Developing a machine learning model to identify protein-protein interaction hotspots to facilitate drug discovery

Rohit Nandakumar [Corresp., 1], **Valentin Dinu** [1]

[1] Department of Biomedical Informatics, Arizona State University, Tempe, Arizona, United States

Corresponding Author: Rohit Nandakumar
Email address: rnandaku@asu.edu

Throughout the history of drug discovery, an enzymatic-based approach for identifying new drug molecules has been primarily utilized. Recently, protein-protein interfaces that can be disrupted to identify small molecules that could be viable targets for certain diseases, such as cancer and the human immunodeficiency virus, have been identified. Existing studies computationally identify hotspots on these interfaces, with most models attaining accuracies of ~70%. Many studies do not effectively integrate information relating to amino acid chains and other structural information relating to the complex. Herein, 1) a machine learning model has been created and 2) its ability to integrate multiple features, such as those associated with amino-acid chains, has been evaluated to enhance the ability to predict protein-protein interface hotspots. Virtual drug screening analysis of a set of hotspots determined on the EphB2-ephrinB2 complex has also been performed. The predictive capabilities of this model offer a precision-recall score of 0.605 and an AUROC of 0.846. Virtual screening of a set of hotspots identified by the machine learning model developed in this study has identified potential medications to treat diseases caused by the overexpression of the EphB2-ephrinB2 complex, including prostate, gastric, colorectal and melanoma cancers which are linked to EphB2 mutations. The efficacy of this model has been demonstrated through its successful ability to predict drug-disease associations previously identified in literature, including cimetidine, idarubicin, pralatrexate for these conditions. In addition, nadolol, a beta blocker, has also been identified in this study to bind to the EphB2-ephrinB2 complex, and the possibility of this drug treating multiple cancers is still relatively unexplored.

Developing a Machine Learning Model to Identify Protein-Protein Interaction Hotspots to Facilitate Drug Discovery

Rohit Nandakumar[1]; Valentin Dinu, PhD[1]

[1]Department of Biomedical Informatics, Arizona State University, Arizona, United States of America

Corresponding Author:

Rohit Nandakumar

Address: Biomedical Informatics - ASU

Mayo Clinic, SC. Johnson Research Bldg

13212 E Shea Boulevard

Scottsdale AZ 85259

Email Address: rnandaku@asu.edu

ABSTRACT

Throughout the history of drug discovery, an enzymatic-based approach for identifying new drug molecules has been primarily utilized. Recently, protein-protein interfaces that can be disrupted to identify small molecules that could be viable targets for certain diseases, such as cancer and the human immunodeficiency virus, have been identified. Existing studies computationally identify hotspots on these interfaces, with most models attaining accuracies of ~70%. Many studies do not effectively integrate information relating to amino acid chains and other structural information relating to the complex. Herein, 1) a machine learning model has been created and 2) its ability to integrate multiple features, such as those associated with amino-acid chains, has been evaluated to enhance the ability to predict protein-protein interface hotspots. Virtual drug screening analysis of a set of hotspots determined on the EphB2-ephrinB2 complex has also been performed. The predictive capabilities of this model offer a precision-recall score of 0.605 and an AUROC of 0.846. Virtual screening of a set of hotspots identified by the machine learning model developed in this study has identified potential medications to treat diseases caused by the overexpression of the EphB2-ephrinB2 complex, including prostate, gastric, colorectal and melanoma cancers which are linked to EphB2 mutations. The efficacy of this model has been demonstrated through its successful ability to predict drug-disease associations previously identified in literature, including cimetidine, idarubicin, pralatrexate for these conditions. In addition, nadolol, a beta blocker, has also been identified in this study to bind to the EphB2-ephrinB2 complex, and the possibility of this drug treating multiple cancers is still relatively unexplored.

INTRODUCTION

Drug discovery is the scientific process where new drugs and small molecules are developed and identified to treat certain conditions. Throughout most of the history of drug discovery, an enzymatic-based (lock and key) approach for identifying new drug molecules was utilized (Bakail & Ochsenbein, 2016). As a result, many drugs targeting G-protein coupled receptors (GPCRs), which interact via this approach, constitute about 34% of the drugs in the market today (Hauser et al., 2017).

Protein-protein interfaces have been of particular interest in regards to drug discovery, such as the EphA4-EphrinB2 complex, which is considered to be conformationally flexible (Ma & Nussinov, 2014). Protein-protein interfaces can be stabilized or disrupted to identify small molecules that could be viable targets for certain diseases such as cancer and the human immunodeficiency virus (HIV). Identifying residue hotspots on these protein-protein interfaces and repurposing existing drugs to target these new hotspots can lead to novel drug targets, ultimately leading to new therapeutic treatments (Scott et al., 2016). Although protein-based drug discovery (as opposed to enzymatic-based drug discovery) is a relatively new and emerging field, recent studies have shown promising results in regards to its potential in a wide range of fields from drug discovery to drug repositioning. For example, the SpotOn study has produced remarkable results in regards to identifying hotspots that are viable for drug discovery, and AnchorQuery, which identifies small molecule protein-interaction inhibitors. (Moreira et al., 2017; Koes, Dömling & Camacho, 2018)

In addition, PPI-based peptide drug discovery has been used to identify new therapeutic targets by disrupting PPIs. Major advances in docking simulations and models in recent years have yielded to be effective in more accurately identifying peptide-protein interactions. Although peptide-based PPI drug discovery does have its challenges, such as limited bioavailability and solubility of peptides, this emerging field highlights potentially exciting advances in computationally aided protein-protein interaction based discovery techniques with the use of interfering peptides. (Lee et al., 2019)

Currently, only 10-14% of the human proteome is considered to be “druggable”, and most targets with published leads are in the rhodopsin-like GPCR family, with a smaller number in cation channels and protein kinases (Hopkins & Groom, 2002; López-Cortés et al., 2019). Druggability is the ability for a drug to bind to a specific target. As protein-based drug discovery is a relatively new field compared to traditional drug discovery, more research is needed to identify new hotspots on protein-protein interfaces. Existing studies do computationally identify hotspots on these interfaces, but most of the models developed only attain accuracies of around 70% (Kim, Chivian & Baker, 2004; Tuncbag, Keskin & Gursoy, 2010). Moreover, many studies do not effectively integrate information relating to amino acid

chains and other structural information relating to the complex/interface, and/or have completely different approaches to predict the likelihood of hotspots on a particular interface.

For example, molecular dynamics (MD) simulations have been used to elucidate the mechanisms of protein interactions and their viability for drug discovery. This strategy has mixed results however - although the approach of molecular dynamics simulations have relatively high predictive power, these simulations are computationally expensive (Cukuroglu et al., 2014). In contrast, knowledge-based machine learning techniques have the advantage of providing accurate results based on the properties/features of a specific interaction. Machine learning and other statistical approaches allow for a high predictive power of hotspot detection, while being computationally efficient, provided that the features inputted into the model are relevant.

This leads to the proposed research question, "Can the development of a machine learning model lead to the discovery of new druggable targets and new drug-disease associations?" The hypothesis was that the integration of different protein-protein interaction features will lead to promising new hotspots. In addition, new drug-disease associations could potentially be identified from these hotspots to treat deadly diseases such as cancer.

To test this hypothesis, 1) a machine learning model was developed and 2) its ability to integrate multiple features, including structural information, such as that associated with amino-acid chains, to enhance the ability to predict protein-protein interface hotspots was evaluated. In addition, virtual drug screening of a set of hotspots identified by the machine learning model developed herein was performed in order to identify potentially new drug-disease associations. Phase 1 consisted of developing the machine learning model to identify potential protein-protein interface hotspots that could be viable as a drug target, using the cancer-associated EphB2-ephrinB2 protein complex (PDB code: 1KGY) for illustration. Phase 2 of this project aimed to identify small molecules that could act as inhibitors or disruptors to the hotspots identified for further analysis in Phase 1.

The machine learning model developed in Phase 1 achieved a precision-recall score of 0.605 and an area under receiver operating characteristic (AUROC or AUC) of .846 on the testing test, and identified

residues 1122-1126 on this complex as potential hotspot residues. This information was then used to generate a pharmacophore in Phase 2 which identified nine drug candidates to disrupt the EphB2-ephrinB2 complex. Out of these candidates, further literature review identified four drug candidates that could treat diseases that are overexpressed by this complex: cimetidine, idarubicin, pralatrexate, and nadolol. Although nadolol has been relatively unexplored in its potential of treating certain cancers, a drug with a similar chemical makeup, propranolol, has been identified to treat multiple cancers including colon cancer, which is linked to the overexpression of the EphB2-ephrinB2 complex, (Pantziarka et al., 2016) (Işeri et al., 2014), and thus highlights significant repositioning opportunities for nadolol.

METHODS

Dataset Collection and Feature Aggregation

As a starting point, the dataset and codebase from the SpotOn study (Moreira et al., 2017) were acquired. This study was selected as the starting point for its high effectiveness in identifying potential hotspots that could aid in drug discovery. The SpotOn database already has information regarding amino acid composition, solvent-accessible surface area (SASA) information, position-specific scoring matrices (PSSMs), the number of amino acids at 2.5 and 4.0 Angstrom, the number of nearby hydrophobic residues, the total change in solvent accessible surface area, the number of interfacial residues, pseudo-amino acid composition, and scales-based descriptors of 2D and 3D descriptors from the protr R package (see below) for a total of 881 features.

In order to add more information to this dataset to better aid model prediction, the protr R package (Xiao et al., 2015) was used to add more features related to amino acid composition, dipeptide composition, etc., to the already pre-existing data. Additionally, data related to pair potential, complex/monomer accessible surface area, residue information, amino acid information, etc. were extracted from the HotPoint database (Tuncbag, Keskin & Gursoy, 2010) and then added to the pre-existing dataset. This data was added to add more information regarding the entire protein complex, as evidenced by most of protr's features, and to add residue specific features such as pair potential that could improve predictive power. The addition of new features in the protr R package and the HotPoint database led to a total of 2323 features.

Upon further investigation of the SpotOn dataset, we found that chains I of proteins with PDB code 2FTL, 3SG8, and 1CH0 do not exist as specified in the Protein Data Bank. In the SpotOn study, these chains are specified, and features were derived for these chains; however, in this study, as additional information is added and these chains could not be identified, these chains have been removed from our dataset. This leads to a total of 520 protein residues, lower than SpotOn's 534 protein residues.
In order to derive features on our prediction dataset with the EphB2-ephrinB2 complex (PDB code: 1KGY), we first downloaded the structure from the Protein Data Bank, and ran this structure through the SpotOn's codebase/pipeline to collect features specific to the SpotOn study. Then, we sequentially added additional features unique to this study, such as from the protr's R package and features from the HotPoint database.

Preprocessing and Feature Engineering

Similar to the SpotOn study, both the training and testing sets were normalized, and the testing set was normalized using mean and standard deviation of the training set. In addition, before the model was run, data balance had to be accounted for, and oversampling was performed in order to retain the properties of the majority class without sacrificing the information available in this class (More, 2016). SMOTE, or synthetic minority oversampling technique, was performed with k=5 nearest neighbors. (Chawla et al., 2002) To account for multicollinearity, principal component analysis was also performed. This leads to four different combinations: a pipeline without any changes to the training data, a pipeline with only SMOTE applied, a pipeline with only PCA applied, and a pipeline with both SMOTE and PCA applied.

Before the model was trained, the dataset was first subjected to feature engineering. Three existing features that were selected for further exploration are the number of intermolecular contacts within 4.0 Angstroms (#Dist-4.0), the number of hydrophobic contacts (#Hydrophobic), and the pair potential of a specific residue (Pair Potential). We hypothesized that an increase of hydrophobic contacts would cause a

decrease in hydrophobic pair potential due to the attractive interaction because of the hydrophobic effect (Israelachvili & Pashley, 1982). As a result, we multiplied both variables and multiplied by -1 to amplify the effects of this association and accounting for the inverse correlation. In addition, we hypothesized that the number of intermolecular contacts will increase the pair potential as this may lead to many body potentials, which are mostly repulsive at short distances (Byggmästar, Granberg & Nordlund, 2018). To model this association, #Dist-4.0 and #Hydrophobic are multiplied to amplify the effects as well. These two new engineered variables were named *#Dist-4.0 * Pair Potential* and *-#Hydrophobic * Pair Potential*. This lead to a total of 2323 features on the training and testing datasets, as well as our dataset containing residue information on the crystal structure of the EphB2-ephrinB2 complex (PDB code: 1KGY).

Machine Learning Model Selection

Five different machine learning models were selected in order to evaluate and develop a model: linear support vector classifier (LSVC), XGBoost (XGB), a random forest classifier (RF), K Nearest Neighbors (KNN), multilayer perceptron neural network (MLP), and a Gaussian Naïve Bayes (GNB). This data was then split into a training:testing set ratio of 80:20. 10-fold cross validation was performed on the training set to prevent overfitting. GridSearch was performed in order to identify the best combination of hyperparameters/parameters that could yield the best results. The following hyperparameters/parameters were tested: LSVM, with C equal to 1, 10, 50, 100, 500, 1e3, 5e3, 1e4, 5e4, 1e5; RF, with the number of estimators equal to 50, 100, 150, 250, 350, 500, and maximum depth of 5, 6, 7, 8, 9, 10; XGB, with a learning rate of .001, .01, .1, the number of estimators as 50, 100, 150, 200, and maximum depth of 4, 5, 6; KNN, with n neighbors of 1, 3, 5, 10, 15, 20; a multilayer perceptron model of hidden_layer_sizes (10, 10, 10), (50, 1), (10, 10), (10, 1), and alpha of 0.0001, 0.0002, 0.0005, 0.001; and GNB with variance smoothing of 1e-8, 1e-7, 1e-6, 1e-5, and 1e-4. The metric used to identify the best model from these sets of parameters on the validation set is precision-recall, as it is incredibly robust in dealing with imbalanced data. Four different run conditions on the four different pipelines was also run and the results are

compared. The run conditions on the highest scoring pre-processing dataset will be used to build an ensemble model, similar to the SpotOn study. If the ensemble model has a higher predictive capability than any individual model, the ensemble model will then be used to predict hotspots on the EphB2-ephrinB2 complex, as this complex has been overexpressed in many cancer cells, most notably in prostate, gastric, colorectal and melanoma cancers. (Pasquale, 2010) PyMol was utilized to visualize the hotspots predicted on the EphB2-ephrinB2 complex.

Small Molecule Selection

A cluster of hotspots was identified and LigandScout (Wolber & Langer, 2005) was used to create an apo-site pharmacophore. Virtual screening was then performed on this pharmacophore to identify possible new drug indications. To perform the drug screening, an approved Drugbank (Wishart et al., 2008) database that has a library of all molecules that have molecular weight from 150 to 500 daltons was used. These small molecules were then ranked by the LigandScout software to identify molecules that most strongly conform to the pharmacophore based on the chemical and structural properties of that molecule. The drug-disease associations were then verified with scientific literature to assess the validity and efficacy of the model, and then we identified new drug-disease associations that have not been previously identified by cross-referencing existing scientific literature.

RESULTS

Phase 1

Table 1: Average test metrics of algorithms tested on pre-processing pipelines

194 The average test metrics of each of the six algorithms tested on the 4 different pre-processing pipelines
195 are shown in Table 1. As the preprocessing pipeline where only SMOTE is applied has the highest
196 precision-recall, F1-score, MCC, and Kappa – all metrics that account for class imbalanced data – the top
197 algorithms from this pipeline are used in order to create an ensemble model.
198
199 **Table 2: Best Individual Algorithms in SMOTE-only pipeline**
200 In Table 2 are the best individual algorithms tested in the SMOTE only pipeline. The best set of
201 hyperparameters were selected using GridSearch as follows: the support vector classifier with C=1000,
202 the random forest classifier with maximum depth of 9 trees and the total number of estimators at 250
203 trees, an XGBoost classifier with learning rate .01, maximum depth of 4, and 150 estimators, K-nearest
204 neighbors with 5 neighbors, a multi-layer perceptron classifier with alpha as .0005 and two layers of 10
205 neurons each, and a Gaussian Naïve Bayes of variable smoothing of .001.
206
207 **Table 3: Comparison of our study vs SpotOn**
208 *This data was adapted from the SpotOn study
209 The results of the SMOTE only pipeline were compared with SpotOn's highest pre-processing procedure,
210 which was the upsampling of their dataset. Although the precision-recall statistic was not provided by the
211 SpotOn study, other class imbalance-sensitive metrics, such as F1 and MCC, were provided. Our
212 algorithms outperform that of SpotOn's ScaledUp processing step in class imbalance-sensitive metrics
213 and sensitivity.
214
215
216 **Table 4: Different ensemble classifiers (stacking and voting) were tested**
217 The top ranking algorithms in the SMOTE only pipeline are used to develop an ensemble classifier to
218 achieve better performance compared to any single algorithm. Different ensemble algorithms are tested:
219 stacking, where a meta-classifier is used to combine the predictive power multiple base classifiers, and

voting, a simple ensemble method where each of the six algorithms tested votes on a specific data point, and a simple majority vote is used to predict the classification of that data point. In this case, the meta-classifier used during stacking is a Logistic Regression classifier where C=5. Each individual model is used as a base model separately with the meta-classifier, and all models are combined with the meta-classifier. All ensemble models are run on the SMOTE only pipeline. In the voting ensemble, hard voting was implemented, and all six algorithms are subjected to majority voting. Here, the best performing classifier was the stacking classifier where all models are combined with the meta-classifier. However, the precision-recall score of this ensemble method is still lower than that of the top individual model, the MLPClassifier in the SMOTE only pipeline.

Table 5: Comparison of our study to other studies

* Columns 2 through 7 are adapted from the SpotOn study to perform the side-by-side comparison among the algorithms

A comparison of the accuracy and performance of the model developed herein, shown in bold, compared with SpotOn. In our model, the multilayer perceptron classifier was our top performing algorithm, and was thus used to develop to predict hotspots with high accuracies. The SpotOn study (Moreira et al., 2017) was used in order to identify the testing accuracies of the SpotOn study and those of the other studies as well. The other studies that are compared to are SpotOn, SBHD213, Robetta23, KFC2-A24, KFC2-B, and CPORT25. (Kim, Chivian & Baker, 2004; Martins et al., 2014) (de Vries & Bonvin, 2011; Zhu & Mitchell, 2011)

Figure 1: Feature importances of the top tree-based classifier

The top features in the top ranking tree-based classifier (random forest). Features near the bottom of the graph have higher feature importances.

As the highest ranking classifier, the multilayer perceptron model, is considered a “black box”, and the interpretability of the predictions of the model are difficult to understand, the top tree-based classifier – the random forest - was used to identify features. In order to identify the most relevant features, highest

ranking tree-based classifier from the SMOTE only pipeline, the random forest classifier, was used in order to analyze top features, and to understand the significance of adding new features to the existing dataset as provided by the SpotOn study. Five out of the top fifteen features (Pair Potential, Relative Complex ASA, Complex ASA, and the engineered features *Dist-4.0*Pair Potential* and *-Pair Potential * Hydrophobic*), were added in this study exclusively, and highlights the improvement in predictive capabilities of the addition of these features.

Figure 2: The EphB2-ephrinB2 complex with highlighted residues using PyMol

Residues 1112 and 1122-1126 are highlighted as shown in green as surface markers. The rest of the complex is in pink.

The chain E of the EphB2-ephrinB2 complex associated with cancer cells. PyMol (Delano WL, 2002) was used to derive the complex and highlight residues 1112 and 1122-1126. Predicted druggable hotspot residues are shown as more visible surface markers (in green), and the other residues are shown in pink or light red. Residues 1122-1126 were selected for further investigation for drug screening as consecutive residues may be used as initial fragments in drug screening. (Modell, Blosser & Arora, 2016) These residues were then utilized to create the apo-site pharmacophore as shown in Figure 3, and the 26-feature pharmacophore in Figure 4.

To determine whether this approach accurately predicts new hotspots in comparison with existing models, analysis was performed comparing the predictive capability of the existing models with the model developed herein. In this study, a multilayer perceptron model is utilized to predict new hotspots, and performed better overall compared to most other protein-protein interface models, as shown in Table 5. However, our model did perform worse than the existing SpotOn study.

In context, sensitivity is the ability for the model to identify the hotspots and the specificity/recall is the ability for the model to identify the non-hotspots, and both of these statistics are defined as:

$$Recall = Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

Precision is defined as:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

F1, MCC, Kappa, and Precision-Recall are all metrics that are robust in dealing with data imbalance. They are defined as:

$$f1 = 2 * \frac{precision * recall}{precision + recall}$$

$$MCC = \frac{True\ Positive * True\ Negative - False\ Positive * False\ Negative}{\sqrt{(True\ Pos + False\ Pos)(True\ Pos + False\ Neg)(True\ Neg + False\ Pos)(True\ Neg + False\ Neg)}}$$

$Kappa = (p_o - p_e)/(1 - p_e)$ where p_o is the probability of agreement assigned to any sample, and p_e is the expected/hypothetical probability of chance agreement.

$Precision - Recall = \sum_n (R_n - R_{n-1})P_n$ where P_n and R_n are precision and recall, respectively, at the n^{th} threshold.

All of these calculations are calculated using the Scikit-learn package in Python. In Figure 3, the predicted hotspot residues of the EphB2-ephrinB2 complex associated with cancer cells are shown as more pronounced surface markers. The EphB2-ephrinB2 complex was selected for its role in a variety of cancers, as detailed in the discussion section.

Phase 2

In phase 2 of this project, virtual drug screening was utilized to identify novel drug-disease associations using the hotspots previously identified. An apo-site grid was implemented on hotspot residues 1122, 1123, 1124, 1125, and 1126 as identified via the machine learning model on the EphB2-ephrinB2 complex in Figure 3. This grid was then utilized to develop the pharmacophore.

Figure 3: Apo-Site Grid for residues 1122-1126

294

295 Apo site pharmacophore of residues 1122-1126. The gray parts of the grid indicate the levels of buriedness and surface area.

296 An apo-site grid was developed and implemented on hotspot residues 1122, 1123, 1124, 1125, and 1126 as
297 identified via the machine learning model on the EphB2-ephrinB2 complex. This grid was developed by
298 first calculating the pockets of hotspot residues 1122-1126 on LigandScout (Wolber & Langer, 2005). This
299 grid was then utilized to develop the pharmacophore in Figure 4.

300

301

302 **Figure 4: Pharmacophore model of residues 1122-1126**

303 This figure shows the 26-feature pharmacophore developed using an apo-site grid derived using hotspot
304 residues 1122, 1123, 1124, 1125, and 1126 identified in Figure 3 via the machine learning model. A
305 pharmacophore identifies the key parts of the molecular features that define the function and shape of a
306 specific ligand, and includes features such as H-bond acceptors and donors, hydrophobic and aromatic
307 rings, etc. This pharmacophore is then used to identify drugs that fit its features. The scoring of this
308 screening procedure follows a pharmacophore-fit scoring function as provided in LigandScout. A
309 maximum number of two features are omitted from this multi-feature pharmacophore to identify small
310 molecule hits, and the best matching conformation is selected.

311

312 **Figure 5: Structure and relative structure of cimetidine in relation to the developed pharmacophore**

313 Cimetidine, currently an acid reflux medication, was identified via virtual screening to potentially bind to
314 the EphB2-ephrinB2 complex associated with cancer cells. The right image is cimetidine in relation to
315 the 26-feature pharmacophore developed as shown in Figure 4. A pharmacophore-fit score of 43.86 was
316 achieved during drug screening. Further literature review identified cimetidine as a potential
317 repositioning target for many different types of cancers, including melanoma, gastric, and colorectal
318 cancers. (Pantziarka et al., 2014)

319

Figure 6: Structure and relative structure of idarubicin in relation to the developed pharmacophore

Idarubicin, a chemotherapy medication that's currently used to treat breast cancer, was identified via virtual screening to potentially bind to the EphB2-ephrinB2 complex, where the expression of the complex is associated with cancer cells. The pharmacophore fit score of this small molecule is 45.46. This drug was also found to treat cancers liked to the EphB2-ephrinB2 complex such as melanoma and leukemia. (Martoni et al., 1986) (Jabbour et al., 2017) The right image is idarubicin in relation to the pharmacophore developed as shown in Figure 4.

Figure 7: Structure and relative structure of pralatrexate in relation to the developed pharmacophore

Pralatrexate, a T-cell lymphoma medication, was identified via virtual screening to potentially bind to the EphB2-ephrinB2 complex, where the expression of the complex is associated with cancer cells. This small molecule has a pharmacophore fit score of 47.41, and literature review suggests that this drug could potentially treat breast cancer and prostate cancer. (Yu, Zhao & Gao, 2018) (Serova et al., 2011) The right image is pralatrexate in relation to the pharmacophore developed as shown in Figure 4.

Figure 8: Structure and relative structure of nadolol in relation to the developed pharmacophore

Nadolol, a beta blocker, was identified via virtual screening to potentially bind to the EphB2-ephrinB2 complex, where the expression of the complex is associated with cancer cells. This small molecule has a pharmacophore fit score of 45.97, and literature review suggests that beta blockers could potentially treat a variety of cancers, including breast cancer and pancreatic cancer. (Ishida et al., 2016) A close relative of this drug, propranolol, can induce apoptosis in liver cancer cells. (Wang et al., 2018) This research suggests nadolol's potential role in mitigating the effects of other cancers as well. The right image is nadolol in relation to the pharmacophore developed as shown in Figure 4.

346

347 Virtual drug screening identified nine drugs (pralatrexate, chlortetracycline, nadolol, imipenem,
348 idarubicin, valganciclovir, conivaptan, cimetidine, and barnidipine) that bind to the pharmacophore
349 shown in Figure 4. Further analysis via literature review identified four drug candidates to potentially
350 treat various types of cancers: cimetidine, idarubicin, pralatrexate, and nadolol. Figure 5 shows the
351 possibility for cimetidine, an antacid, to bind with the EphB2-ephrinB2 complex, and scientific literature
352 identified the possibility for this drug to potentially treat melanoma, gastric, and colorectal cancers
353 (Pantziarka et al., 2014). Figure 6 identifies the possibility for idarubicin, a chemotherapy drug used to
354 treat leukemia, to bind with the EphB2-ephrinB2 complex, and literature review identified the possibility
355 for this drug to potentially treat melanoma and leukemia (Martoni et al., 1986) (Jabbour et al., 2017).
356 Figure 7 demonstrates the possibility for pralatrexate, a T-cell lymphoma medication to bind to the
357 EphB2-ephrinB2 complex.

358

359 DISCUSSION

360 In this paper, we presented our development of a machine learning approach for identifying druggable
361 hotspots at protein-protein interfaces. Our algorithm builds on previously existing methods, most notably
362 the SpotOn study. Our approach combines molecular features that have not previously been combined,
363 such as the molecular descriptors used in the SpotOn and HotPoint studies, and additional information
364 related to amino acid composition as provided by the protr module. It applies various machine learning
365 techniques, such as 10-fold cross-validation, feature engineering, and ensembling techniques, including
366 voting and stacking. A multilayer perceptron classifier with two hidden layers of 10 neurons each and an
367 alpha of .0005 was used in order to achieve an AUROC of .846 and a precision-recall score of .605.

368 In order to find the most optimal pipeline, all four pipelines were run, and the pipeline that used
369 only SMOTE during the pre-processing step was chosen the most optimal pipeline due to its high
370 precision-recall score. The average metrics of all classifiers in each of the pre-processing steps are
371 recorded in Table 1. Furthermore, the results of each top performing classifier in the SMOTE only pre-

processing step are illustrated in Table 2. In Table 3, the average of the metrics of each individual algorithm in the most optimal pipeline, the SMOTE-only pre-processing pipeline, are compared with the average of each top-performing model in the ScaledUp pre-processing dataset in SpotOn, the highest performing dataset in that study. SpotOn-specific metrics are provided by the study itself. The individual models of our study performed better than the individual models of SpotOn as highlighted in Table 3. After this step, ensemble methods such as stacking and voting were implemented to potentially achieve even better results than any single model. The results of performing this step are shown in Table 4.

Although our models outperform that of SpotOn's individual models without any type of ensembling, the results of our approach are lower on three out of four metrics than the top performing ensemble model from the SpotOn study, as illustrated in Table 5. This may be due to one of many reasons. Even though there was an increase in the total number of features as compared to the SpotOn study, the slight decrease in the total number of samples could potentially negatively affect predictive performance. Another reason could be that the models tested are not diverse enough from each other to significantly boost performance via ensembling. Two of the models in this study are tree-based methods (random forest and gradient boosting). A greater diversity of these models would probably have boosted performance during stacking or voting, as a greater variety of base models have been shown to boost predictive performance. (Whalen & Pandey, 2013)

To illustrate our approach, we applied this model to analyze the EphB2-ephrinB2 complex, which has been overexpressed and associated with multiple types of cancer, including prostate, gastric, colorectal and melanoma cancers. (Pasquale, 2010) As the overexpression of the EphB2-ephrinB2 complex is associated with these cancers, further analysis for drug discovery could aid in identifying possible new hotspots that potentially aid in drug discovery in the fight against cancer (Barquilla & Pasquale, 2015). In addition, the viability for the EphB2-ephrinB2 complex, and more specifically the EphB2 receptor, for drug discovery has been examined, and it was determined that small molecules could potentially disrupt and/or bind to the ephrin binding pocket. (Chrencik et al., 2007) (Noberini, Lamberto & Pasquale, 2012)

The effectiveness of introducing new engineered features was demonstrated by the feature importances of our top tree-based classifier, the random forest classifier (Figure 1). Our algorithm identified a set of residue hotspots (Figure 2). These hotspots were then used to generate a pharmacophore model (Figure 4). This model was used to identify drugs with similar characteristics that could be potentially used to modulate the molecular functions of the EphB2-ephrinB2 complex. The identified drugs included compounds already used for cancer treatment, such as pralatrexate, a T-cell lymphoma medication, as well as non-cancer medication, such as cimetidine, an antacid, and nadolol, a beta blocker that can treat cardiac conditions. Literature review suggests that pralatrexate can potentially treat breast cancer and prostate cancer, and highlights the possibility for this small molecule to treat other conditions. (Yu, Zhao & Gao, 2018) (Serova et al., 2011) Figure 8 identifies nadolol, a beta blocker that can treat cardiac conditions, as a candidate to bind to the EphB2-ephrinB2 complex. Literature review strongly supports that beta blockers can be repositioned to treat other cancers, such as cancer, and has identified a close relative of nadolol, propranolol, as a potential treatment against multiple cancers, including colon cancer. (Işeri et al., 2014)

Conclusion

The model developed herein in phase one compares favorably with those developed in prior studies and offers enhanced predictive ability for identifying new druggable hotspots, including possible druggable hotspots for cancer-related protein interfaces. The predictive capabilities of the model developed herein are high, offering a high AUROC and overall predictive performance to date. Herein, a multilayer feedforward perceptron model with alpha .0005 and two layers of ten neurons was developed to successfully identify hotspots.

Phase two of this project aims to identify possible drugs for repositioning. Structural properties of the identified hotspot residues, such as H-bond acceptors and donors, were identified as feature sets to aid in drug development. The efficacy of the model developed herein has been demonstrated through its

successful ability to predict drug-disease associations previously identified in literature, including cimetidine, idarubicin, and pralatrexate. Importantly, nadolol has been uniquely identified in this study to potentially treat conditions caused by the overexpression of the EphB2-ephrinB2 complex. This work aims to yield better predictions in terms of hotspot discovery by primarily increasing the sheer amount of data that is available regarding protein-protein interactions. As a consequence, this work has shown that the increases in predictive power as a result of this addition of data.

Possible avenues for future work include drug development using the pharmacophores identified in this study to treat these diseases. Hopefully, by identifying hotspot residues with unparalleled accuracy and identifying possible drug repositioning opportunities, traditional drug development based on these residues and repositioned drugs could yield new and effective treatments for diseases such as cancer. In addition, adding additional novel features and data for hotspot identification, especially those that directly correlate with the extent of how energetically favorable residues are, could further improve model performance. Another avenue for future work would be to streamline the workflow of both phases. Phase one is automated with the help of the machine learning model. However, phase two requires manual input of the hotspot residues as identified in phase one to identify potential drug candidates. A more streamlined process would improve functionality and ease of use.

REFERENCES

Bakail M, Ochsenbein F. 2016. Targeting protein–protein interactions, a wide open field for drug design. *Comptes Rendus Chimie* 19:19–27. DOI: 10.1016/j.crci.2015.12.004.

Barquilla A, Pasquale EB. 2015. Eph Receptors and Ephrins: Therapeutic Opportunities. *Annual review of pharmacology and toxicology* 55:465–487. DOI: 10.1146/annurev-pharmtox-011112-140226.

Byggmästar J, Granberg F, Nordlund K. 2018. Effects of the short-range repulsive potential on cascade damage in iron. *Journal of Nuclear Materials* 508:530–539. DOI: 10.1016/j.jnucmat.2018.06.005.

449 Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. 2002. SMOTE: Synthetic Minority Over-sampling
450 Technique. *Journal of Artificial Intelligence Research* 16:321–357. DOI: 10.1613/jair.953.
451 Chrencik JE, Brooun A, Recht MI, Nicola G, Davis LK, Abagyan R, Widmer H, Pasquale EB, Kuhn P.
452 2007. Three-dimensional Structure of the EphB2 Receptor in Complex with an Antagonistic
453 Peptide Reveals a Novel Mode of Inhibition. *The Journal of biological chemistry* 282:36505–
454 36513. DOI: 10.1074/jbc.M706340200.
455 Cukuroglu E, Engin HB, Gursoy A, Keskin O. 2014. Hot spots in protein–protein interfaces: Towards
456 drug discovery. *Progress in Biophysics and Molecular Biology* 116:165–173. DOI:
457 10.1016/j.pbiomolbio.2014.06.003.
458 Delano WL. 2002. The PyMOL molecular graphics system.
459 Hauser AS, Attwood MM, Rask-Andersen M, Schiöth HB, Gloriam DE. 2017. Trends in GPCR drug
460 discovery: new agents, targets and indications. *Nature Reviews Drug Discovery* 16:829–842.
461 DOI: 10.1038/nrd.2017.178.
462 Hopkins AL, Groom CR. 2002. The druggable genome. *Nature Reviews Drug Discovery* 1:727–730.
463 DOI: 10.1038/nrd892.
464 Işeri OD, Sahin FI, Terzi YK, Yurtcu E, Erdem SR, Sarialioglu F. 2014. beta-Adrenoreceptor antagonists
465 reduce cancer cell proliferation, invasion, and migration. *Pharmaceutical Biology* 52:1374–1381.
466 DOI: 10.3109/13880209.2014.892513.
467 Ishida J, Konishi M, Ebner N, Springer J. 2016. Repurposing of approved cardiovascular drugs. *Journal*
468 *of Translational Medicine* 14. DOI: 10.1186/s12967-016-1031-5.
469 Israelachvili J, Pashley R. 1982. The hydrophobic interaction is long range, decaying exponentially with
470 distance. *Nature* 300:341–342. DOI: 10.1038/300341a0.
471 Jabbour E, Short NJ, Ravandi F, Huang X, Xiao L, Garcia-Manero G, Plunkett W, Gandhi V, Sasaki K,
472 Pemmaraju N, Daver NG, Borthakur G, Jain N, Konopleva M, Estrov Z, Kadia TM, Wierda WG,
473 DiNardo CD, Brandt M, O'Brien SM, Cortes JE, Kantarjian H. 2017. A randomized phase 2

study of idarubicin and cytarabine with clofarabine or fludarabine in patients with newly diagnosed acute myeloid leukemia. *Cancer* 123:4430–4439. DOI: 10.1002/cncr.30883.

Kim DE, Chivian D, Baker D. 2004. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research* 32:W526-531. DOI: 10.1093/nar/gkh468.

Koes DR, Dömling A, Camacho CJ. 2018. AnchorQuery: Rapid online virtual screening for small-molecule protein–protein interaction inhibitors. *Protein Science : A Publication of the Protein Society* 27:229–232. DOI: 10.1002/pro.3303.

Lee AC-L, Harris JL, Khanna KK, Hong J-H. 2019. A Comprehensive Review on Current Advances in Peptide Drug Development and Design. *International Journal of Molecular Sciences* 20. DOI: 10.3390/ijms20102383.

López-Cortés A, Cabrera-Andrade A, Cruz-Segundo CM, Dorado J, Pazos A, Gonzáles-Díaz H, Paz-y-Miño C, Pérez-Castillo Y, Tejera E, Munteanu CR. 2019. Prediction of druggable proteins using machine learning and functional enrichment analysis: a focus on cancer-related proteins and RNA-binding proteins. *bioRxiv*:825513. DOI: 10.1101/825513.

Ma B, Nussinov R. 2014. Druggable Orthosteric and Allosteric Hot Spots to Target Protein-protein Interactions. *Current pharmaceutical design* 20:1293–1301.

Martins JM, Ramos RM, Pimenta AC, Moreira IS. 2014. Solvent-accessible surface area: How well can be applied to hot-spot detection? *Proteins* 82:479–490. DOI: 10.1002/prot.24413.

Martoni A, Pacciarini MA, Piana E, Pannuti F. 1986. A pilot study of oral idarubicin in metastatic melanoma. *Chemioterapia: International Journal of the Mediterranean Society of Chemotherapy* 5:414–415.

Modell AE, Blosser SL, Arora PS. 2016. Systematic Targeting of Protein-Protein Interactions. *Trends in pharmacological sciences* 37:702–713. DOI: 10.1016/j.tips.2016.05.008.

More A. 2016. Survey of resampling techniques for improving classification performance in unbalanced datasets. *arXiv:1608.06048 [cs, stat]*.

499 Moreira IS, Koukos PI, Melo R, Almeida JG, Preto AJ, Schaarschmidt J, Trellet M, Gümüş ZH, Costa J,
500 Bonvin AMJJ. 2017. SpotOn: High Accuracy Identification of Protein-Protein Interface Hot-
501 Spots. *Scientific Reports* 7:8007. DOI: 10.1038/s41598-017-08321-2.
502 Noberini R, Lamberto I, Pasquale EB. 2012. Targeting Eph Receptors with Peptides and Small
503 Molecules: Progress and Challenges. *Seminars in cell & developmental biology* 23:51–57. DOI:
504 10.1016/j.semcdb.2011.10.023.
505 Pantziarka P, Bouche G, Meheus L, Sukhatme V, Sukhatme VP. 2014. Repurposing drugs in oncology
506 (ReDO)-cimetidine as an anti-cancer agent. *Ecancermedicalscience* 8:485. DOI:
507 10.3332/ecancer.2014.485.
508 Pantziarka P, Bouche G, Sukhatme V, Meheus L, Rooman I, Sukhatme VP. 2016. Repurposing Drugs in
509 Oncology (ReDO)—Propranolol as an anti-cancer agent. *ecancermedicalscience* 10. DOI:
510 10.3332/ecancer.2016.680.
511 Pasquale EB. 2010. Eph receptors and ephrins in cancer: bidirectional signaling and beyond. *Nature*
512 *reviews. Cancer* 10:165–180. DOI: 10.1038/nrc2806.
513 Scott DE, Bayly AR, Abell C, Skidmore J. 2016. Small molecules, big targets: drug discovery faces the
514 protein–protein interaction challenge. *Nature Reviews Drug Discovery* 15:533–550. DOI:
515 10.1038/nrd.2016.29.
516 Serova M, Bieche I, Sablin M-P, Pronk GJ, Vidaud M, Cvitkovic E, Faivre S, Raymond E. 2011. Single
517 agent and combination studies of pralatrexate and molecular correlates of sensitivity. *British*
518 *Journal of Cancer* 104:272–280. DOI: 10.1038/sj.bjc.6606063.
519 Tuncbag N, Keskin O, Gursoy A. 2010. HotPoint: hot spot prediction server for protein interfaces.
520 *Nucleic Acids Research* 38:W402–W406. DOI: 10.1093/nar/gkq323.
521 de Vries SJ, Bonvin AMJJ. 2011. CPORT: a consensus interface predictor and its performance in
522 prediction-driven docking with HADDOCK. *PloS One* 6:e17695. DOI:
523 10.1371/journal.pone.0017695.

524 Wang F, Liu H, Wang F, Xu R, Wang P, Tang F, Zhang X, Zhu Z, Lv H, Han T. 2018. Propranolol
525 suppresses the proliferation and induces the apoptosis of liver cancer cells. *Molecular Medicine*
526 *Reports* 17:5213–5221. DOI: 10.3892/mmr.2018.8476.
527 Whalen S, Pandey G. 2013. A Comparative Analysis of Ensemble Classifiers: Case Studies in Genomics.
528 In: *2013 IEEE 13th International Conference on Data Mining*. 807–816. DOI:
529 10.1109/ICDM.2013.21.
530 Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M. 2008.
531 DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*
532 36:D901–D906. DOI: 10.1093/nar/gkm958.
533 Wolber G, Langer T. 2005. LigandScout: 3-D pharmacophores derived from protein-bound ligands and
534 their use as virtual screening filters. *Journal of Chemical Information and Modeling* 45:160–169.
535 DOI: 10.1021/ci049885e.
536 Xiao N, Cao D-S, Zhu M-F, Xu Q-S. 2015. protr/ProtrWeb: R package and web server for generating
537 various numerical representation schemes of protein sequences. *Bioinformatics* 31:1857–1859.
538 DOI: 10.1093/bioinformatics/btv042.
539 Yu L, Zhao J, Gao L. 2018. Predicting Potential Drugs for Breast Cancer based on miRNA and Tissue
540 Specificity. *International Journal of Biological Sciences* 14:971–982. DOI: 10.7150/ijbs.23350.
541 Zhu X, Mitchell JC. 2011. KFC2: a knowledge-based hot spot prediction method based on interface
542 solvation, atomic density, and plasticity features. *Proteins* 79:2671–2683. DOI:
543 10.1002/prot.23094.
544
545
546 AUTHOR CONTRIBUTIONS
547 Rohit Nandakumar conceived and performed the experiments as stated in this study, analyzed the data,
548 and co-authored the corresponding paper.
549 Dr. Valentin Dinu provided revisions and co-authored to this paper.

ACKNOWLEDGEMENTS

I would like to thank the researchers who conducted the SpotOn study, especially Ms. Irina Moreira, for providing the code and existing dataset that this study is built on top of.

I would also like to thank Dr. Michael McKelvy of Basha High School for his extensive feedback on my poster and project.

In addition, I would like to thank Mr. Thomas Lemker for his assistance in using the LigandScout software.

SUPPLEMENTARY INFO

All data used in this study is provided as the supplementary materials.

Table 1(on next page)

Average test metrics of algorithms tested on pre-processing pipelines

1 **Table 1: Average test metrics of algorithms tested on pre-processing pipelines**

Test	Precision-Recall	Precision	Recall	F1	AUROC	Accuracy	MCC	Kappa	Specificity
ONLY SMOTE	0.455	0.542	0.708	0.605	0.754	0.779	0.474	0.460	0.800
RAW	0.421	0.560	0.597	0.559	0.712	0.774	0.427	0.416	0.827
NO SMOTE, PCA	0.438	0.551	0.653	0.582	0.733	0.776	0.451	0.438	0.814
SMOTE, PCA	0.413	0.503	0.674	0.572	0.732	0.764	0.427	0.416	0.792

2

Table 2(on next page)

Best Individual Algorithms in SMOTE-only pipeline

1 **Table 2: Best Individual Algorithms in SMOTE-only pipeline**

Test	Precision-Recall	Precision	Recall	F1	AUROC	Accuracy	MCC	Kappa	Specificity
SVC	0.478	0.500	0.917	0.647	0.821	0.769	0.547	0.497	0.725
RF	0.521	0.667	0.667	0.667	0.783	0.846	0.567	0.567	0.900
GBC	0.477	0.625	0.625	0.625	0.756	0.827	0.513	0.513	0.888
KNN	0.306	0.359	0.583	0.444	0.635	0.664	0.236	0.222	0.688
MLP	0.605	0.704	0.792	0.745	0.846	0.875	0.665	0.663	0.900
Gaussian	0.344	0.400	0.667	0.500	0.683	0.692	0.318	0.297	0.700

2

Table 3(on next page)

Comparison of our study vs SpotOn

Table 3: Comparison of our study vs SpotOn

Test	SMOTE only	SpotOn's ScaledUp*
Accuracy	0.779	0.79
F1	0.605	0.52
AUROC	0.754	0.83
MCC	0.475	0.38
Sensitivity	0.708	0.48
Specificity	0.800	0.88

*This data was adapted from the SpotOn study

Table 4(on next page)

Different ensemble classifiers (stacking and voting) were tested

1 **Table 4: Different ensemble classifiers (stacking and voting) were tested**

2

Test Metrics	Precision-Recall	Precision	Recall	F1	AUROC	Accuracy	MCC	Kappa	Specificity
SVC (Stacking) w/ Logistic Regression	0.421	0.536	0.625	0.577	0.731	0.789	0.439	0.437	0.838
RF (Stacking) w/ Logistic Regression	0.541	0.696	0.667	0.681	0.790	0.856	0.588	0.588	0.913
GBC (Stacking) w/ Logistic Regression	0.558	0.667	0.750	0.706	0.819	0.856	0.613	0.611	0.888
KNN (Stacking) w/ Logistic Regression	0.487	0.615	0.667	0.640	0.771	0.827	0.527	0.526	0.875
MLP (Stacking) w/ Logistic Regression	0.523	0.621	0.750	0.679	0.806	0.837	0.576	0.571	0.863
Gaussian (Stacking) w/ Logistic Regression	0.508	0.600	0.750	0.667	0.800	0.827	0.558	0.552	0.850
All (Stacking)	0.569	0.708	0.708	0.708	0.810	0.865	0.621	0.621	0.913

w/ Logistic Regression									
Voting Classifier	0.462	0.6	0.625	0.612	0.75	0.817	0.493	0.493	0.875

3

Table 5(on next page)

Comparison of our study to other studies

Table 5: Comparison of our study to other studies

	Our model	SpotOn*	SBHD2*	Robetta*	KFC2-A*	KFC2-B*	CPORT*
AUROC	**0.846**	0.91	0.69	0.62	0.66	0.67	0.54
Sensitivity	**0.792**	0.98	0.7	0.29	0.53	0.28	0.54
Specificity	**0.900**	0.84	0.71	0.88	0.81	0.96	0.47
F1-score	**0.745**	0.96	0.62	0.39	0.56	0.42	0.42

* Columns 2 through 7 are adapted from the SpotOn study to perform the side-by-side comparison among the algorithms

Figure 1

Feature importances of the top tree-based classifier

The top features in the top ranking tree-based classifier (random forest). Features near the bottom of the graph have higher feature importances.

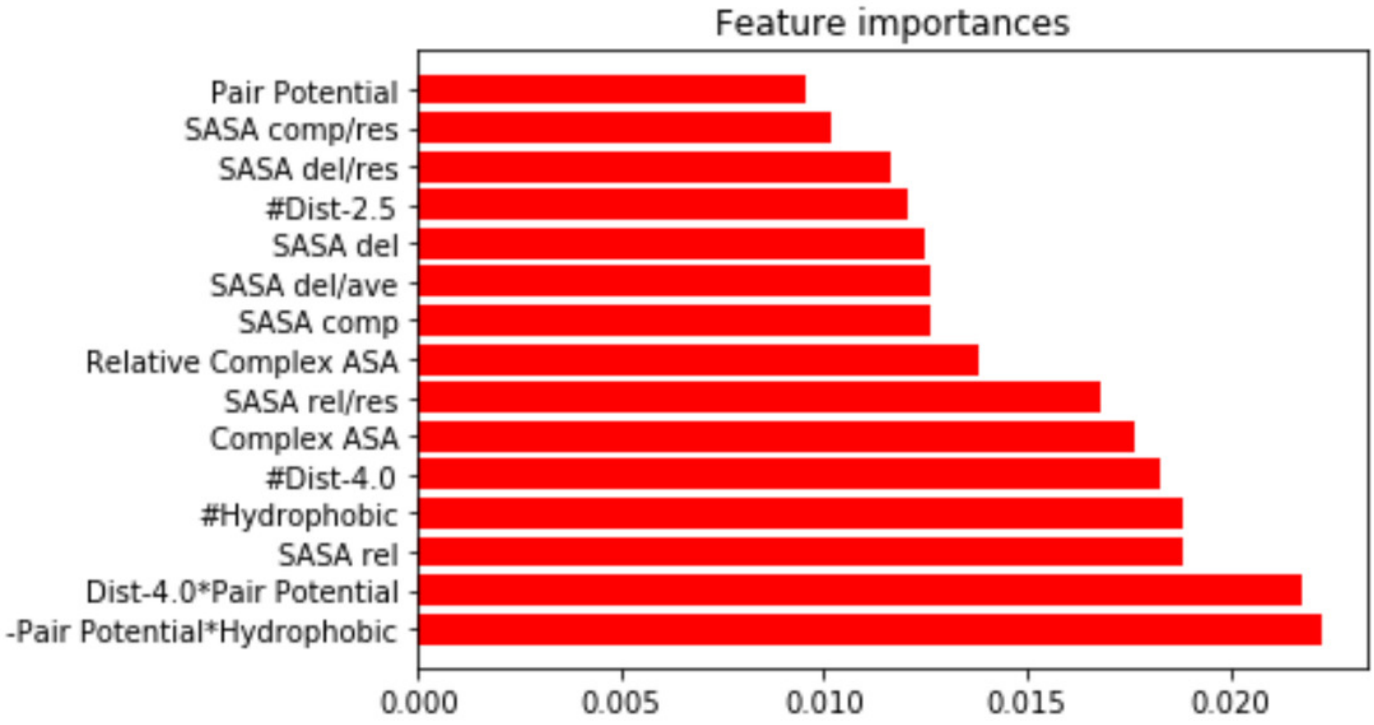

Figure 2

The EphB2-ephrinB2 complex with highlighted residues using PyMol

Residues 1112 and 1122-1126 are highlighted as shown in green as surface markers. The rest of the complex is in pink.

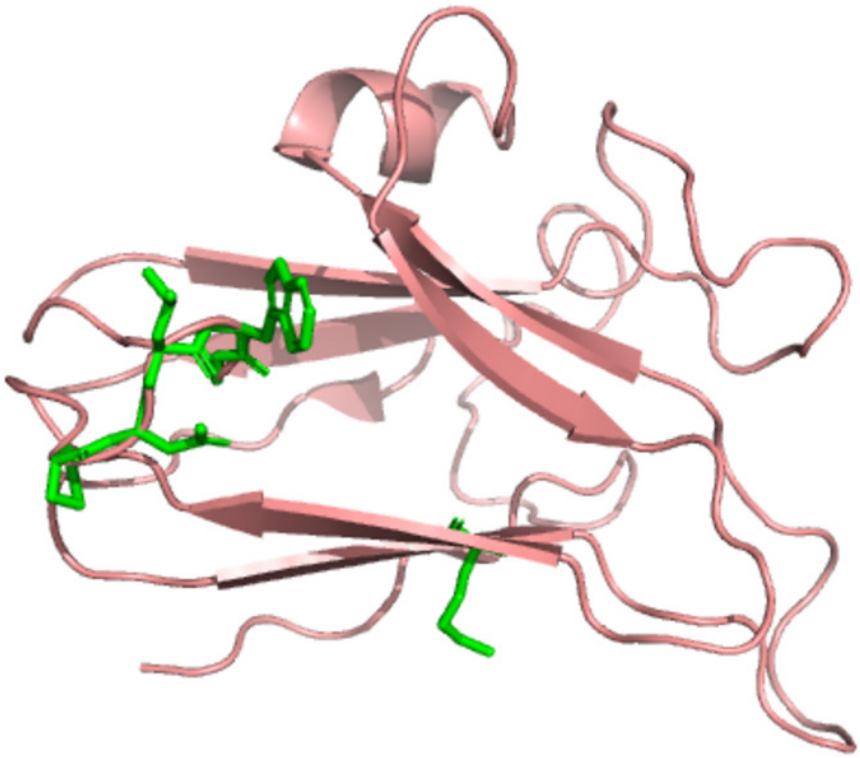

Figure 3

Apo-Site Grid for residues 1122-1126

Apo site pharmacophore of residues 1122-1126. The gray parts of the grid indicate the levels of buriedness and surface area.

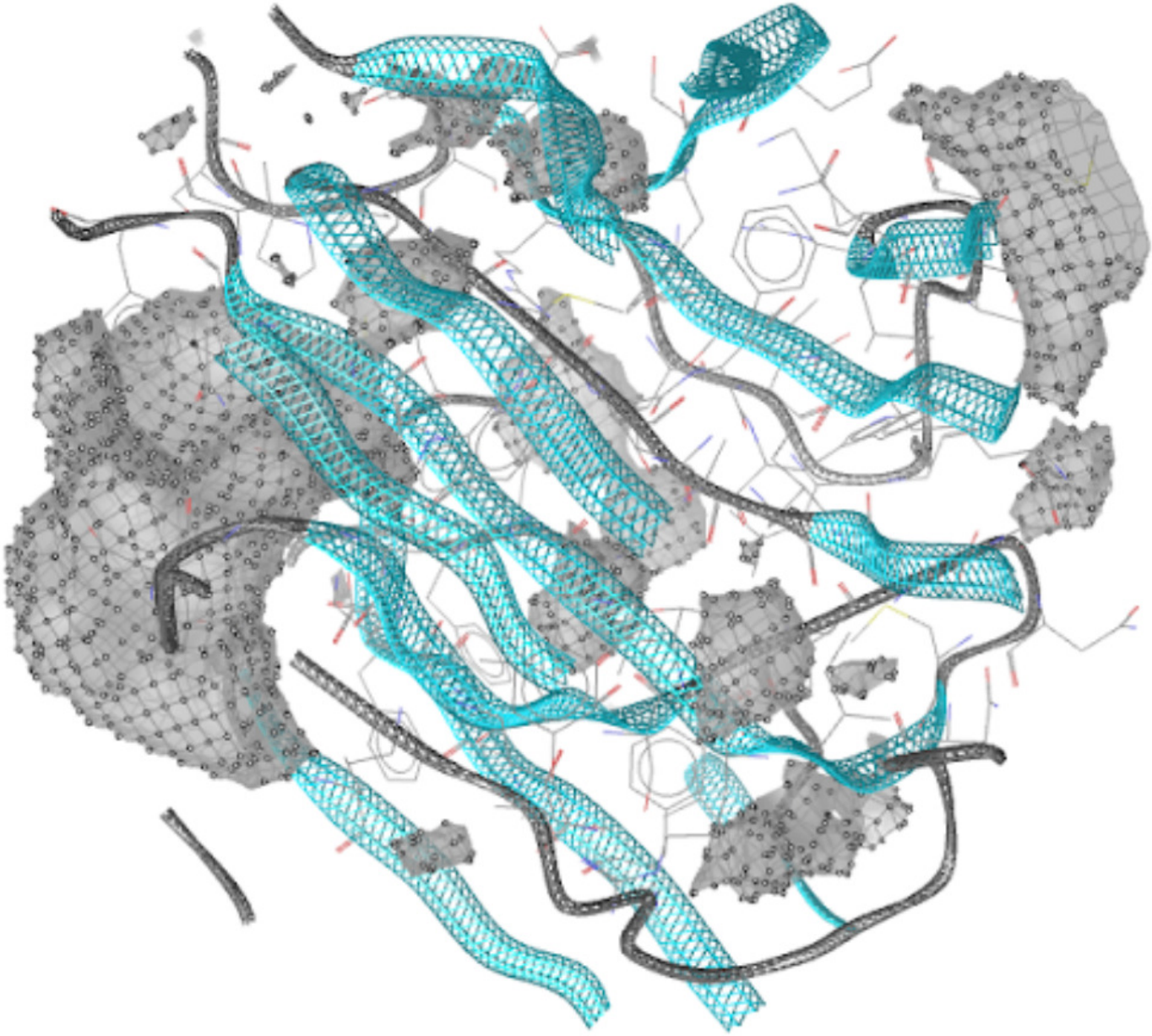

Figure 4

Pharmacophore model of residues 1122-1126

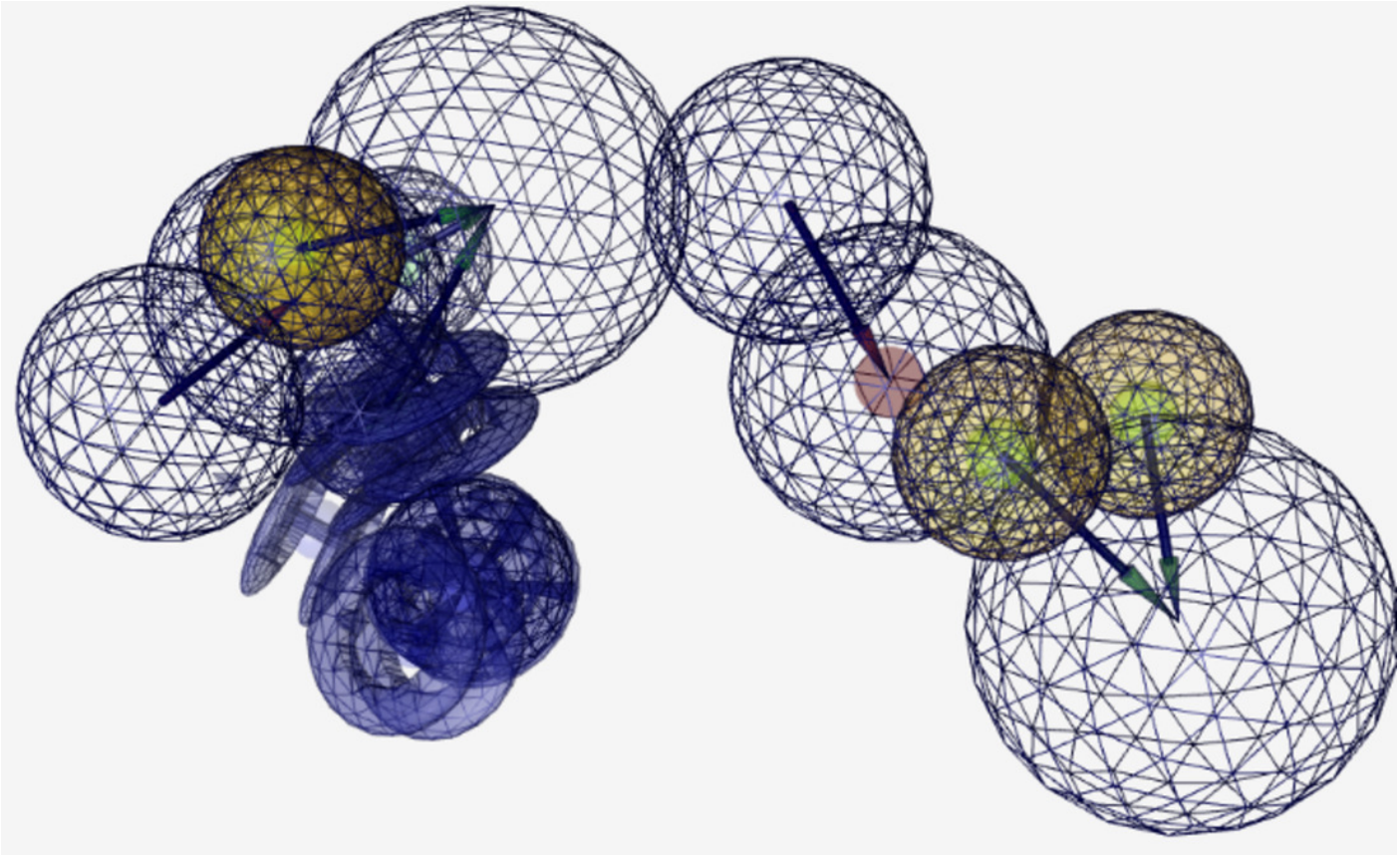

Figure 5

Structure and relative structure of cimetidine in relation to the developed pharmacophore

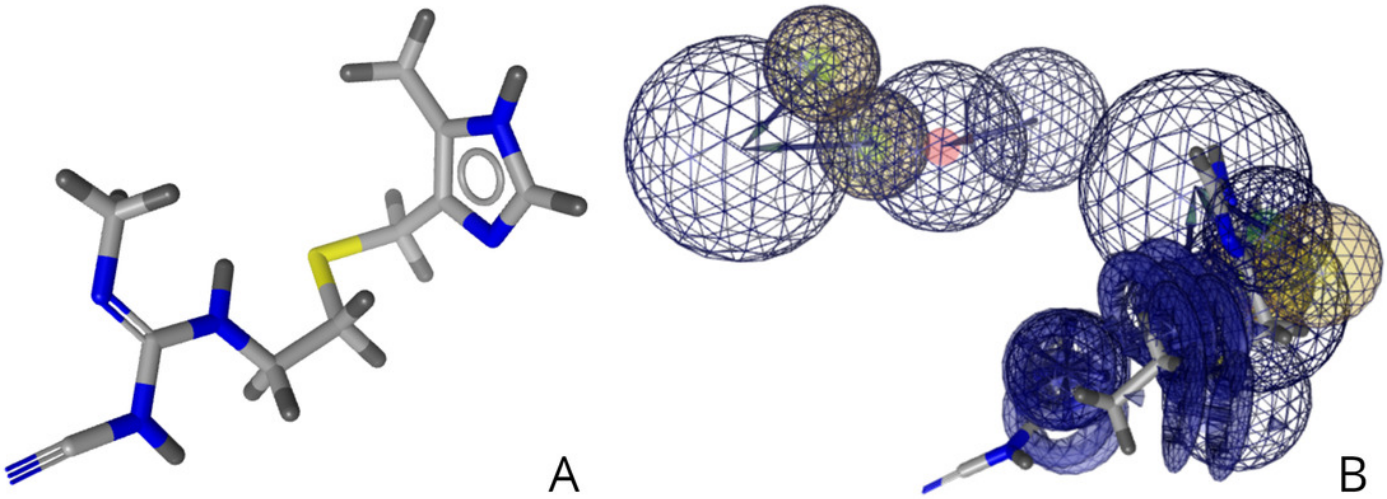

Figure 6

Structure and relative structure of idarubicin in relation to the developed pharmacophore

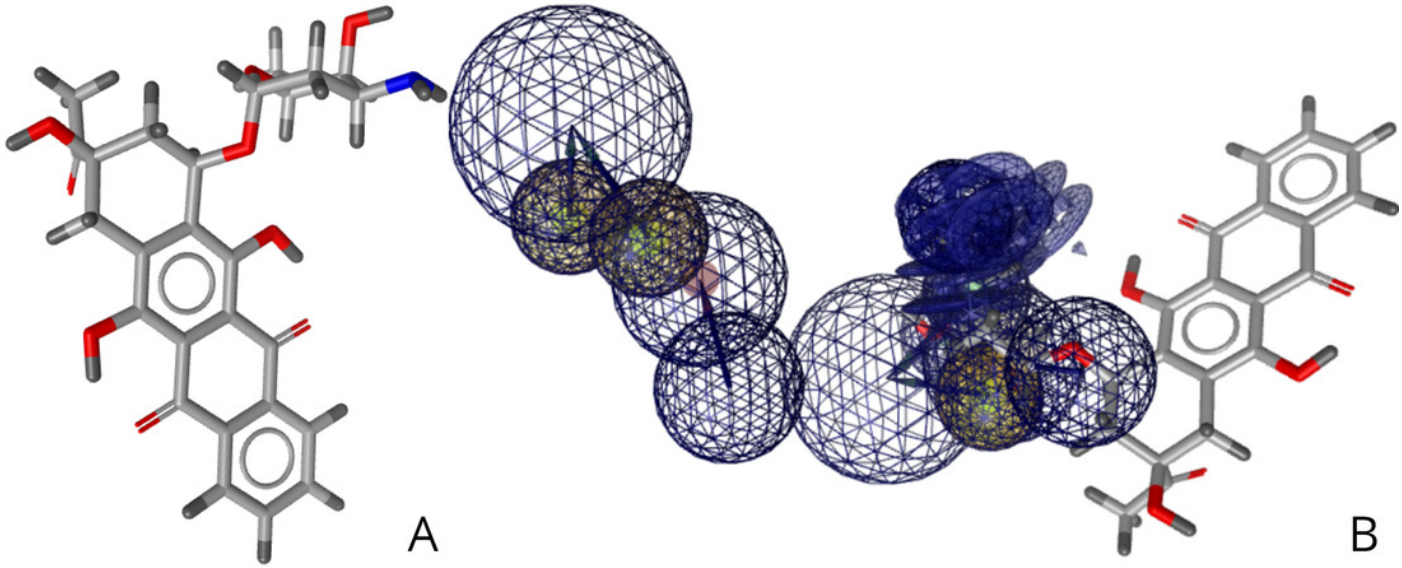

Figure 7

Structure and relative structure of pralatrexate in relation to the developed pharmacophore

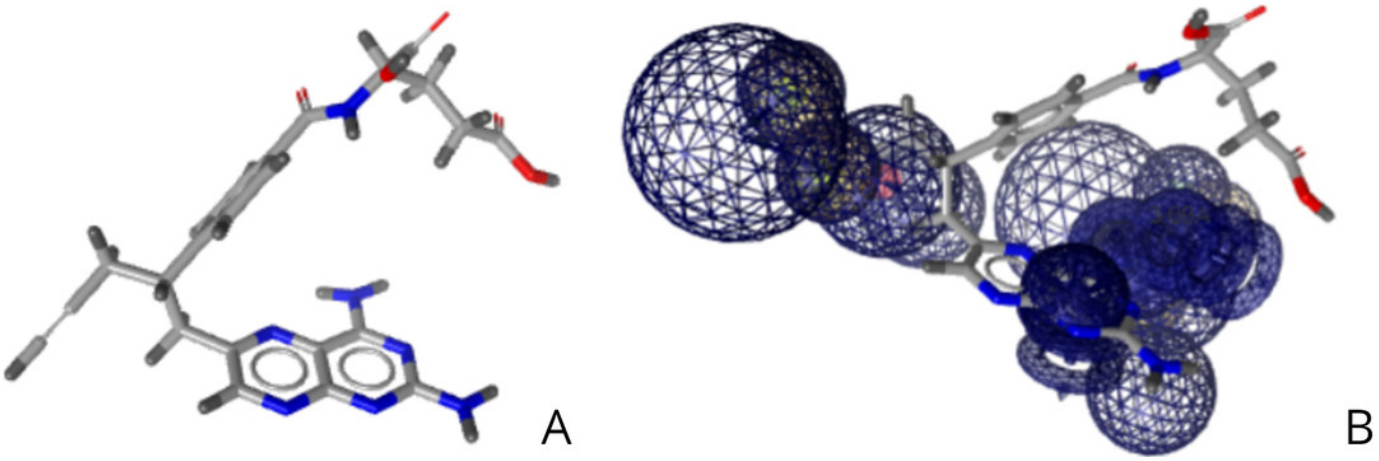

Figure 8

Structure and relative structure of nadolol in relation to the developed pharmacophore