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

By
Parnika Agrawal

Senior Honors Thesis
Department of Biostatistics
University of North Carolina at Chapel Hill

Spring 2023

Approved by:

Di Wu

Dr. Di Wu, Thesis Advisor

David Williams

Dr. David Williams, Committee Member

Fei Zou

Dr. Fei Zou, Committee Member

Abstract

The MYND domain of the ETO2 protein is a novel target for drugs aimed at treating sickle cell disease and related blood disorders.[1,2] This study explored the application of automated quantitative structure-analysis relationship (QSAR) modeling, a machine learning application of *in-silico* drug discovery, to this target protein system using Schrödinger's AutoQSAR software. The protein target in this study currently has no known drug-like binders, allowing the assessment of conducting every stage of lead discovery *in-silico*. A training set was generated using a preliminary docking study, from which QSAR models were built and verified across varying data splitting ratios. The most favorable of these models was subject to further testing to assess overfitting and ligand-inclusion/exclusion dependency, and a test set of QSAR predictions was evaluated for accuracy. The use of AutoQSAR modeling for this system was found to be unsuccessful, likely associated with the lack of verified drug-like binders in the training set.

Introduction and Background

Sickle Cell Anemia and a Novel Therapeutic Target

Sickle cell anemia (SCA) is an inherited red blood cell disorder characterized by a lack of healthy red blood cells needed to carry oxygen throughout the body, due to the deformity of typically disc-shaped cells into a sickle-shape.[3] The shape of healthy red blood cells is optimal for oxygen diffusion and travel through blood vessels, but sickle-shaped blood cells lack these characteristics and become easily stuck in vessels. This leads to symptoms such as anemia, episodes of pain, swelling of hands and feet, and frequent infections.[3]

Over 100,000 Americans are currently affected by SCA, and greater than 300,000 babies are born each year with SCA globally, with the majority of global cases occurring in Nigeria, the Democratic Republic of the Congo, and India.[4,5] The number of people with sickle cell disease is expected to increase by 30 percent by the year 2050.[6]

Diagnosis and treatment are vitally important to improving outcomes of SCA, and have demonstrated significant success. Since 1970, sickle cell patients' life spans in the United States have increased from 20 years old to the majority living past 50, largely attributable to improved diagnosis and treatment.[7]

One treatment that has been explored over the last several decades is the use of the body's native fetal hemoglobin (HbF) to compensate for SCA deficiencies. In a red blood cell, hemoglobin is the protein that carries oxygen.[3] Adult hemoglobin becomes dominant in humans by 2 months postnatal, but HbF is naturally produced until 6 months.[8]

The presence of HbF was noticed to affect SCA symptoms when infants with the condition did not show symptoms, and their blood cells did not sickle or deform as extensively.[8]

Studies have further shown that any increase in HbF was met with improvements in symptoms, leaving reactivation of the gene controlling HbF to be a promising treatment to explore.[8]

HbF activation and deactivation is controlled by proteins that interact directly with the gene.[9] Research has identified one such protein complex, the NuRD complex, which results in the silencing of HbF when recruited to the globin regulation gene.[1,2] Several proteins, specifically transcription factors, are involved in NuRD recruitment.

ETO2 is one such protein that recruits the NuRD complex–the MYND domain of the ETO2 protein binds to NuRD by recognizing a polyproline-leucine motif (Figure 1).[2] Disrupting this interaction blocks NuRD-dependent HbF silencing. This known binding site on MYND may be a novel drug-target, but the site is not typical of one. The binding pocket is shallow, and binds to a peptide rather than a small ligand or drug-like molecule. Still, such interactions involved in transcriptional regulation have been the recent focus of ligand-based drug discovery.[10]

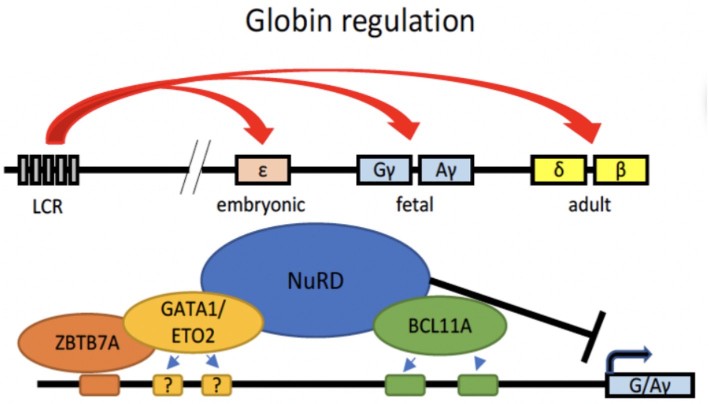

Figure 1. Visualization of interaction between NuRD complex and ETO2 protein, as well as NuRD recruitment to globin regulation gene for silencing of HbF.[2]

Drug Discovery: Traditional Methods

There has been an increase in focus on epigenetic targets such as the one described to influence HbF regulation above in the field of drug discovery and development.[10] Drug discovery is a historically lengthy and costly process, typically spanning decades for a single drug.[11] The process typically begins with basic research to identify and select a specific protein or pathway involved in a condition of interest, followed by lead discovery, during which systematic or exploratory searches are conducted to identify small drug-like compounds that are believed to interact with the protein or pathway of interest with reasonable specificity (*figure 2*). While these preliminary stages of drug discovery themselves can last years, the process following a successful lead discovery is still a lengthy one, involving further structural modifications, preclinical development, clinical development, and FDA approval.[11]

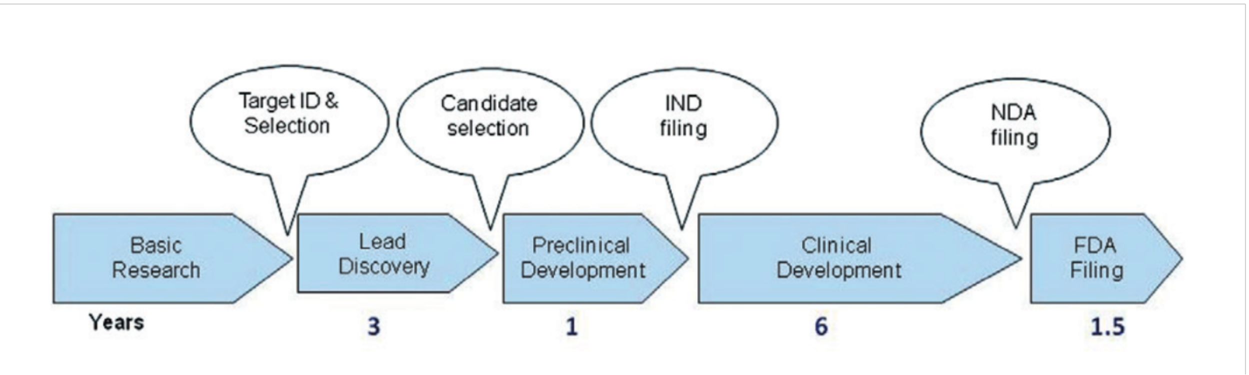

Figure 2. *Figure from Hughes et. al. 2011.*[11] Traditional pathway of drug discovery, spanning years beginning from basic science research for target identification, all the way to clinical development and FDA filing. *This study focuses on the lead discovery and candidate identification part of the pathway*.

This study focuses on the lead discovery stage of the process, attempting to use newer technology to improve upon some of the constraints and limitations often faced during this stage of drug discovery. Traditionally, lead discovery involves iterations of high throughput screening of existing chemical libraries, structure-based filtering, systematic compound design and synthesis, and *in vitro* and *ex vivo* mechanistic assays.[11] The high throughput screening approach does not typically require as much knowledge about the type of lead that may interact with a protein target. However, when conducted physically in a lab using a complex assay system, there are substantial cost and resource limitations on the number of libraries that can reasonably be screened.[11,12]

Several lead discovery techniques exist to supplement and guide high throughput screening to reduce this cost and resource strain. These methods may include but are not limited to: focused screens, in which previously identified successful compounds or compound classes influence library selection; structural aided design, in which crystal structures of the protein target with docked compounds are used to strategically introduce modifications; nuclear magnetic resonance (NMR) screens in which smaller fragments to be used as building blocks are prepared with protein targets with known NMR structures to search for binding activity.[11,12] The drawback with each of these techniques is that either a considerable amount of information must be known about the protein target and its potential binders, or a large number of physical resources may be utilized on a dead-end lead search.[11,12]

Drug Discovery: A Shift to *In-silico* Methods

With costs of traditional drug discovery and rapid technological advances, *in-silico* or computer-aided drug discovery methods have become an attractive alternative for lead discovery to many of the methods described above.[13,14] Structure-based *in-silico* drug discovery is made possible in part by improvements in 2D and 3D molecular digital representations. To begin, structural information regarding a target protein is collected, typically using nuclear magnetic resonance (NMR) or X-ray crystallography.[15] This information is then coded into 3D representations of the protein, built around data collected by NMR or x-ray crystallography regarding individual molecular interactions. In some rare cases, predictions regarding the protein's structure can be made entirely *in-silico* as well, but it is currently seen as ideal to begin with physical confirmation of this first stage.[15] One such *in-silico* 3D representation of this study's protein target created using NMR data can be seen in figure 3.

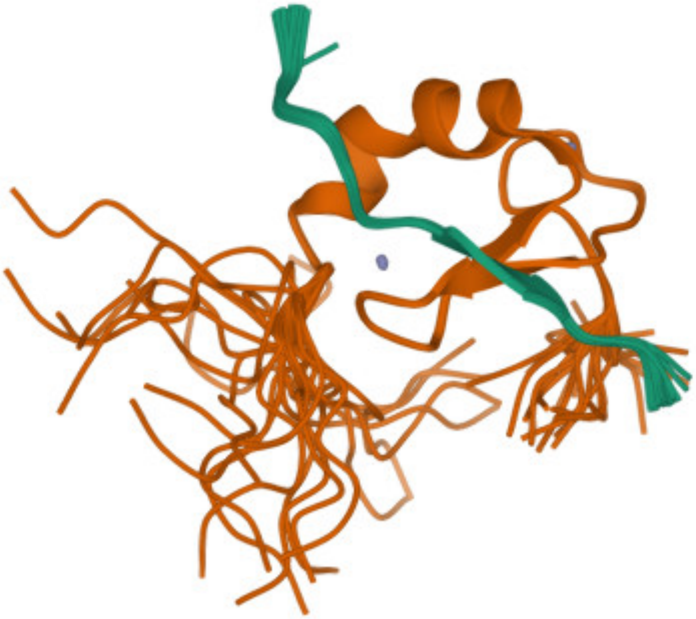

Figure 3. *From Liu et. al. 2007;[16]* 3D structure of the MYND domain (represented in orange), with its native peptide binder represented in green. Experimental data to determine the structure was collected using NMR.

Once a digital representation of the protein target has been created, virtual ligands must be selected for docking studies, to predict interactions between these ligands and the target protein. Virtual representations of ligands for docking studies are largely available for public use, with molecular information typically stored in Spatial Data Files (SDF) format or Simplified Molecular Input Line Entry System (SMILES) format. The chemical space of virtual representations of ligands and molecules to be tested in these docking studies has grown at an unprecedented rate, making *in-silico* drug discovery that much more appealing.[17] Libraries of upwards of 3 million compounds have existed since the 1990s, with libraries today pushing beyond billions of compounds–the most recent expansion is the new "eXplore" trillion-sized chemical space.[17,18]

The ZINC compound library is one such example of publicly accessible chemical space, frequently used in docking studies.[19] This database provides an enumeration of over 14 billion commercially available chemical compounds, many of which belong to make-on-demand libraries. Of these 14 billion compounds, over 230 million have calculated 3D structures, ready for docking. These compounds can be further filtered and explored based on chemical characteristics and purchasability.[19]

Once structural information on the protein target has been digitized and a chemical library has been selected, *in-silico* docking studies can be conducted to investigate the predicted relationship and binding affinity between these compounds and the protein target. Docking relies on information on the protein target's binding sites, or pockets of the protein for ligands to bind and produce a desired output.[15] The binding site may be known from physical experiments–such as X-ray crystallography of the protein target with a co-crystalized ligand–but may also be predicted by many docking softwares.[15] *In-silico* docking uses algorithms to predict the ideal orientation of each ligand within the hypothesized or known binding sites. This predicted orientation is then coupled with calculations of electrostatic and van der Waals interactions to either produce a predicted binding affinity or some type of ranked docking score.[15] Different docking software use different algorithms and therefore have different recommended thresholds for what score represents successful binding for the desired effects, or a "hit."

While molecular docking outputs are not typically seen as a final outcome of the binding relationship between a ligand and protein target, the *in-silico* docking method allows for a preliminary exploration of an unprecedented volume of chemical space at a minimal cost.[15,17] Outputs of a docking study can be used to inform further decisions in a drug discovery process and drastically narrow down compounds for physical screening.

Over the last several decades, *in-silico* docking has been integrated heavily into the drug discovery process, and has led to the successful discovery of drugs now used to treat conditions including diabetes, cancer, and viral and bacterial infections.[15] Some specific examples which are still used today are Lopinavir/Kaletra and Ritonavir/Norvir, both HIV protease inhibitors used to treat HIV/AIDS, and Sorafenib (Nexavar), a vascular endothelial growth factor (VEG-F) receptor kinase inhibitor used to treat several types of cancers.[15]

Drug Discovery: Introduction of Machine Learning and Quantitative Structure-Activity Relationship Modeling

In-silico drug discovery begins to address many of the cost and efficiency concerns that emerge from traditional drug discovery methods, but the size explosion of chemical space introduces a new limitation onto to the field—the computational inability or infeasibility to screen the entirety (or even a fraction) of the growing chemical space. As a method of addressing this problem, a specific form of regression and machine learning has gained popularity in the drug discovery field.[14,15,20]

This method, Quantitative Structure−Activity Relationship (QSAR), is essentially a form of regression or classification to model the relationship between the structure and activity of compounds.[20] Given information about predicted or existing binders of a target protein, QSAR uses mathematical models and information extracted about various molecular descriptors to predict biological activities, toxicity profiling, binding energies, and kinetic rates.[20] In using mathematical modeling to predict these various biological characteristics, QSAR is able to mimic many predictions made by direct molecular docking studies, at a fraction of the computational power.[14,15,20]

As QSAR models become increasingly integral to the field of drug discovery, the elegance and sophistication of QSAR modeling improve as well.[20] Initial QSAR models–the first developed in the 1960s–were relatively crude, and had low predictive power and generalizability.[20] The increase in data on chemical space and related biological activities has served to improve QSAR drastically. Still, sophisticated QSAR methods required substantial machine learning and chemical knowledge–even beyond the scope of many researchers exploring drug discovery.[20,21]

This was seen in the research community as a need for automated QSAR modeling. Several automated QSAR models have now been developed, including QSAR Workbench, AstraZeneca's AutoQSAR, and Schrödinger's AutoQSAR.[21] AutoQSAR methods take in a learning or training set of compounds associated with particular values for a dependent variable (some property of interest, relating to activity). Then via an automated process, hundreds of descriptors and fingerprints relating to a compound's structure are computed to be used as independent variables in model building.[21] Models are automatically built, validated, and refined using these calculated descriptors, until a final model is produced.[21] This final QSAR model is then applied in a predictive manner to a much wider array of compounds—structural information about the compounds are used to calculate descriptors like those used as independent variables in the model building stage. This information is finally used to generate predictions on the activity or dependent variable of choice.[21]

Summary of the Drug Discovery Pathway and Relevance

The drug discovery pathway has continued to become increasingly complex and sophisticated as the scope for what is chemically and technologically possible increases. Improved sophistication and automation of *in-silico* drug discovery could significantly reduce the time and resource investment required by drug discovery, and make drug discovery more accessible to academic settings as opposed to being exclusive to large pharmaceutical companies.

Purpose and Research Question

This project aimed to investigate and assess the applicability of Schrödinger's AutoQSAR software to the exploration and discovery of novel therapeutics for blood disorders such as sickle cell anemia. This is a unique target and application of the software because: (1) the target protein's structure was calculated using NMR data rather than X-ray crystallography with a co-crystallized ligand; (2) The learning or training set consisted of theoretical binders identified using a preliminary *in-silico* docking study as opposed to known drug-like binders for the binding site; (3) The identified binding pocket is a non-traditional pocket, characteristically having a more shallow binding groove than typical, and natively binding to peptides rather than drug-like compounds. The successful use of automated QSAR methods this early in the drug discovery process would move even more of the drug discovery process to *in-silico* methods and remove additional cost, resource, and time burdens of physically validating preliminary leads or binders, making the drug discovery process even more efficient and accessible.

Methods

Target Protein Preparation

The NMR structure of the target protein, the MYND domain, was prepared for docking using Schrödinger's Protein Preparation Wizard software.[22,23] The Protein Preparation Wizard prepares PDB files of target proteins for successful docking by ensuring structural correctness, incorporating missing hydrogen atoms, fixing ionization states, determining optimal histidine protonation states, and making any other relevant adjustments to the initial file.[22]

The MYND domain with a truncated structure of its known binder bound in the binding pocket of interest in the form of a PDB file was used as the initial structure for the protein preparation. The protein and ligand then went through the automated process of pre-processing, refinement, hydrogen optimization, and production of a final minimized structure.

Following this, a grid was generated to be used as the site for virtual docking. The minimized structure of the MYND domain and ligand were used to identify the intended grid, selected based on the binding pocket identified by the known binder. The intended size of the grid for docking was set to 15Å, to create a window wide enough for testing a variety of ligand types and conformations, but narrow enough to still be computationally efficient.

Training Set Generation: Preliminary Docking Study

A preliminary docking study was conducted on the prepared protein domain using Schrödinger's Virtual Screen Workflow. The Virtual Screening Workflow involves ligand library preparation using the LigPrep Software, followed by a series of docking calculations.[24];[25–27] The docking calculations occur in stages from most general to most specific, with only the top percentage of scorers from each stage advancing on to the next. These stages involve the Glide HTVS, Glide SP, and Glide XP software, in order from most general (and least computationally expensive) to most specific (and most computationally expensive) calculations.

The National Cancer Institute's Developmental Therapeutics Program (NCI DTP) provides a library of 265,242 structures in SDF format, which was one source used to construct the ligand library for the preliminary screen.[28] The NCI DTP library was selected as it is a free, publicly accessible library of compounds curated for the intended purpose of cancer therapeutic development, and was predicted to be largely made up of drug-like compounds that could interact with targets similar to the protein target in this study.

The other set of compounds selected to complete the ligand library for preliminary screening was the Enamine Hit Locator Library of 460,160 compounds.[29] Enamine constructed this library to be representative of its screening library of over 3.9M compounds. The inclusion of this library in preliminary screening offered the opportunity to evaluate ligand types different from those in the NCI library. Additionally, the ZINC chemical space was the ultimate intended target of exploration for this drug discovery study, and the largest contributor to this library is Enamine.[30] This makes the Enamine Hit Locator Library advantageous to include in the preliminary docking analysis–inclusion would make the preliminary docking data (and thus the QSAR models) more representative of libraries to be screened in the future.

Once these libraries were compiled, they were prepared using the Schrödinger LigPrep software.[24] The LigPrep software generates energy minimized 3D structures from 2D inputs of compounds, such as the SDF inputs used from NCI and Enamine. LigPrep was selected as the method of ligand preparation as opposed to other 2D to 3D ligand converters since LigPrep has built in checks to ensure chemical correctness and energy minimization beyond a typical one-to-one conversion, and also generates outputs specifically compatible with downstream Schrödinger software, such as the Glide docking software to be used later in the study.

Prepared ligands were then docked on the prepared protein target using Schrödinger Glide.[25–27] Docking scores were calculated for the compounds as followed: docking scores were calculated for all compounds using Glide HTVS; Glide SP scores were then calculated for the top 25% of Glide HTVS scorers; Glide XP scores were calculated for the top 10% of Glide HTVS scorers; finally, docking scores were outputted for the top 10% of Glide XP scorers.

Glide docking scores at each level are calculated using scoring functions that take into account shape and other properties of the receptor (target protein) and ligand. An exhaustive list of ligand torsions are generated to examine various docking conformations and poses at different areas of the established grid, or binding site, on the target protein.[27] The OPLS34 and OPLS2005 force fields are used to refine and minimize ligand conformations within the grid. Finally, docking scores are calculated by factoring in electrostatic and van der Waals interactions between the posed and minimized ligands and the target protein's binding site.

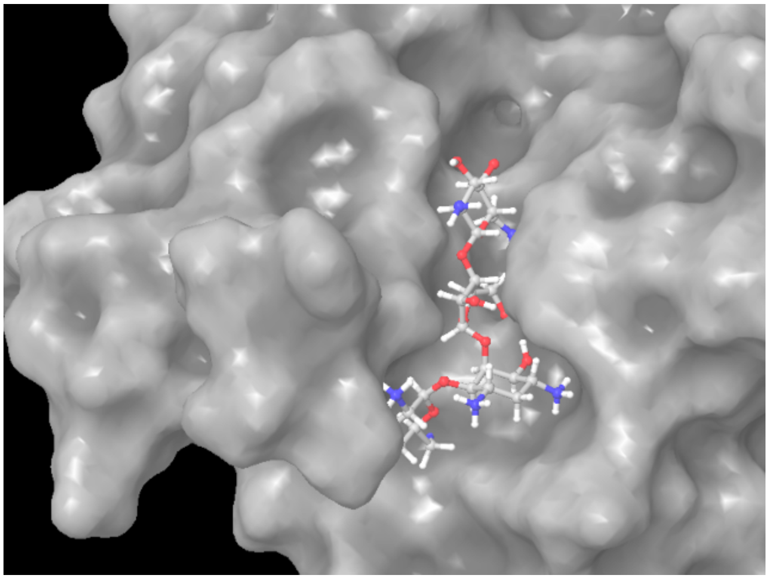

Figure 4. Example of a ligand (seen in a colored ball-and-stick representation), docked into the binding pocket of a protein target using Schrödinger Glide Docking.[25–27]

The specificities of the docking algorithm vary by level. While Glide HTVS and SP docking both use the same scoring function, HTVS scoring samples fewer intermediate conformations and has a less thorough refinement process, reducing computational power required.[27] Glide XP docking involves much more extensive sampling compared to Glide SP docking. The XP algorithm begins with SP sampling, but then follows an anchor-and-grow algorithm, in which part of the ligand is anchored and the remaining ligand is assembled into varying conformations from this point.[26] The XP scoring function is also more demanding in terms of ligand-receptor shape complementarity, and is the most thorough of the 3 in terms of minimization, intended to weed out false positives allowed by SP docking.[26]

At this time, a threshold docking score was also decided as the cutoff for what would be considered a “hit” from *in-silico* docking. The Glide software does not have any inherent value for what should be considered a good binder or a hit as this varies from system to system, but Schrödinger rather encourages the selection of a threshold depending on project goals, preliminary docking information, and any information available about known binders. [26,27]

QSAR Model Building and Validation

The preliminary data generated from *in-silico* docking was used as a training set to generate QSAR models using the Schrödinger AutoQSAR software.[21,31] AutoQSAR consolidates into one workflow the processes of descriptor generator and feature selection from given ligands, creation of QSAR models based on these characteristics, and the validation and selection of top QSAR models.[21,31] Figure 5 demonstrates an overview of the AutoQSAR workflow.

AutoQSAR uses Schrödinger’s Canvas cheminformatics package to generate topology-based descriptors–physicochemical properties, graph-theoretical indices and functional group counts–of the provided training set of ligands.[21,32] Descriptors are then filtered by variance and redundancy relative to the entire training set of ligands, clustered, and reduced using absolute Pearson correlation matrices.[21,32] Canvas is also used to generate fingerprints encoded into an addressable bit space of 2^{32} for each ligand in the training set, to be used as independent variables.

AutoQSAR then constructs multiple linear regression (MLR) and kernel-based partial least squares (KPLS) models using the determined descriptors and fingerprints as independent variables, and a given characteristic of the ligands as dependent variables.[21,31] In this study, docking score was this dependent variable–docking scores from the preliminary docking study were used to create the QSAR models, and the QSAR models were used to predict docking scores of other compounds. Docking scores are a continuous outcome variable, but can be later dichotomized into the binary classifiers of “hits” versus “misses” based on the predetermined threshold. While some results are presented in terms of binary outcomes, all statistical analysis was conducted on the original continuous variables to avoid bias.

AutoQSAR MLR models are arithmetic averages of the top five linear least squares models, identified via 1000 steps of Monte Carlo simulated annealing. The number of variables is limited to 10 + 10·log(N) of the training set.[21,31] AutoQSAR KPLS models are constructed from latent variables–mutually orthogonal vectors combining independent variables with different weights.[21] To prevent overfitting, factors are no longer added before the coefficient of determination (R^2) exceeds 0.9.[21]

As these described processes are all automated in the AutoQSAR workflow, the only manual determinations made were the selection of the training set (previously described) and how this training set would be split into subsets for model training, model testing or internal validation, and holdout or external validation. The optimal ratio for splitting data for machine learning is 64:16:20, with 64% of the data used for training, 16% for testing, and 20% for external validation.[33]

As a way to further evaluate the performance of AutoQSAR models and select the best suited models, 3 different sets of models were constructed using 3 different ratios for splitting data. The first, Model A, was created using the optimal 64:16:20 ratio. A second model, Model B, was created using a 70:15:15 ratio; and a final model, Model C, was created using a 60:20:20 ratio. Given a training set and a predetermined data splitting ratio, AutoQSAR determines which compounds from the training set will be included in which subset (training, testing, or holdout) based on a random seed.[21]

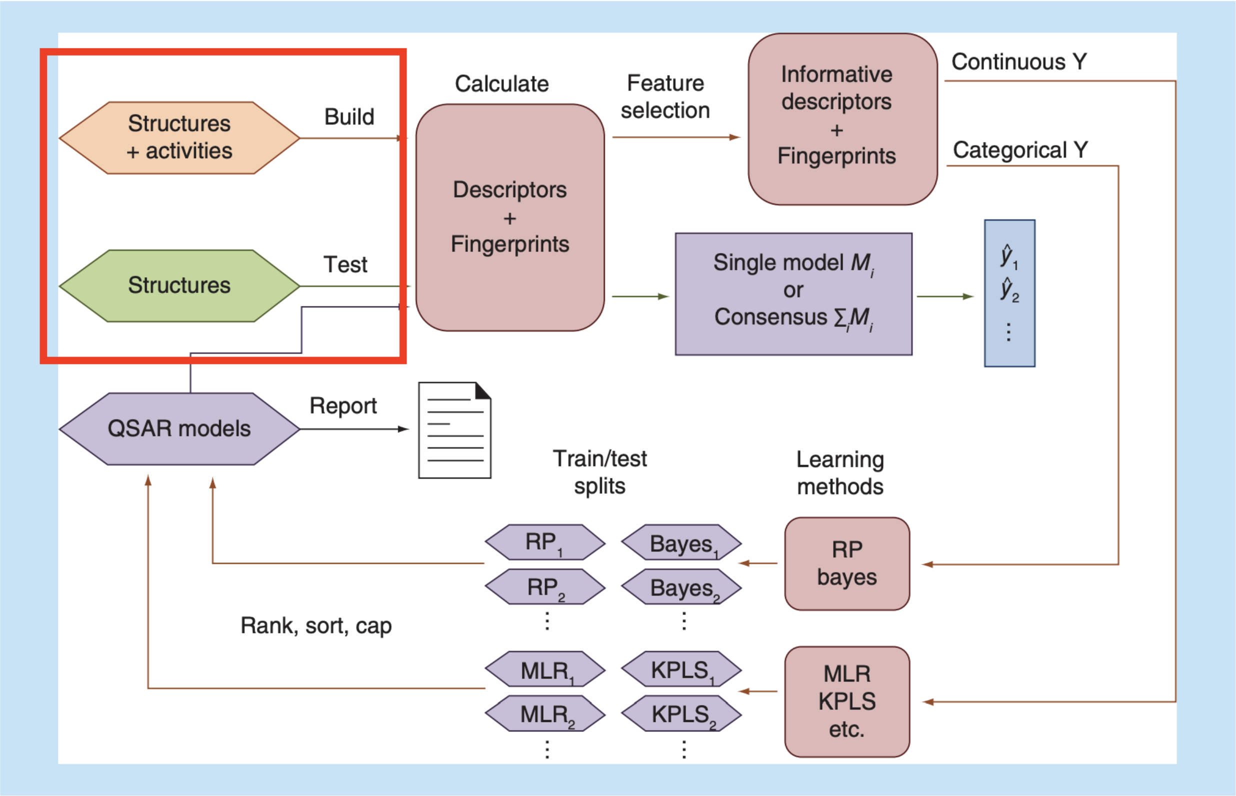

Figure 5. *Adapted from Dixon et. al. 2011.*[21] AutoQSAR workflow overview; Red box added to indicate which components of the workflow are manually determined by the researcher (selection of training set, and division of training set into training, testing, and validation components).

QSAR Model Comparison and Selection

Once the 3 models were built, they were compared on the basis of model predictive ability from an external evaluation set (Q^2), root-mean-square error over the external test set (RMSE), area under the curve of a receiver operating characteristic curve (ROC-AUC score), and area under the curve of a precision-recall curve (PR-AUC score) to determine which model would be used for generating predictions. All statistical analysis was conducted using RStudio.[34]

Comparison of Q^2 values provides information regarding the extent to which data can be predicted by each model, with the most favorable model being the one with the highest Q^2 value. For Schrodinger AutoQSAR, Q^2 is calculated as follows[21]:

$$Q^2 = \frac{variance\ in\ prediction\ errors\ over\ test\ set}{variance\ in\ test{-}set\ dependent\ variables} \quad (1)$$

Comparison of RMSE values gives information on the typical difference between predictions made by the model and the true values (the values provided in the training set) in the same units as the dependent variable (docking scores). The most favorable model is the one with the lowest RMSE. Direct comparisons of Q^2 and RMSE values were made across models. A one-way ANOVA was conducted for both Q^2 and RMSE, to evaluate if there was any statistically significant difference between the models as well.

AUC scores from ROC and PR curves were also used to compare models. ROC curves represent the tradeoff between specificity and sensitivity of a model, while PR curves represent the tradeoff between precision and sensitivity. AUC scores represent the probability that a randomly selected positive instance is ranked higher than a randomly selected negative instance. An AUC value close to 1 indicates a high predictive power/high discrimination between negative and positive cases, while an AUC value below 0.5 indicates that the model's predictions are essentially due to random chance.

ROC curves visualize how correctly classified positive cases vary with incorrectly classified negative cases. Graphically, 1 – specificity (or false positive rate) is plotted against sensitivity (or the true positive rate). The calculations for these values are as follows[35]:

$$true\ pos.rate\ =\ sensitivity\ =\ \frac{true\ positives}{true\ positives + false\ negatives} \quad (2)$$

$$false\ pos.rate\ =\ 1\ -specificity\ =\ 1 - \frac{false\ negatives}{false\ negatives + true\ positives} \quad (3)$$

ROC curves are most commonly used for representing, evaluating, and comparing machine learning models.[36] However, PR curves give better information when outcomes are not balanced–that is, the positive class is a rare outcome.[36] PR curves plot precision against recall, with the calculations for these values as follows:[37]

$$precision\ = \frac{true\ positives}{true\ positives + false\ positives} \quad (4)$$

$$recall\ (= sensitivity)\ = \frac{true\ positives}{true\ positives + false\ negatives} \quad (5)$$

While it was expected that the positive class of "hits" would be a rare outcome among the training data, both ROC and PR curves were created and evaluated. Construction of ROC and PR curves require a binary outcome variable. In this study, the outcome variable of docking scores itself was continuous, but as described earlier, the ultimate goal in calculating docking scores was to determine which compounds should be considered "hits" and which ones should not. Thus, outcomes were dichotomized into binary variables based on the predetermined threshold of docking scores–ligands with docking scores more negative than the threshold were considered "hits" and assigned a value of 1 to be placed in the positive outcome class. The remaining ligands were assigned a value of 0, and placed in the negative outcome class.

Additionally, a new variable to represent docking scores was created, scaling the original value by (-1). Glide docking scores are represented as negative numbers, with values that are more negative representing stronger binding. However, calculations for ROC and PR curves assume greater values of the outcome variable to be associated with the positive class. Following this data cleaning and preparation, ROC and PR calculations (and calculations of associated AUC scores) were made using R software packages plotROC and ROCR respectively, and plotted using ggplot2.[35,37,38] Finally a one-way ANOVA was also conducted for the ROC-AUC scores across the 3 sets of models, and the PR-AUC scores across the 3 sets of models, to determine if there was a statistically significant difference across the models.

Finally, a one-way ANOVA was conducted for each measure across all models at a significance level of 0.05, to determine if the differences between models was statistically significant. Following the results of the comparisons and statistical analysis, one data splitting ratio would be selected for remaining analysis. Priority was given to factors with statistically significant differences across models in deciding which model setup was the most favorable.

Additional Verification of Models: Inclusion Dependency

Once a data splitting ratio was selected, a second round of model building was conducted to test the robustness of the models against the inclusion or exclusion of certain ligands from the training set. 5 models were built, all using the selected data splitting method, varying only the random seed used to allocate specific ligands to training, testing, and external validation sets.

Following the construction of 5 such models, a one-way ANOVA was conducted on the Q^2 values, to test for a statistically significant difference in the models dependent on ligand inclusion/exclusion. A statistically significant difference would indicate that the models change significantly based on small changes in the inclusion of certain ligands, indicating a weak model.

Sample Predictions

Next, sample docking score predictions were generated using the selected and verified QSAR model for the initial ligand library used to build the model (the combination of the NCI DTP compound set and Enamine Hit Locator Library). These predictions were aimed to be a final validation of the model's predictive ability, aimed specifically at the ability for the model to accurately predict information about 2D ligands, since this would be the structure of any larger-scale screening performed with the selected QSAR models. For this reason, predictions were generated using 2D structures of the ligand library, rather than 3D structures.

The predictions were analyzed using RMSE and capture of top scorers. RMSE was calculated over the difference between predicted scores and scores from the preliminary docking study, to quantify the average difference between predicted and "true" values. RMSE values were calculated overall, as well as by outcome classification in the preliminary study (separately ligands classified as positive, or "hits," and for ligands classified as negative, or "misses").

Capture of top scorers refers to the percentage of ligands that were classified as hits in the preliminary study that would also be classified as hits based on predictive values. This serves as a practicality check on the models–even if a small degree of error exists between the specific docking scores outputted by the model, the model would still be practically effective if it captures a sufficient number of hits. Similar QSAR models and applications of AutoQSAR have achieved 90-95% capture of top scorers, which will serve as a standard for comparison.[21,39]

Excluded methods: Larger-Scale Predictions

Had the previous methods been successful, the next steps would have been to generate docking score predictions on larger-scale libraries, to use the finalized QSAR models to explore a wider segment of chemical space. The library of choice would have been the ZINC20 library, offering 230 million ready-to-dock ligand structures.[19] Predictions would have initially been generated for the 13 million compound subset of ready-to-purchase compounds, to enable easy purchase and *in-vitro* verification of predicted hits. Depending on success, the screen could be scaled up to include a greater portion of the ready-to-synthesize library.

Results

The following results are from the 4 distinct analyses conducted throughout the study methods: (1) preliminary docking study (2) analysis of varying data splitting ratios in QSAR model building; (3) analysis of overfitting and inclusion/exclusion dependency; (4) test predictions.

Preliminary Docking Study

The preliminary docking study resulted in the outputted docking scores for the top scoring compounds among the NCI DTP and Enamine Hit Locator Libraries. The calculated docking scores range from -5.2931 to -13.569. Based on the preliminary docking data, a threshold of -10 was selected as the cutoff for outcomes that were considered "hits"–if a ligand had a docking score ≤ -10, it was included in the positive outcome of hits; otherwise, it was included in the negative outcome class of misses. Based on this threshold, only 4.1% of compounds included in the preliminary dataset are counted as hits.

Data Splitting Ratios

3 sets of QSAR models were constructed using 3 different data splitting ratios, to determine if the optimal 64:16:20 data splitting ratio for machine learning held true for AutoQSAR models as well. Table 1 includes a summary of measures used to compare and assess the predictive ability of the models. Table 2 contains the results of one-way ANOVAs for each of these measures. At a predetermined significance level of 0.05, only the measures of Q^2 and RMSE had statistically significant differences across models, and were thus given priority in selecting the most favorable model.

The AUC scores of the PR models were ignored in determining the most favorable model, since scores below 0.5 indicate that there is less than a 50% chance that a randomly selected positive instance is ranked higher than a randomly selected negative instance, and model predictions are due to random chance. Because all models had scores below 0.5, this served as an initial indicator of the poor predictive power of the models.

Based on these considerations, While Model B had the smallest RMSE value, Model A had a Q^2 value closest to 1.0 and an ROC-AUC score closest to 1.0 as well. Model A, the model built using a 64:16:20 data splitting ratio, was selected for further analysis.

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

	Q^2	RMSE	ROC-AUC	PR-AUC
Model A (64:16:20)	0.87106	0.59421	0.92435	0.38665
Model B (70:15:15)	0.87029	0.58852	0.92219	0.40152
Model C (60:20:20)	0.85689	0.61476	0.91984	0.39659

Table 2. One-way ANOVA results for comparison metrics across 3 different data splitting ratios.

		Df	Sum Sq	Mean Sq	F value	Pr (>F)
Q^2	Model	2	0.001270	0.0006349	16.63	1.96e-05
	Residuals	27	0.001031	0.0000382		
RMSE	Model	2	0.003811	0.0019054	9.097	0.000955
	Residuals	27	0.005655	0.0002094		
ROC-AUC	Model	2	0.0001018	5.088e-05	1.755	0.192
	Residuals	27	0.0007828	2.899e-05		
PR-AUC	Model	2	0.001147	0.0005737	1.347	0.277
	Residuals	27	0.011503	0.0004261		

The ROC curves and PR curves also provide valuable information regarding the models. The ROC curves for all 3 model setups are shown in figure 6. The curves appear to be closely approaching the point of a perfect test (100% sensitivity and 100% specificity) and are graphically far from the line of random performance, with AUC scores close to 1.0, indicating strong predictive power of the models.

However, given that the outcomes in the study were imbalanced, with positive class outcomes being rare, PR curves should be given precedence as an indicator of model predictive ability. Figure 7 shows the PR curves for each of the 3 models. The curves appear far from the point of a perfect test and *below* the line of random chance, with AUC scores for each version of each model falling below 0.5. Not only does this indicate that the model has a poor predictive ability, but also is indicative of potential overfitting due to inflated Q^2 values.

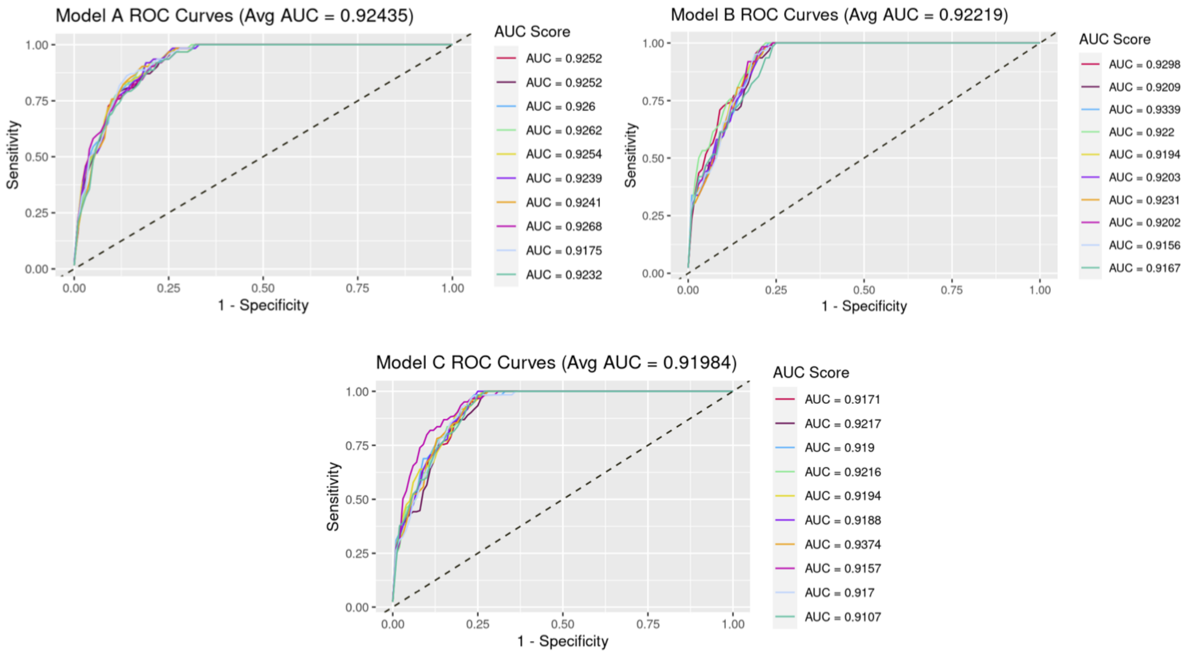

Figure 6. ROC Curves for each set of QSAR models, trained using different data splitting ratios. Black dotted line indicates the line of random chance. Color-coded ROC curves are associated with particular AUC scores shown in graph legends.

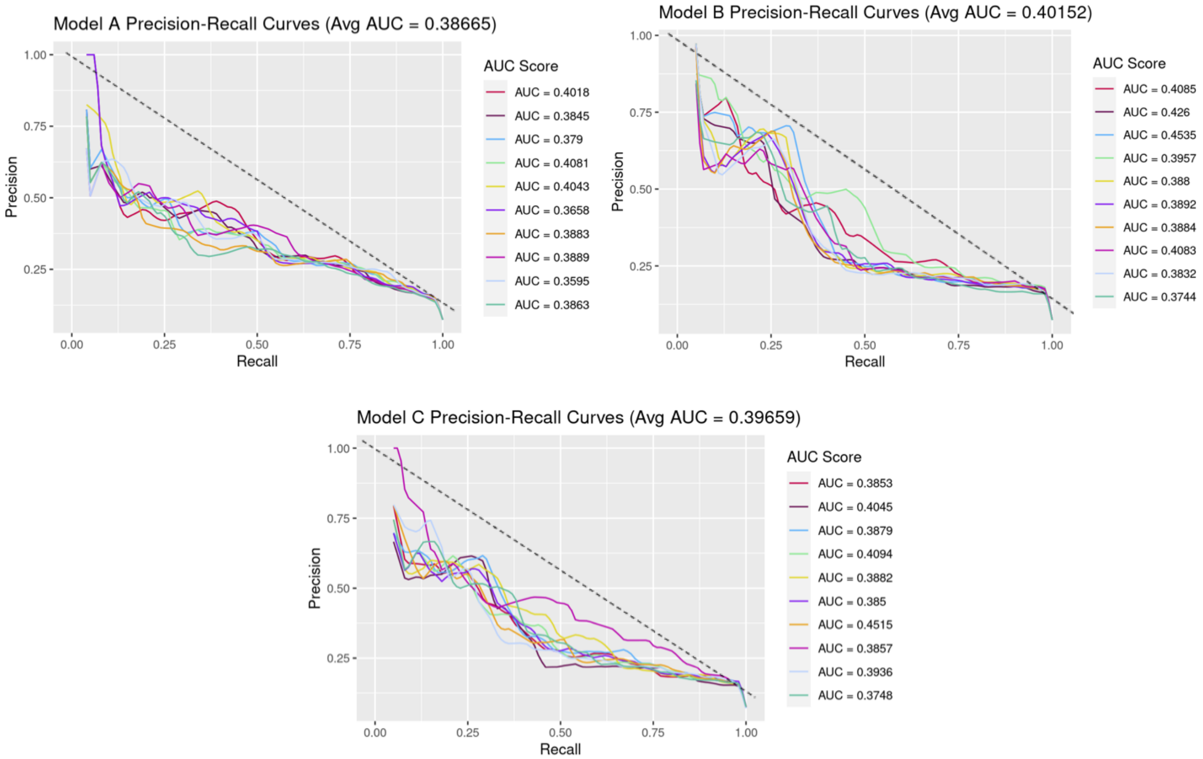

Figure 7. PR curves for each set of QSAR models, trained using different data splitting ratios. Black dotted line indicates the line of random chance. Color-coded ROC curves are associated with particular AUC scores shown in graph legends.

Overfitting and Inclusion/Exclusion Dependency

Following the selection of the 64:16:20 data splitting ratio, 5 QSAR models were then built using this same ratio, with 5 different random seeds used to categorize ligands from the training set into training, test, and external validation subsets. Table 3 includes the results of a one-way ANOVA on the Q^2 values for these models. At a significance level of 0.05, there is a significance difference between the predictive powers of each model. This demonstrates that the models change significantly depending on small changes in the inclusion or exclusion of certain ligands. This is characteristic of model overfitting, and further reduces the validity of the generated QSAR models and their predictions.

Table 3. One-way ANOVA results of Q2 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		

Test Predictions

Although the previous results already present two indicators of poor predictability from the QSAR models, the third and final planned analysis was carried out as confirmation. Test predictions were generated using 2D structures of the same ligands used to build the selected QSAR model. Table 4 contains a summary of relevant results from these predictions. While the overall RMSE value is deceptively low (an average error of 0.6 for a metric ranging from roughly -5 to -13), RMSE values separated by positive and negative class outcomes (hits and misses) paint a clearer picture. The RMSE was 1.4 for compounds that should have been classified as hits at a predetermined threshold of -10, while RMSE was only 0.5 for compounds that should have been classified as misses. This demonstrates that the model has difficulty discerning positive cases, and has a greater error in predicting hits.

Further, the percent of hits captured in the predictive screen was only 64%, meaning only 64% of compounds labeled as hits based on the preliminary docking study were identified as hits based on the QSAR predictions. Compared to other models cited in industry use which achieve 90-95% capture of top scorers, this provides further evidence that the generated QSAR models would not serve a practical predictive purpose.[21,39]

Table 4. Measures of accuracy of sample prediction; RMSE values of predicted docking scores versus preliminary docking scores using QSAR Model A, and percent capture of top-scorers

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

Discussion

Ultimately, AutoQSAR modeling for this particular protein system was unsuccessful at this time. The low AUC-PR scores, susceptibility of the models to change on the basis of randomized ligand inclusion or exclusion, high error upon external verification in predicted docking scores for compounds that should have been labeled hits, and low ability to capture top scorers all indicate a low predictive ability and low practical application of the models.

Several factors may have factored into the failure of these specific AutoQSAR model, many of which do not directly relate to the inapplicability of automated QSAR, but rather other factors throughout the study. First, the starting structure of the target protein (MYND domain of ETO2), was obtained from NMR data as opposed to crystallography with a co-crystallized ligand. While NMR data is an acceptable form of structural information for target proteins under Schrödinger’s protocol, crystallography based structures are more commonly seen in literature on *in-silico* docking. The use of an NMR structure may have introduced a degree of uncertainty in foundational information regarding the binding pocket.

Additionally, the binding pocket was non-traditional when compared to binding pockets on other drug targets. The binding pocket of MYND is more shallow than most, and natively binds to a peptide. This study was specifically exploring drug-like compounds as binders for future pharmaceutical or therapeutic use, but the binding pocket may not have been best suited for these interactions.

Likely even more significantly, no drug-like binders were known or verified at any point during the study. This meant that the training set used to build the models may have been flawed, and compounds that were listed as hits may not have been true binders of the MYND domain. Between the non-traditional binding pocket and lack of verification of preliminary docking data, it is possible that scores from the preliminary docking study were falsely driven by excessive weight given to factors such as pi-stacking or hydrogen, which would not have been present to the same degree in an *in-vitro* interaction of the same ligand and target. This would further interfere with the AutoQSAR calculated descriptors and patterns found in these descriptors, since many ligands included in the model may not have truly been hits and should not be contributing to this pattern. Finally, the high variability and susceptibility of the models to change depending on the inclusion and exclusion of certain ligands serves as further confirmation that the descriptor patterns found across ligands were not very robust.

The use of fully automated QSAR modeling may have also imposed further limitations on the study. While it can be argued that machine learning/artificial intelligence-based models and software have a better predictive power and are able to discern a greater number of the descriptors to be factored into a QSAR model, it is possible that structural chemists with greater field experience may have been able to discern more *relevant* information regarding the chemical descriptors that should have been factored into the model. Low predictability of docking scores for the same ligands used to build the model when using a 2D ligand format as opposed to the original 3D format the model was built from further highlights potential issues in AutoQSAR's descriptor calculations. It is possible that these calculated descriptors are overly reliant on characteristics of the 3D posed ligands, as opposed to inherent molecular characteristics.

This being said, there are several limitations to generalizing the results of this study. Drug discovery is so highly system-specific that the results and issues encountered in this study may be entirely different to those encountered by another study following an identical protocol with a new protein system. Several variables played a role in the outcomes of the study–the non-traditional binding pocket, the lack of existing drug-like binders, the use of an NMR structure, to name a few–and without independently testing changes in each of them, it is impossible to say which ones or if all played a role in the outcomes of the study.

Future studies may be conducted to continue the exploration of AutoQSAR with this protein system, with some modifications. Verification of hits from preliminary *in-silico* docking, using traditional methods such as NMR, may be useful prior to building QSAR models. This would allow the use of a smaller number of verified hits as opposed to a larger number of unverified hits, providing a stronger foundation for the model. The patterns picked up by the AutoQSAR descriptors may have a greater degree of accuracy and precision given this. Additionally, beginning with an x-ray crystallography structure of the target protein with a co-crystallized ligand may provide more robust information regarding the binding pocket, which may be especially essential given the shallow and non-traditional nature of the pocket.

While the results of this study did not produce a successful QSAR model for future use in generating docking predictions, it did reveal many limitations of the protein system that may need to be overcome before future *in-silico* experiments are conducted with this system. It also demonstrated some of the reasons that, while *in-silico* drug discovery may serve as a supplement to traditional methods, it is not yet advanced enough to become a complete replacement for traditional methods of drug discovery.

References

1. Townsend PA, Kozhevnikova MV, Cexus ONF, Zamyatnin AA, Soond SM. BH3-mimetics: recent developments in cancer therapy. *J Exp Clin Cancer Res*. 2021;40(1):355. doi:10.1186/s13046-021-02157-5

2. Guo X, Plank-Bazinet J, Krivega I, Dale RK, Dean A. Embryonic erythropoiesis and hemoglobin switching require transcriptional repressor ETO2 to modulate chromatin organization. *Nucleic Acids Res*. 2020;48(18):10226-10240. doi:10.1093/nar/gkaa736

3. 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

4. What is Sickle Cell Disease? | CDC. Accessed January 27, 2022. https://www.cdc.gov/ncbddd/sicklecell/facts.html

5. Kato GJ, Piel FB, Reid CD, et al. Sickle cell disease. *Nat Rev Dis Primers*. 2018;4:18010. doi:10.1038/nrdp.2018.10

6. How Common Is Sickle Cell Disease? Accessed January 27, 2022. https://sickle-cell.com/statistics

7. *The State of Sickle Cell Disease: 2016 Report*. The American Society of Hemotology; 2016:1-31.

8. Akinsheye I, Alsultan A, Solovieff N, et al. Fetal hemoglobin in sickle cell anemia. *Blood*. 2011;118(1):19-27. doi:10.1182/blood-2011-03-325258

9. Demirci S, Leonard A, Essawi K, Tisdale JF. CRISPR-Cas9 to induce fetal hemoglobin for the treatment of sickle cell disease. *Mol Ther Methods Clin Dev*. 2021;23:276-285. doi:10.1016/j.omtm.2021.09.010

10. Díaz-Eufracio BI, Naveja JJ, Medina-Franco JL. Protein-Protein Interaction Modulators for Epigenetic Therapies. *Adv Protein Chem Struct Biol*. 2018;110:65-84. doi:10.1016/bs.apcsb.2017.06.002

11. 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

12. Zhou S-F, Zhong W-Z. Drug design and discovery: principles and applications. *Molecules*. 2017;22(2). doi:10.3390/molecules22020279

13. Nazarova AL, Katritch V. It all clicks together: In silico drug discovery becoming mainstream. *Clin Transl Med*. 2022;12(4):e766. doi:10.1002/ctm2.766

14. Brogi S, Ramalho TC, Kuca K, Medina-Franco JL, Valko M. Editorial: In silico Methods for Drug Design and Discovery. *Front Chem*. 2020;8:612. doi:10.3389/fchem.2020.00612

15. Shaker B, Ahmad S, Lee J, Jung C, Na D. In silico methods and tools for drug discovery. *Comput Biol Med*. 2021;137:104851. doi:10.1016/j.compbiomed.2021.104851

16. 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

17. Bender BJ, Gahbauer S, Luttens A, et al. A practical guide to large-scale docking. *Nat Protoc*. 2021;16(10):4799-4832. doi:10.1038/s41596-021-00597-z

18. Neumann A, Marrison L, Klein R. Relevance of the Trillion-Sized Chemical Space "eXplore" as a Source for Drug Discovery. *ACS Med Chem Lett*. Published online March 16, 2023. doi:10.1021/acsmedchemlett.3c00021

19. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. ZINC: a free tool to discover chemistry for biology. *J Chem Inf Model*. 2012;52(7):1757-1768. doi:10.1021/ci3001277

20. Soares TA, Nunes-Alves A, Mazzolari A, Ruggiu F, Wei G-W, Merz K. The (Re)-Evolution of Quantitative Structure-Activity Relationship (QSAR) Studies Propelled by the Surge of Machine Learning Methods. *J Chem Inf Model*. 2022;62(22):5317-5320. doi:10.1021/acs.jcim.2c01422

21. 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

22. Sastry GM, Adzhigirey M, Day T, Annabhimoju R, Sherman W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J Comput Aided Mol Des*. 2013;27(3):221-234. doi:10.1007/s10822-013-9644-8

23. Schrödinger, LLC. *Schrödinger Release 2022-3: Protein Preparation Wizard*. Schrödinger, LLC; 2021.

24. Schrödinger, LLC. *Schrödinger Release 2022-3: LigPrep*. Schrödinger, LLC; 2021.

25. Schrödinger, LLC. *Schrödinger Release 2022-3: Glide*. Schrödinger, LLC; 2021.

26. Friesner RA, Murphy RB, Repasky MP, et al. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem*. 2006;49(21):6177-6196. doi:10.1021/jm051256o

27. Friesner RA, Banks JL, Murphy RB, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem*. 2004;47(7):1739-1749. doi:10.1021/jm0306430

28. NCI Database Download Page. Accessed April 2, 2023. https://cactus.nci.nih.gov/download/nci/

29. Hit Locator Library - Enamine. Accessed April 2, 2023. https://enamine.net/compound-libraries/diversity-libraries/hit-locator-library-200

30. Warr WA, Nicklaus MC, Nicolaou CA, Rarey M. Exploration of ultralarge compound collections for drug discovery. *J Chem Inf Model*. 2022;62(9):2021-2034. doi:10.1021/acs.jcim.2c00224

31. Schrödinger, LLC. *Schrödinger Release 2022-3: AutoQSAR*. Schrödinger, LLC; 2021.

32. Schrödinger, LLC. *Schrödinger Release 2022-3: Canvas*. Schrödinger, LLC; 2021.

33. Joseph VR. Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*. Published online April 4, 2022. doi:10.1002/sam.11583

34. RStudio Team (2020). *RStudio: Integrated Development for R*. RStudio, PBC; 2020.

35. Sachs MC. plotROC: A Tool for Plotting ROC Curves. *J Stat Softw*. 2017;79. doi:10.18637/jss.v079.c02

36. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*. ACM Press; 2006:233-240. doi:10.1145/1143844.1143874

37. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. *Bioinformatics*. 2005;21(20):3940-3941. doi:10.1093/bioinformatics/bti623

38. Wickham H. *Ggplot2: Elegant Graphics for Data Analysis (Use R!)*. 2nd ed. Springer; 2016:276.

39. Carracedo-Reboredo P, Liñares-Blanco J, Rodríguez-Fernández N, et al. A review on machine learning approaches and trends in drug discovery. *Comput Struct Biotechnol J*. 2021;19:4538-4558. doi:10.1016/j.csbj.2021.08.011