American Journal of Applied Bio-Technology Research
(AJABTR)
Vol. 5, July 2026, pp. 40-54 DOI: doi.org/10.15864/ajabtr.515

Research Article

Structure-Based Molecular Docking Analysis of Fluorescent Ligands Targeting Key Cardiovascular Biomarker Proteins

Shalini Basu Roy, Urjoswati Maity and Ushnish Saha

Department of Biotechnology, School of Bio-Sciences and Technology, Vellore Institute of Technology, Vellore, Tamil Nadu, India.
shalini.basuroy2024@vitstudent.ac.in
urjoswati.maity2024@vitstudent.ac.in
ushnish.saha2024@vitstudent.ac.in

Received on 30th September 2025
Revision submitted on May 2026
Published on 28th July 2026

Abstract

Cardiovascular diseases (CVDs) continue to be among the major causes of death worldwide. Accurate diagnostic methods are thus highly desirable and require development. The development of fluorescence biosensors is an approach worth pursuing, provided that the choice of suitable fluorophores is made. In this study, we propose a computational strategy for the discovery and ranking of fluorescent probes targeting proteins related to CVDs. The transcriptome of the GEO database GSE116250 was studied to reveal the differential expression of genes in two types of cardiomyopathy: dilated and ischemic. Three targets were then chosen based on this information: *Atrial Natriuretic Peptide (ANP)*, *Natriuretic Peptide B* (*BNP*), and *Cardiac-specific isoform of Troponin I (TNNI3)*. A collection of ~300 fluorescent compounds retrieved from PubChem and ChEMBL databases was characterized by ECFP4 fingerprint and MACCS descriptors and encoded by graph convolutional networks. Random Forest was used for machine learning with stratified 5-fold cross-validation for evaluation alongside SVM and XGBoost models. Molecular docking of compounds with protein targets using AutoDock Vina and predicted protein structure from AlphaFold was followed by ranking using normalized weighting of docking scores and ADMET properties. Strong Pearson's correlation coefficients of docking and ML scores were confirmed for all three proteins tested ($r \approx -0.98$, $p < 10^{-33}$). Among the probes tested, *BP-Fluor 594 DBCO* and *5-FAM DBCO* emerged as the best performers. These findings demonstrate the potential of using this computational model as an approach for designing fluorescent probes in cardiology.

Index Terms: Cardiovascular biomarkers, Cheminformatics, Fluorescent probes, Machine learning, Molecular docking.

Introduction

This paper elaborates an integrative computational system for the prioritization of fluorescent molecules as cardiovascular biomarkers using molecular docking and machine learning (ML). Current research indicates that emerging technologies of ML and Artificial Intelligence (AI) are revolutionizing the discovery of cardiovascular protein biomarkers by improving the use of fluorescence-based biosensing technology and integrating the two[1]. Using ML algorithms, robust data sets (including rich proteomics, genomics and clinical information from potential biomarkers) have been used to enhance accurate diagnosis and prediction of cardiovascular disease risk, at a level of precision that has never before been attained, with traditional statistical techniques occasionally falling short[2,3].

Using the capabilities of structure mining and accessing chemical databases, subtly optimizing potential ligands and dyes can improve detection sensitivity and specificity on fluorescence-based platforms[4]. Public databases of curated data have aided the rapid discovery of probes for fluorescence-based biosensors through chemical fingerprint availability, SMILES representation and graph-based molecular embeddings. Fluorescence-based biosensors are fast and comparatively inexpensive with convenient point-

American Journal of Applied Bio-Technology Research
(AJABTR)
Vol. XX, July 2026, pp. 1111-1111 DOI: xxxxxxxxxxxxxxxxxx

of-care testing. However, as of now, the selection of probes for use in fluorescence-based biosensors is primarily determined through empirical means; to date, there is limited systematic integration of ML and protein-to-ligand docking for selection of fluorescent molecules for cardiovascular biomarkers. This combination of interpretable ML algorithms used in conjunction with biosensors provides an additional promise for the future, but much work lies ahead in the areas of robustness and clinical utility across heterogeneous populations.

In previous ML-assisted studies, emphasis was placed on biomarker discovery from gene expression data or therapeutic ligand design[5,6]; however, none of them have expanded their research to the selection of fluorescent probes for POC biosensors. Similarly, while performing docking studies, researchers used target proteins based on an assumption rather than identifying them through transcriptomic analysis. It is important to note that the novelty of the current study is in integrating a transcriptomics-driven biomarker selection, fluorescence-specific ligand screening, machine learning classification, as well as molecular docking, incorporated into one pipeline in order to discover a fluorescent probe against cardiovascular diseases.

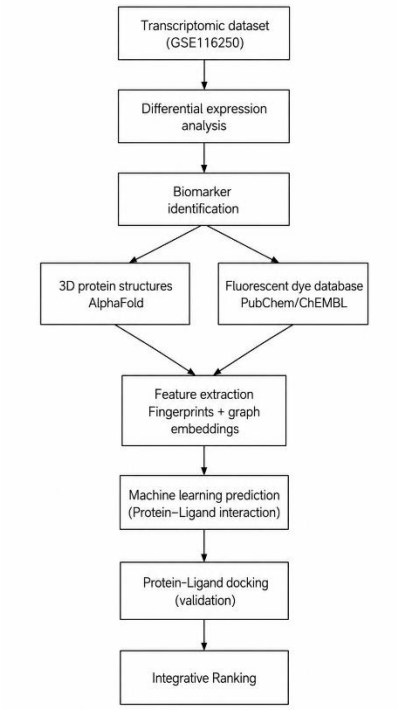

Fig 1. Integrated Workflow for Biomarker Discovery and Fluorescent Probe Identification. To detect differentially expressed proteins, we perform transcriptome analysis (GEO GSE116250). The obtained data along with the predicted protein structure (AlphaFold 3) and ligand information (PubChem/ChEMBL) serve as an input for feature extraction, where fingerprints and graph embeddings are derived. Such descriptors are utilized in the interaction predictions, and next protein–ligand docking, resulting in selection of optimal fluorescent probes.

Methodology

The process involved various stages like data collection and pre-processing of gene expression data sets associated with heart diseases from the database, biomarker identification and selection, features generation and encoding, prediction by machine learning techniques (Random Forest, SVM, and XGBoost), virtual screening and molecular docking of ligands for the selected protein biomarkers, and finally integration of docking and ADMET predictions to choose the prospective fluorescent probe molecules.

Data Acquisition and Preprocessing

Transcriptomic data regarding cardiovascular disease was downloaded from the Gene Expression Omnibus (GEO) database, specifically from the dataset GSE116250. This dataset comprises poly-A enriched RNA-seq profiles derived from 64 human left ventricle tissue samples, including 14 non-failing donor hearts, 37 dilated cardiomyopathy (DCM) cases, and 13 ischemic cardiomyopathy (ICM) cases.

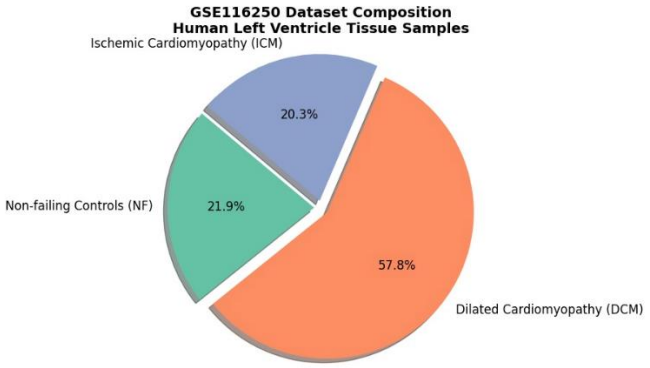

Fig 2. Composition of GSE116250 Human Left Ventricle Tissue Samples. The pie chart shows the distribution of samples in the GSE116250 dataset, comprising Dilated Cardiomyopathy (DCM, 57.8%), Non-failing Controls (NF, 21.9%), and Ischemic

American Journal of Applied Bio-Technology Research
(AJABTR)
Vol. XX, July 2026, pp. 1111-1111 DOI: xxxxxxxxxxxxxxxxxx

Cardiomyopathy (ICM, 20.3%), highlighting the predominance of DCM samples in the dataset.

Raw RPKM values were subjected to log2 transformation after cleaning the data. First, a low-expression filter retained only genes with log2 RPKM > 0 in at least 14 samples (the size of the smallest group), reducing the gene set from 57,974 to 18,881 genes. Non-coding RNAs, pseudogenes, and lncRNAs were removed using pattern-based filtering. Two contrasts were defined: DCM vs non-failing and ICM vs non-failing.

Second, limma-based differential expression analysis was performed using empirical Bayes moderation. Multiple testing correction was applied using the Benjamini-Hochberg FDR method. An adjusted p-value < 0.05 and |log2 fold change| ≥ 1 yielded 1,105 DEGs in DCM and 1,936 DEGs in ICM. After differential analysis, *Natriuretic Peptide A* or *NPPA* (also called *Atrial Natriuretic Peptide, ANP*) and *Natriuretic Peptide B* or *NPPB* (also called *B-type Natriuretic Peptide, BNP*) were among the most significantly upregulated genes in both contrasts, with logFC values ranging from 3.59 to 4.62 and adj.P.Val < 0.05 across both DCM and ICM comparisons. Established literature supports this selection as cardiac wall stress causes ANP and BNP to be secreted directly into the bloodstream. This, along with their low molecular weight, gives these two peptides potential to be ideal point-of-care biomarkers since they can be detected without need for invasive tissue biopsy. They possess epitope regions that can allow the attachment of fluorescent markers and also fall within detectable levels of fluorescence-based biosensors. Fluorescent immunoassays for BNP detection have been documented in multiple literatures, hence proving photophysical feasibility. So, the choice of ANP and BNP not only guided by gene expression analysis from dataset GSE116250, but also from literature evidence.

On the other hand, the selection of our third and final candidate biomarker *Troponin I, Cardiac Type* or *TNNI3* (*cardiac-specific isoform of Troponin I*)was mainly due to prior research. TNNI3 showed modest dysregulation in DCM (logFC = 0.44, adj.P.Val = 6.85×10^{-3}) and was below significance threshold in ICM, and was retained on clinical grounds. Cardiac Troponin I is considered the most extensively used POC cardiac biomarker globally. Unlike ANP and BNP, whose selection was largely based on their transcriptomics studies, TNNI3 was chosen primarily due to its clinical and photophysical validation in literature based on their usefulness as an analyte at the protein stage in the blood stream. Critically, all three proteins are soluble, circulating, and detectable at nanomolar concentrations- properties that are prerequisites for fluorescence-based POC assay development.

We downloaded the 3D structures of the biomarker proteins from the AlphaFold 3 Protein Structure Database for structural analysis. Although the structures predicted by AlphaFold can be highly accurate and are used extensively for computational studies, there is always a chance that they will be different from those derived from experiments in certain regions. This could affect the geometrical configuration of the binding pocket. Therefore, the results of the docking obtained in this study are only hypothetical analyses and require experimental validation.

We opted for relaxed PDB structures rather than rigid ones, as relaxed ones have optimized side-chain conformations ready for docking. Preprocessing involved removal of water molecules and hetero-ligands where available, addition of polar hydrogens, and assignment of charges for software-compatible input.

In parallel, around 300 fluorescent small molecules were also chosen from PubChem and ChEMBL chemical databases using keyword queries regarding fluorescence and dye activity. Canonical SMILES strings were compiled and each molecule standardized removing duplicates or invalid molecules. Cheminformatics tools transformed the SMILES into 3D conformers, energy-minimized for stable conformer, and output as PDB files. All protein and ligand molecules were converted as PDBQT files for dock analysis.

Feature Extraction and Representation

American Journal of Applied Bio-Technology Research
(AJABTR)
Vol. XX, July 2026, pp. 1111-1111 DOI: xxxxxxxxxxxxxxxxxx

To facilitate machine learning–based predictive modeling, we needed to convert each fluorescent molecule from its chemical representation into numerical features preserving relevant structural and physicochemical information. Two complementary methods were adopted:

Molecular Fingerprints

Canonical SMILES representations of all 300 fluorescent dyes selected were used to build Extended Connectivity Fingerprints (ECFP4) and MACCS structural keys[7]. These fingerprints represent each molecule as a binary vector, in which each digit is an indicator of the presence or absence of a specifically defined chemical substructure or topological surroundings. The ECFP4 circular fingerprints represent local neighborhoods of atoms up to four bonds in radius and are optimized specifically for identifying structural motifs associated with fluorescence activity. MACCS keys, on the other hand, encode a pre-defined set of 166 chemical patterns and allow for rapid screening of molecules against a standardized vocabulary of characteristics[8].

Graph Embeddings

Embedding vectors for each fluorescent compound were created using GCNs (Graph Convolutional Networks) implemented in the framework of PyTorch Geometric library[9]. Firstly, each molecule was expressed as a graph with atom as node and bonds as edges with help of canonical SMILES and RDKit library. Each node was characterized by three features: atom type (one-hot encoded, i.e. H = 0, C = 1, N = 2, O= 3, s = 4, P = 5, F = 6, CI = 7, Br = 8, I = 9), atom degree, and atom charge. In order to have symmetrical edges, bonds were presented as bi-directional. Finally, after passing the data into the SimpleGCN neural network that consists of Graph Convolutional layers (GCNConv)[10] and Global Mean Pooling layer, 64-dimensional embedding vectors were obtained as output of the network. As model trained only in the inference mode, the described architecture can be considered as an unsupervised molecular graph encoder. These embeddings represent structural information on the atomic level including connectivity, aromaticity, hybridization, number of bonds, among others. All the representations (both fingerprint-based and graph-based) were joined into the common feature matrix considering both interpretability and expressiveness, with topological and substructural patterns conserved.

Component	Details
Framework	PyTorch Geometric
Graph construction	RDKit (MolFromSmiles)
Node features	Element type (integer encoded), atomic degree, formal charge
Edge representation	Bidirectional bonds
GNN layer type	Graph Convolutional Network (GCNConv)
Pooling method	Global mean pooling
Output dimensions	64 per molecule
Mode	Inference (eval, no gradient)
Input molecules	300 fluorescent compounds

Table 1: Architecture and implementation details of the GNN pipeline

Machine Learning Framework

The aim of the ML model was to predict the likelihood of a given fluorescent molecule which behave as an efficient probe for the discovery of cardiovascular biomarker. It was necessary to classify and rank by uniting chemical feature data and supervised learning techniques.

Dataset Preparation

The union of significant protein-coding DEGs from both DCM and ICM comparisons formed the final feature set of 1,500 genes that we used for the ML pipeline.

Algorithm Applied

Random Forest was employed as the primary

classification framework for distinguishing between DCM, ICM, and non-failing control samples based on this DEG-filtered transcriptomic feature set. To validate and benchmark this approach, Support Vector Machine with RBF kernel and XGBoost were additionally implemented under identical experimental conditions, enabling a rigorous multi-algorithm comparison.

Pre-processing Steps and Feature Engineering

All three models were trained on the final feature matrix comprising 1,500 DEG-selected protein-coding genes across 64 samples, using stratified 5-fold cross-validation with balanced class weights to account for the inherent class imbalance between DCM (n=37), ICM (n=13), and non-failing control (n=14) groups.

Training and Validating the Algorithm

The Random Forest model was setup with 500 estimators and balanced class weights. SVM used an RBF kernel, setting C=1.0, having balanced class weights, and calculating probabilities to allow computing ROC-AUC. Finally, XGBoost used 300 estimators, learning rate=0.05, maximum tree depth of 3, subsample ratio=0.8, and column subsampling of 0.8 to make the trees shallower and less complex relative to the small dataset size. All the models were trained and tested using stratified 5-fold cross validation, with mean and standard deviation reported across folds for accuracy, F1 macro, and weighted ROC-AUC.

Performance Evaluation

The SVM yielded the best performance on classification metrics (accuracy = 90.6% ± 5.7%, F1 macro = 0.876 ± 0.080, ROC-AUC = 0.960 ± 0.057) followed by XGBoost (accuracy = 82.8% ± 5.7%, F1 macro = 0.710 ± 0.122, ROC-AUC = 0.914 ± 0.080) and Random Forest (accuracy = 82.8% ± 5.7%, F1 macro = 0.698 ± 0.134, ROC-AUC = 0.939 ± 0.048). While SVM achieved the highest classification performance, the convergent ROC-AUC above 0.91 across all three models confirms that the disease signal captured by the RF-based framework is genuine and robust, independent of the choice of classifier.

The confusion matrix(Fig 3) indicated that there is an adequate ability of the classifier to classify different disease classes but not a good ability to classify specific cardiac subtypes. For instance, the model exhibited a lower predictive ability for one disease subgroup indicating that the chosen biomarkers can only measure general cardiovascular stress responses.

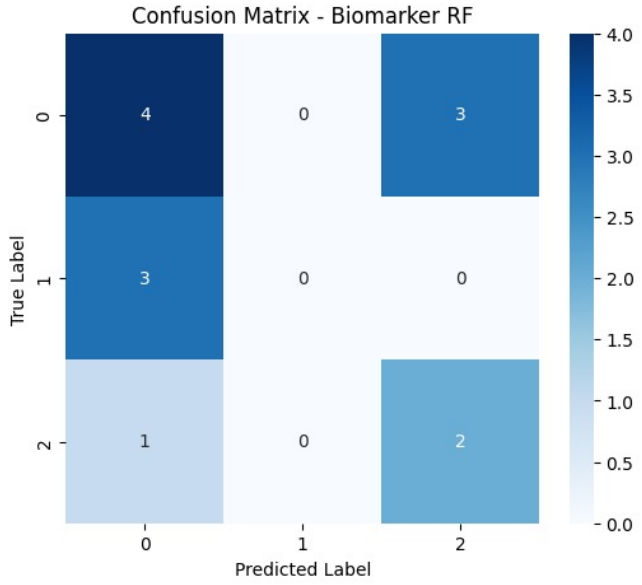

Fig 3. Confusion Matrix of Biomarker-Focused Random Forest Classifier. Classification results for the Random Forest model using three biomarker genes *(ANP, BNP, TNNI3)* that have been selected for classification using the test data set (n=13). The rows in the table below correspond to the true class labels while the columns to predicted class labels where 0 – Non-failure, 1 – DCM and 2 – ICM.

Heterogeneous discrimination ability was evident from the multiclass ROC test, with higher classification accuracy in Class 2 with AUC around 0.80 while others had comparatively lower AUC. Precision recall test was also characterized by heterogeneity in predictive accuracy, with mean precision being 0.62, 0.38, and 0.63 for classes 0, 1, and 2, respectively.

All three models showed a perfect training accuracy, which is characteristic of overfitting due to the small training set of size n=64 and high feature dimensionality. The generalization capabilities of all the three models were tested

using 5-fold stratified cross validation that allowed for an unbiased estimation of the model's ability on the test data.

Feature Importance and Biomarker Analysis

Feature importance was computed using Random Forest and its Gini coefficient difference, as well as XGBoost's importance based on its gain and SHAP values. Permutation-based importance was not applied to SVM, as this approach is known to be unreliable for kernel-based models in high-dimensional, small-sample settings.

Among the top-ranking SHAP genes were CCDC69, NQO1, ADAMTSL5, and FNDC1, all of which were independently found to be significant DEGs via the limma method, thus corroborating the independent feature selection capabilities of the ML and statistical techniques. Nevertheless, the computational ranking of these genes was not enough to include them into the list of potential fluorescent probes, since a number of them do not have an experimentally determined protein structure, validated clinical application, or circulating forms required for constructing a fluorescence-based POC test. The machine learning classification model was used rather to confirm the biological validity of the gene expression data set. This way, NPPA and NPPB were selected for developing fluorescent probes due to the constant upregulation of the two genes in the transcriptomic data, their ability to circulate, and their existing clinical applications, while TNNI3 was selected due to its diagnostic significance, as outlined previously.

Protein-Ligand Docking and Integrative Ranking

While ML models provided fluorescence applicability predictions, protein–ligand docking was required to validate and quantify physical interactions between candidate dyes and biomarker proteins.

Protein Preparation

The biomarker proteins, *Natriuretic Peptide A* (*NPPA/ANP*), *Natriuretic Peptide B* (*NPPB/BNP*), and *Cardiac Troponin I* (*TNNI3*), were extracted from the AlphaFold Protein Structure Database in relaxed PDB format. These structures were subsequently processed for docking by eliminating water molecules and any non-essential ligands, before adding polar hydrogens and assigning Gasteiger charges. The resulting structures were finally saved in PDBQT format for further docking studies.

Ligand Preparation

Approximately 300 fluorescent dyes were converted from their SMILES notation to 3D PDB structure, then optimized using molecular mechanics, protonated, and converted to PDBQT format. Hydrogens were made explicit for proper ionization and docking requirements. The library of ligands used was relatively small, but the selection of the compounds included was primarily guided by their structural variety and applicability in cardiovascular biomarker imaging applications. In an effort to prevent overfitting and bias when analyzing the data with machine learning algorithms, several approaches including normalization, reduction of redundancies, stratified sampling and cross-validation were applied. However, the relatively small dataset used is definitely one drawback that should be considered in the future research involving fluorescent ligands.

Docking Simulations and Validation

AutoDock Vina (v1.5.7) software was used to perform docking studies[11-13]. In an initial collection of 300 potential ligands, selection of 50 ligands for docking was based on the following characteristics: (1) association with pathways related to cardiovascular diseases, (2) biological action against target proteins, (3) drug-likeness verified using Lipinski's Rule of Five, and (4) exclusion of structurally redundant ligands to guarantee diversity of chemicals in the final set. The rest of the 250 ligands were not considered because of poor drug-likeness features, redundancy of structure, or no known interaction with the selected cardiovascular targets. Docking studies were carried out for all pairs of biomarkers-ligands, where the docking grid was designed in such a way that it encompassed the predicted active or functional sites calculated using AlphaFold prediction or known from the literature. The structure with minimal binding free energy

(kcal/mol) was considered as the representative binding mode for each ligand.

Parameters for docking grid were set up independently for each target protein using AutoDock MGLTools (v1.5.7). Since experimentally validated ligand-binding pockets for the AlphaFold predicted structure of ANP, BNP, and TNNI3 were absent, experiments were performed by adopting a blind docking strategy which would help identify energetically feasible binding modes without bias. Search space for the docking study was defined by a 40 × 40 × 40 Å grid box centered at the geometric centroid of each protein structure as determined by AutoDock MGLTools. Grid box centers were: x = −8.100, y = −0.122, z = 10.495 Å (for ANP); x = −1.388, y = −6.267, z = 9.342 Å (for BNP); and x = 7.630, y = 7.736, z = −16.384 Å (for TNNI3). An exhaustiveness value of 8 and energy range of 3 kcal/mol were applied uniformly across all docking runs, consistent with standard AutoDock Vina settings for small-molecular screening. Considering that ANP, BNP, and TNNI3 serve as peptide hormones and structural cardiac proteins, this blind docking approach is particularly appropriate as no small-molecule binding pockets of these proteins have been crystallographically as of yet.

The docking procedure was validated by the use of Fluorescein 5-isothiocyanate (FITC), a known and characterized fluorophore that is part of the fluorophore screening library, which was independently redocked against the target biomarker proteins, TNNI3, ANP, and BNP, using identical grid parameters and exhaustiveness settings The absence of a co-crystallized experimental structure for these biomarker targets required the evaluation of the docking procedure through an analysis of top ranked binding scores and inter-pose RMSD values.

Fluorescent Ligand ML Framework

A machine learning model under the supervision category was developed independently on the fluorescence ligand data set to predict and prioritize binding. The Random Forest model was trained based on the feature matrix created using ECFP4 fingerprints, MACCS keys, and graph embeddings mentioned above, consisting of 500 decision trees with seed 42.

The fluorescence activity of each molecule was not considered as a continuous value, but it was determined during the database curation process. ~300 candidate drugs were collected via searches in PubChem and ChEMBL database through keywords related to fluorescence and dye activity, limiting the drug library to only those compounds that were known to be fluorescent. Further, within this subset, ML classifier labels (Label = 1 or 0) were generated from AutoDock Vina scoring results, where the best candidates (docking score ≤ -7.0 kcal/mol) were labeled as Label = 1 and the others as Label = 0. Hence, the machine learning algorithm was trained to predict the binding affinity in molecules which have been confirmed to be fluorescent.

Scoring, ADMET Analysis, and Statistical Evaluation

The interactions and binding affinities were evaluated to see the most significant interactions among the target proteins, which are the ANP, BNP, and TNNI3 biomarkers. All the fifty shortlisted fluorescent ligands went through ADMET and drug-likeness analysis using parameters provided by SwissADME. These parameters include gastrointestinal absorption, bioavailability score, blood-brain barrier permeability, number of violations from the Lipinski rule, number of PAINS and Brenk alerts, lead-likeness violations, and synthetic accessibility. In order to determine the ranking of ligands on the basis of priorities, the method of normalized weighted ranking was used for assessing docking properties and ADMET properties. Binding affinity properties of docking and pharmacological properties were normalized using the methodology of Min-Max normalization, followed by determination of weight ranking score to verify whether any ligand has biomarker binding affinity and drug-like nature. Hierarchy clustering and heatmap comparison were performed using normalized docking scores along with some physicochemical properties like Consensus Log P, TPSA, molecular weight, synthetic accessibility, and ADMET total scores. Using this procedure, it was determined that the ligands showing similar behavior in terms of pharmacology and biomarker binding affinity with ANP, BNP, and TNNI3

American Journal of Applied Bio-Technology Research
(AJABTR)
Vol. XX, July 2026, pp. 1111-1111 DOI: xxxxxxxxxxxxxxxxxxx

biomarkers and dissimilar binding affinities with respect to these three biomarkers were determined. Docking scores were used as orthogonal validation of ML predictions, so shortlisted molecules had not only favorable fluorescence predictions but also physically plausible biomarker protein binding interactions.

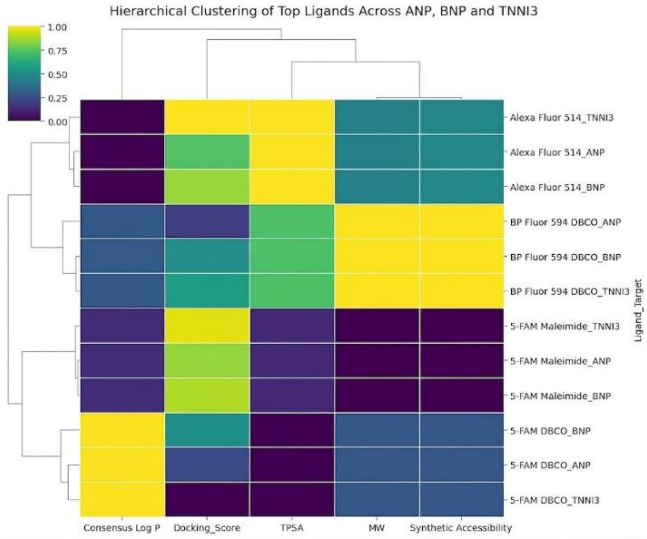

Fig 4. Heatmap of hierarchical clustering of the top-ranked ligands in terms of docking and physicochemical parameters. A hierarchical clustering heatmap displaying the similarities among the top ligands (*5-FAM DBCO*, *BP Fluor 594 DBCO*, *5-FAM Maleimide*, and *Alexa Fluor 514*) across the *ANP*, *BNP*, and *TNNI3* targets based on normalized docking scores and specific ADMET parameters (Consensus Log P, TPSA, Molecular Weight, and synthetic accessibility).

Integrative Ranking

To ensure that only the best performing ligands are ranked higher based on their fluorescence, the scores obtained from the docking studies and the ADMET pharmacokinetic parameters were integrated within a weighted scoring system based on Min–Max Normalization. The ligands with high binding ability along with good drug-likeness and bioavailability properties had high scores. Moreover, hierarchical clustering and heatmap were used to analyze and visualize similarities between ligands based on their normalized docking scores and physicochemical properties for all three biomarkers.

The combined integrative score *Si* for each ligand *i* was calculated using the following formula:

$$S_i = w_1 \cdot D_i + w_2 \cdot A_i$$

with D_i and A_i being the docking and ADMET scores normalized using the Min–Max normalization approach, respectively. Since the purpose of this research is to identify high-affinity fluorophore binders rather than their ADMET profile, the docking affinity receives a greater weight (w_1 = 0.6, w_2 = 0.4).

Validation of Random Forest Predictions Against AutoDock Vina

In order to measure the inverse relationship between machine learning predictions and docking scores, Spearman and Pearson correlation coefficient calculations were performed for the binding affinities of AutoDock Vina and the predicted binding affinities of Random Forest prediction in the training set for each biomarker target. Strong correlations between AutoDock Vina docking score and the predicted Random Forest binding score were found for all three target cardiovascular biomarkers(Fig 4).

Results

Re-docking of the reference fluorophore FITC against the three biomarker proteins validated the effectiveness and reproducibility of the docking protocol. Independent docking runs resulted in identical top-ranked affinity scores for TNNI3 (−6.7 kcal/mol), ANP (−6.3 kcal/mol), and BNP (−6.0 kcal/mol). In turn, an assessment of the inter-pose RMSD showed that the second-best alternative binding modes for TNNI3, ANP, and BNP occurred at distances of 2.614 Å, 9.200 Å, and 3.882 Å from the top-ranked binding modes, respectively. Hence, in this particular case, a distinct convergence of the predicted binding modes seemed evident, especially for the ANP target protein. Considering the comparatively large RMSD values in relation to the other biomarkers, it

Protein	First Dock Pose 1	Re-dock Pose 1	Reproducibility	Nearest alternative mode (RMSD)
TNNI3	-6.7 kcal/mol	-6.7 kcal/mol	Identical	2.614 Å (pose 3)
ANP	-6.3 kcal/mol	-6.3 kcal/mol	Identical	9.200 Å (pose 3)
BNP	-6.0 kcal/mol	-6.0 kcal/mol	Identical	3.882 Å (pose 3)

is possible to suggest that in this instance, the secondary binding mode is rather clearly

S.No	Name of the Compound	Binding Affinity (kcal/mol)
1	5-Fam dbco	**-9.1**
2	BP Fluor 594 DBCO	**-8.2**
3	FAM hydrazide, 5-isomer	-7.3
4	Alexa Fluor 514 para-isomer	-6.9
5	5-FAM Maleimide	-6.7
6	ZL-12A probe	-6.6
7	Biotinylated Cholesterol	-6.4
8	Cy3-NHS ester	-6.3
9	C.I. Acid red 29, disodium salt	-6.2
10	Bromophenol Blue sodium salt, Technical grade	-6.2
11	US11142505, Fam-48	-6.1
12	FAM amine, 6-isomer	-6.1
13	Acid Blue 29, Dye content 40%	-6.0
14	CXCR4 probe 1	-6.0
15	Dichlorofluorescein diallyl ether	-5.9
16	BP Fluor 488 amine	-5.9
17	Sulfo-Cy3 Boc-ethylenediamine	-5.9
18	dsGreen, solid form, PCR grade	-5.9
19	BP Fluor 555 Maleimide	-5.8
20	Alexa Fluor 488 para-isomer(2-)	-5.8

distinguishable. Overall, the obtained results prove the efficiency of the applied docking protocol for detecting reproducible small molecule binding poses for all of the three biomarker proteins.

Table 2: Re-docking validation results for representative biomarker targets

The 4 highest ranking ligands consistently on the top for all the three biomarkers, ranked by their predicted affinity, were:

- *BP Fluor 594 DBCO*
- *5-FAM DBCO*
- *Alexa Fluor 514 para-isomer*
- *5-FAM Maleimide*

For Cardiac Troponin I (TNNI3), *5-FAM DBCO* was predicted as the strongest binding ligand with a predicted affinity of -9.1 kcal/mol, followed by *BP Fluor 594 DBCO*, *Alexa Fluor 514 para-isomer* and *5-FAM Maleimide*. A clear negative correlation is observed in the scatterplot graph between docking scores and predicted binding values. As docking score becomes more negative for a ligand, the predicted binding value increases. The top 20 ligands and their binding affinities for TNNI3 have been shown in Table 3.

Table 3. Top 20 ligands docked against TNNI3 (Troponin I Type 3 (Cardiac)) biomarker protein ranked by docking score.

For Natriuretic Peptide B (BNP/NPPB), *BP Fluor 594 DBCO* was identified as the best binding candidate showing an affinity of -8.5 kcal/mol. The scatterplot graph shows a similar negative correlation between docking scores and predicted binding values, however there is a bit more variation compared to TNNI3. The top 20 ligands and their binding affinities for BNP have been shown in Table 4.

Table 4. Top 20 ligands docked against BNP (B-type Natriuretic Peptide) biomarker protein ranked by docking score.

S.No	Name of the Compound	Binding Affinity (kcal/mol)
1	BP Fluor 594 DBCO	**-8.5**
2	5-Fam dbco	**-8.4**
3	5-FAM Maleimide	-6.9
4	Alexa Fluor 514 para-isomer	-6.8
5	Acid Blue 29, Dye content 40%	-6.6
6	ZL-12A probe	-6.5
7	FAM hydrazide, 5-isomer	-6.5
8	C.I. Acid red 29, disodium salt	-6.4
9	Cy3-NHS ester	-6.1
10	Sulfo-Cy3 Boc-ethylenediamine	-5.9
11	US11142505, Fam-48	-5.8
12	CXCR4 probe 1	-5.7
13	Dichlorofluoroscein diallyl ether	-5.7
14	Alexa Fluor 488 para-isomer(2-)	-5.7
15	Bromophenol Blue sodium salt, Technical grade	-5.6
16	FAM amine, 6-isomer	-5.6
17	Sulforhodamine B, Dye content 75%	-5.5
18	dsGreen, solid form, PCR grade	-5.4
19	Azide MegaStokes dye 735	-5.4
20	BP Fluor 488 amine	-5.3

Finally for Natriuretic Peptide A (ANP/NPPA), *BP Fluor 594 DBCO* (affinity = -8.7 kcal/mol), *5-FAM DBCO* and *Alexa Fluor 514 para-isomer* stood out as the highest-ranking ligands with strong predicted affinities. The resulting graph shows inverse linear relationship with most data points near the regression line. A few outliers were observed as exceptions which might indicate alternative binding modes or photophysical properties that weren't fully captured by docking. The top 20 ligands and their binding affinities for ANP have been shown in Table 5.

Table 5. Top 20 ligands docked against ANP (Atrial Natriuretic Peptide) biomarker protein ranked by docking score.

S.No	Name of the Compound	Binding Affinity (kcal/mol)
1	BP Fluor 594 DBCO	**-8.7**
2	5-Fam dbco	**-8.7**
3	Alexa Fluor 514 para-isomer	-7.3
4	FAM hydrazide, 5-isomer	-6.9
5	5-FAM Maleimide	-6.9
6	Biotinylated Cholesterol	-6.8
7	FAM amine, 6-isomer	-6.6
8	ZL-12A probe	-6.4

American Journal of Applied Bio-Technology Research
(AJABTR)
Vol. XX, July 2026, pp. 1111-1111 DOI: xxxxxxxxxxxxxxxxxxx

9	Cy3-NHS ester	-6.4
10	C.I. Acid red 29, disodium salt	-6.2
11	BP Fluor 555 Maleimide	-6.2
12	Acid Blue 29, Dye content 40%	-6.2
13	Sulforhodamine B, Dye content 75%	-6.1
14	6-FAM-PEG3-Azide	-6.0
15	Sulfo-Cy3 Boc-ethylenediamine	-6.0
16	Alexa Fluor 488 para-isomer(2-)	-5.8
17	US11142505, Fam-48	-5.7
18	dsGreen, solid form, PCR grade	-5.7
19	Dichlorofluoroscein diallyl ether	-5.6
20	MK-2 Dye	-5.6

Taken together, these results show that integrating docking with machine learning provides a consistent and combined approach for identifying high affinity biomarker probes.

Table 6. ADMET profiling of the shortlisted fluorescent ligands (5-FAM DBCO, BP Fluor 594 DBCO, 5-FAM Maleimide and Alexa Fluor 514 para-isomer) detailing important physicochemical, pharmacokinetic, drug-likeness and synthetic accessibility parameters.

Ligand Name	MW	TPSA	Consensus Log P	BBB permeant	Lipinski #violations	Bioavailability Score	PAINS #alerts	Brenk #alerts	Synthetic Accessibility
5-FAM DBCO	634.63	125.4	4.85	No	1	0.55	0	1	4.88
BP Fluor 594 DBCO	980.09	233.35	2.5	No	2	0.11	1	3	6.95
5-FAM Maleimide	498.44	142.47	2	No	0	0.55	0	1	4.07
Alexa Fluor 514 para-isomer	713.69	275.5	1.56	No	2	0.11	0	3	5.44

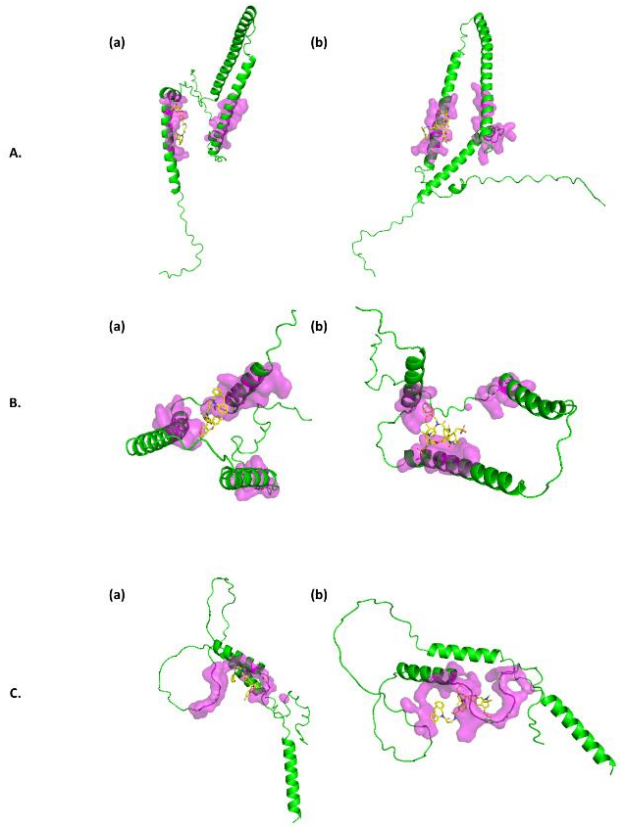

Fig 5. Protein–ligand docking poses for the top predicted fluorescent ligands with cardiovascular biomarker proteins. (A) TNNI3, (B) BNP, and (C) ANP, each docked with (a) *5-FAM DBCO* and (b) *BP Fluor 594 DBCO*. Proteins shown in green (cartoon), binding pockets in magenta (surface), and ligands in yellow (sticks).

For TNNI3, the calculated value of the Spearman's r was -0.904, $p = 2.20 \times 10^{-19}$ and Pearson's r was -0.977, $p = 4.94 \times 10^{-34}$. The Spearman's r for BNP was -0.961, $p = 1.69 \times 10^{-28}$ and Pearson's r was -0.981, $p = 1.26 \times 10^{-35}$. As for ANP, the value of Spearman's r was -0.952, $p = 3.29 \times 10^{-26}$, whereas the Pearson's r was -0.980, $p = 3.54 \times 10^{-35}$. This proves that ligands having high binding affinity scores predicted by Random Forest classifier showed good docking interactions with the target protein structures.

American Journal of Applied Bio-Technology Research
(AJABTR)
Vol. XX, July 2026, pp. 1111-1111 DOI: xxxxxxxxxxxxxxxxxx

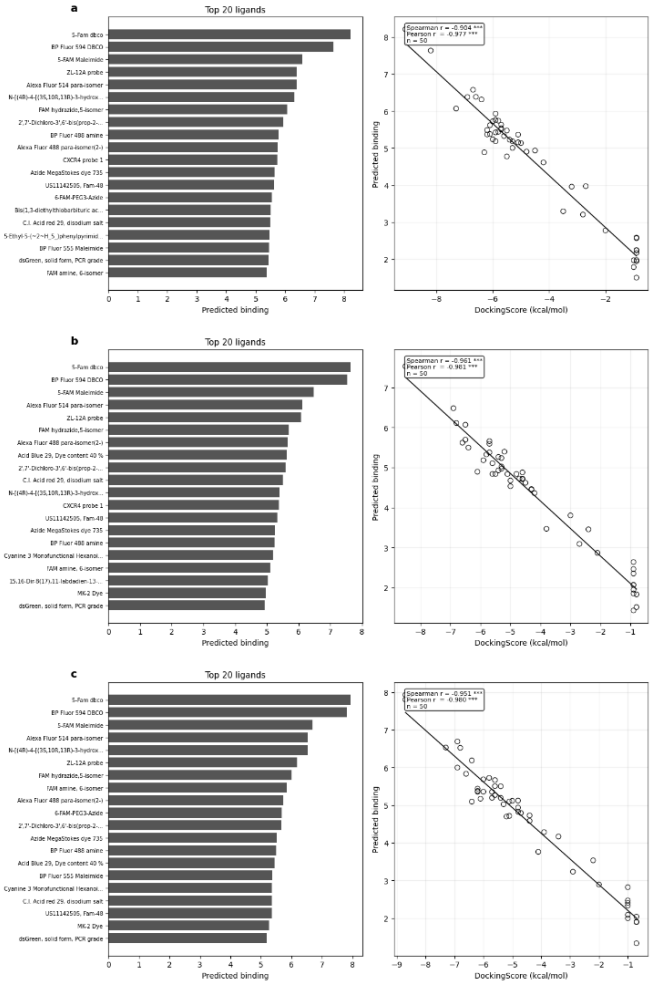

Fig 6. Comparative ML and Docking-Based Ranking of Top 20 Ligands for *TNNI3* (a), *BNP* (b), *and ANP* (c). (a) Bar charts (left) show ML-predicted binding scores; scatterplots (right) illustrate the inverse correlation between docking scores (kcal/mol, x-axis) and ML-predicted binding values (y-axis), where increasingly negative docking scores correspond to higher predicted binding, confirming concordance between the two independent methods. *5-FAM DBCO* and *BP Fluor 594 DBCO* rank highest across all targets.

Discussion

Analysis of ligands of high score and entire pipeline accuracy:

The effectiveness of the integrated pipeline is indicated by the correlation between the docking scores and ML predictions for all three protein targets. In case of TNNI3 target, the accordance between the low binding energy and high binding prediction indicates that the use of structural docking and machine learning to score Troponin I probes is consistent.

For BNP, although consistent results were achieved in both docking method and ML predictions, it can be seen that there is some form of disparity in relation to the higher variance in the scatterplot relative to that of TNNI3. This implies that the ligands have some form of fluorescence, a feature that cannot be completely attributed to binding energy alone. Therefore, the dual assessment process becomes vital when integrating the docking method and machine learning scores.

In the case of ANP, the inverse relationship shown between the two variables in the scatterplot shows the effectiveness of the scoring strategy used in the study in relation to identifying ANP biomarkers. More insight can be gained by analyzing the violin plots.

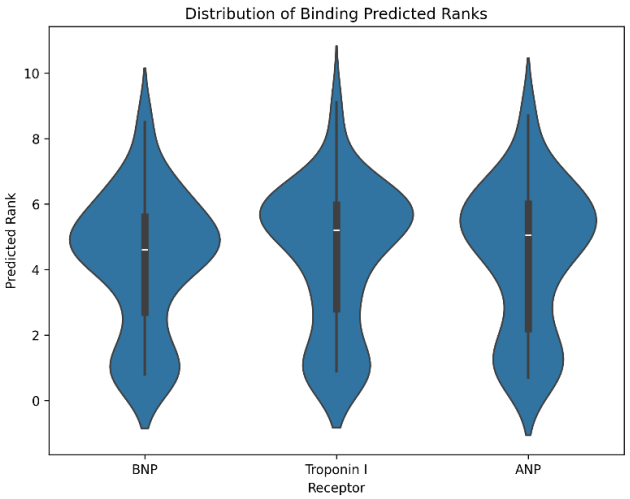

Fig 7. Distribution of Predicted Binding Ranks For BNP, TNNI3, and ANP. All three receptors share similar median ranks (~5) but differ in distribution shape, reflecting target-specific differences in ligand binding behavior.

Despite having similar general binding potential among the three receptors, as evident from their median rankings (~5), the pattern of distributions indicates different behaviour of the receptors in terms of ligand binding affinity. The presence of two peaks on the histogram of *BNP* implies the existence of two types of ligands with either strong or moderate binding potential, implying the presence of receptor specificity in ligand binding sites. The wide dispersion of estimates in the mid and high ranges of ranking values in *ANP* suggests a higher degree of variance in ligand binding

affinities, which might be due to the flexibility of the binding site and photophysical properties of the probes. On the other hand, *Troponin I* showed comparatively similar binding affinity of the probes based on their median rankings.

Both lead fluorophores from this work, *BP Fluor 594 DBCO* and *5-FAM DBCO*, provide complimentary photophysical properties appropriate for biosensor applications for cardiac markers. The *BP Fluor 594 DBCO* belongs to the rhodamine group of fluorescent dyes similar in structure to the Alexa 594 dyes, with their absorption maximum in the 590-594 nm range and emission maximum in the red-orange spectrum. Rhodamine-type dyes of this group are known to be insensitive to pH changes in the 4-10 range and therefore reliable probes in physiologically changing environments[14]. Experimental investigations have also shown that rhodamine analogs like DyLight 594 show increased photostability and higher conjugate quantum yields in comparison to Alexa 594-IgG [15].

The molecule *5-FAM DBCO* is a member of the fluorescein/carboxyfluorescein dye group, in which excitation of the parent dye is at about 490 nm with emission around 515 nm, fitting with excitation by the common 488 nm laser in clinical diagnostic instrumentation[16; 17]. As is common in dyes belonging to the fluorescein family, the fluorescein dianion has very high photoluminescent efficiency, with a quantum yield of about 0.93 and a molar extinction coefficient of about 76,900 M^{-1} cm^{-1} at 490 nm[17], which would allow for highly sensitive fluorescent detection of biomarkers. An important factor to keep in mind is that fluorescein has a pKa for monoanion/dianion transition of about 6.4. Thus, at physiological pH 7.4, almost 90% of the dye molecules will be in the luminescent dianion form, although fluorescence intensity still depends on pH changes in the neighborhood of 6.4[16].

When considered collectively, both *BP Fluor 594 DBCO* and *5-FAM DBCO* have complementary strengths. The first candidate possesses better photostability and independence of pH and is ideal for use in diagnostic purposes, whereas the latter candidate is bright and is compatible with commonly used instrumentation that utilizes the green channel. In terms of their suitability, both of these options need characterization in physiological buffer solutions to determine factors such as quantum yield, photobleaching dynamics, and pH titrations due to their unique DBCO chemistry[18].

Since there were no experimentally determined co-crystals for the biomarker proteins predicted by AlphaFold, conventional re-docking using an existing crystallographic pose was not possible. However, protocol validation was conducted via independent re-docking of the fluorescent marker (FITC), resulting in an exact match for binding energies and similar binding orientations for all three target proteins, and this is consistent with validation approaches reported for other AlphaFold-docking studies.

One critical limitation with regard to the current docking approach is that it involves blind docking based on the geometric center, where the known ligand binding sites of ANP, BNP, and TNNI3 are not known. Although this technique facilitates unbiased screening of the whole protein surface, it fails to assure whether the predicted binding sites are biologically significant or not. This is because the identified binding sites obtained from the blind docking can only reflect surface accessibility but not any allostery or functionality.

Conclusion

This research paper provided a computational pipeline that incorporates both docking and machine learning approaches to systematically identify the candidates for fluorescent probes targeting three cardiovascular biomarkers, *ANP* (*NPPA*), *BNP* (*NPPB*), and *Troponin I* (*TNNI3*). Using chemically curated databases from PubChem and ChEMBL together with protein structures predicted by AlphaFold, it allowed for the screening process of the fluorescent molecules based on a balanced approach of structure binding properties and photochemical features.

The results revealed strong consistency between binding free energies calculated via AutoDock Vina and machine learning-derived binding scores across all three targets. The ligands, *BP Fluor 594 DBCO* and *5-FAM DBCO* were identified as the most promising computational candidates. ECFP4

and MACCS fingerprints, along with molecular graph representations, proved informative as descriptors within the machine learning framework.

On the whole, the above discussion shows that the use of docking combined with machine learning is feasible in terms of time and efficacy for fluorescent probe design for cardiometabolic diagnostics. However, as all findings are computational in nature, experimental validation, including characterization of photophysical properties under physiological conditions, fluorescence performance testing, and assessment of selectivity toward cardiometabolic markers, will be essential before these probes can be considered for biosensor or diagnostic applications.

Acknowledgement

The authors extend sincere gratitude to the Department of Biotechnology, School of Bio-Sciences and Technology, Vellore Institute of Technology (VIT) for providing a collaborative platform and academic support that greatly facilitated the progress of this work. The authors also thank the Biotech Research Society of India (BRSI), VIT Chapter, Faculty Coordinator Dr. Shanthi V for providing the guidance and resources essential for carrying out this research.

References

[1] Wasilewski, T., Kamysz, W., & Gębicki, J. (2024). AI-Assisted Detection of Biomarkers by Sensors and Biosensors for Early Diagnosis and Monitoring. *Biosensors*, *14*(7), 356. https://doi.org/10.3390/bios14070356

[2] DeGroat, W., Abdelhalim, H., Patel, K., Mendhe, D., Zeeshan, S., & Ahmed, Z. (2024). Discovering biomarkers associated and predicting cardiovascular disease with high accuracy using a novel nexus of machine learning techniques for precision medicine. *Scientific reports*, *14*(1), 1. https://doi.org/10.1038/s41598-023-50600-8

[3] Batta, I., Patial, R., Sobti, R. C., & Agrawal, D. K. (2024). Computational Biology in the Discovery of Biomarkers in the Diagnosis, Treatment and Management of Cardiovascular Diseases. *Cardiology and cardiovascular medicine*, *8*(5), 405–414.

[4] Mao, X., Zhao, X., Ni, Z., Xu, X., Liu, Q., & Qu, P. (2026). Harnessing fluorescent probes for molecular diagnosis and theranostics of atherosclerosis. *Drug Delivery*, *33*(1). https://doi.org/10.1080/10717544.2025.2595728

[5] Jiang, X., Luo, Y., Li, Z., Zhang, H., Xu, Z., & Wang, D. (2024). Identification of diagnostic biomarkers and immune cell infiltration in coronary artery disease by machine learning, nomogram, and molecular docking. *Frontiers in immunology*, *15*, 1368904. https://doi.org/10.3389/fimmu.2024.1368904

[6] Lin, S. L., Chen, Y. S., Liu, R. Y., Zhu, M. Y., Zhu, T., Wang, M. Q., & Liu, B. Q. (2024). Synergistic acceleration of machine learning and molecular docking for prostate-specific antigen ligand design. *RSC advances*, *14*(12), 8240–8250. https://doi.org/10.1039/d3ra08550c

[7] Boldini, D., Ballabio, D., Consonni, V., Todeschini, R., Grisoni, F., & Sieber, S. A. (2024). Effectiveness of molecular fingerprints for exploring the chemical space of natural products. *Journal of cheminformatics*, *16*(1), 35. https://doi.org/10.1186/s13321-024-00830-3

[8] Xie, L., Xu, L., Kong, R., Chang, S., & Xu, X. (2020). Improvement of Prediction Performance With Conjoint Molecular Fingerprint in Deep Learning. *Frontiers in pharmacology*, *11*, 606668. https://doi.org/10.3389/fphar.2020.606668

[9] Fey, M., & Lenssen, J.E. (2019). Fast Graph Representation Learning with PyTorch Geometric. *ArXiv*, *abs/1903.02428*. https://doi.org/10.48550/arXiv.1903.02428

[10] Kipf, T., & Welling, M. (2016). Semi-Supervised Classification with Graph Convolutional Networks. *ArXiv*, *abs/1609.02907*. https://doi.org/10.48550/arXiv.1609.02907

[11] Omar, E. D., Mat, H., Abd Karim, A. Z., Sanaudi, R., Ibrahim, F. H., Omar, M. A., Ismail, M. Z. H., Jayaraj, V. J., & Goh, B. L. (2024). Comparative Analysis of Logistic Regression, Gradient Boosted Trees, SVM, and Random Forest

Algorithms for Prediction of Acute Kidney Injury Requiring Dialysis After Cardiac Surgery. *International journal of nephrology and renovascular disease*, *17*, 197–204. https://doi.org/10.2147/IJNRD.S461028

[12] Su, X., Xu, Y., Tan, Z., Wang, X., Yang, P., Su, Y., Jiang, Y., Qin, S., & Shang, L. (2020). Prediction for cardiovascular diseases based on laboratory data: An analysis of random forest model. *Journal of clinical laboratory analysis*, *34*(9), e23421. https://doi.org/10.1002/jcla.23421

[13] Rizvi, S. M., Shakil, S., & Haneef, M. (2013). A simple click by click protocol to perform docking: AutoDock 4.2 made easy for non-bioinformaticians. *EXCLI journal*, *12*, 831–857.

[14] Panchuk-Voloshina, N., Haugland, R. P., Bishop-Stewart, J., Bhalgat, M. K., Millard, P. J., Mao, F., Leung, W. Y., & Haugland, R. P. (1999). Alexa dyes, a series of new fluorescent dyes that yield exceptionally bright, photostable conjugates. *The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society*, *47*(9), 1179–1188. https://doi.org/10.1177/002215549904700910

[15] Sarkar, P., Sridharan, S., Luchowski, R., Desai, S., Dworecki, B., Nlend, M., Gryczynski, Z., & Gryczynski, I. (2010). Photophysical properties of a new DyLight 594 dye. *Journal of photochemistry and photobiology. B, Biology*, *98*(1), 35–39. https://doi.org/10.1016/j.jphotobiol.2009.10.005

[16] Hurley, J. J. M., Meisner, Q. J., Huang, C., & Zhu, L. (2021). Hydroxyaromatic Fluorophores. *ACS omega*, *6*(5), 3447–3462. https://doi.org/10.1021/acsomega.0c04611

[17] Sjöback, R., Nygren, J., & Kubista, M. (1995). Absorption and fluorescence properties of fluorescein. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *51*(6), L7–L21. https://doi.org/10.1016/0584-8539(95)01421-P

[18] Yao, J. Z., Uttamapinant, C., Poloukhtine, A., Baskin, J. M., Codelli, J. A., Sletten, E. M., Bertozzi, C. R., Popik, V. V., & Ting, A. Y. (2012). Fluorophore targeting to cellular proteins via enzyme-mediated azide ligation and strain-promoted cycloaddition. *Journal of the American Chemical Society*, *134*(8), 3720–3728. https://doi.org/10.1021/ja208090p