;162(6):1239-1249. doi:10.1111/j.1476-5381.2010.01127.x
3. Liu Y, Chen W, Gaudet J, et al. Structural basis for recognition of SMRT/N-CoR by the MYND domain and its contribution to AML1/ETO's activity. *Cancer Cell*. 2007;11(6):483-497. doi:10.1016/j.ccr.2007.04.010
4. Dixon SL, Duan J, Smith E, Von Bargen CD, Sherman W, Repasky MP. AutoQSAR: an automated machine learning tool for best-practice quantitative structure-activity relationship modeling. *Future Med Chem*. 2016;8(15):1825-1839. doi:10.4155/fmc-2016-0093
5. Schrödinger, LLC. *Schrödinger Release 2022-3: Protein Preparation Wizard*. Schrödinger, LLC; 2021.
6. Schrödinger, LLC. *Schrödinger Release 2022-3: LigPrep*. Schrödinger, LLC; 2021.
7. Schrödinger, LLC. *Schrödinger Release 2022-3: Glide*. Schrödinger, LLC; 2021.
8. Schrödinger, LLC. *Schrödinger Release 2022-3: AutoQSAR*. Schrödinger, LLC; 2021.
9. Schrödinger, LLC. *Schrödinger Release 2022-3: Canvas*. Schrödinger, LLC; 2021.