JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

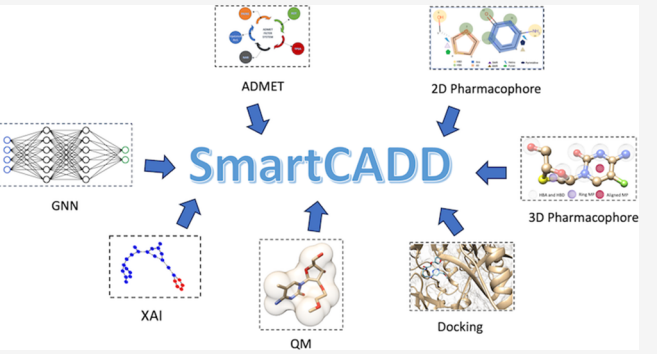
SmartCADD: AI-QM Empowered Drug Discovery Platform with Explainability

Ayesh Madushanka, Eli Laird, Corey Clark, and Elfi Kraka*

Cite This: J. Chem. Inf. Model. 2024, 64, 6799−6813 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Artificial intelligence (AI) has emerged as a pivotal force in enhancing productivity across various sectors, with its impact being profoundly felt within the pharmaceutical and biotechnology domains. Despite AI's rapid adoption, its integration into scientific research faces resistance due to myriad challenges: the opaqueness of AI models, the intricate nature of their implementation, and the issue of data scarcity. In response to these impediments, we introduce SmartCADD, an innovative, open-source virtual screening platform that combines deep learning, computer-aided drug design (CADD), and quantum mechanics methodologies within a user-friendly Python framework. SmartCADD is engineered to streamline the construction of comprehensive virtual screening workflows that incorporate a variety of formerly independent techniques—spanning ADMET property predictions, de novo 2D and 3D pharmacophore modeling, molecular docking, to the integration of explainable AI mechanisms. This manuscript highlights the foundational principles, key functionalities, and the unique integrative approach of SmartCADD. Furthermore, we demonstrate its efficacy through a case study focused on the identification of promising lead compounds for HIV inhibition. By democratizing access to advanced AI and quantum mechanics tools, SmartCADD stands as a catalyst for progress in pharmaceutical research and development, heralding a new era of innovation and efficiency.



INTRODUCTION

The complex and time-consuming nature of the drug discovery process emphasizes the growing need for new and effective drug discovery procedures in modern medicine. This urgency spans various classes of drugs, including antibiotics,[1] cancer treatments,[2] and antivirals,[3] to combat emerging threats like antibiotic resistance and rapid viral mutations. Simultaneously, artificial intelligence (AI) and machine learning (ML) have emerged as pivotal technologies in numerous fields in recent years,[4−6] with their impact being particularly pronounced in the realm of drug design and discovery, enabling more accurate and efficient computation of ADMET properties,[7,8] virtual screening,[9−11] binding free energy predictions,[12] and synthesis route planning.[13,14] However, the performance and precision of AI/ML models are significantly impacted by the availability of high-quality, well-structured data sets. This condition is often difficult to meet due to the challenges of data sparsity and privacy concerns.[15] Moreover, the interpretability and explainability of AI models stand as critical concerns.[16]

Explainable Artificial Intelligence (XAI) is an important tool in computational chemistry, particularly within the interdisciplinary fields of cheminformatics and drug discovery.[17−20] Unlike conventional black-box AI models that offer limited insight into their internal workings, XAI sheds light onto the AI's decision-making process, providing an understandable path from input data to output predictions. This transparency is crucial in scientific research, where understanding the "why" and "how" behind predictions is as important as the predictions themselves.[21] For drug screening processes, XAI methods provide a deeper understanding of the molecular features and interactions that drive the predictions of their AI models. Given the insights provided by these XAI methods, researchers can align the model with biological and chemical intuition, enhancing the fidelity of its predictions. Furthermore, XAI can uncover new knowledge and hypotheses, guiding the design of novel compounds and therapeutic strategies with a level of interpretability that accelerates discovery in the pharmaceutical domain.[22,23] Through bridging the gap between complex AI algorithms and human comprehension, XAI stands as an important part in the evolution of drug discovery.[24] A comprehensive review of XAI techniques,

Received: April 25, 2024
Revised: August 6, 2024
Accepted: August 15, 2024
Published: August 23, 2024

https://doi.org/10.1021/acs.jcim.4c00720
J. Chem. Inf. Model. 2024, 64, 6799−6813

applications and limitations is available in Supporting Information under the title "XAI Review".

We introduce Smart-Computer Aided Drug Design (SmartCADD), an open-source, virtual screening tool that utilizes the combined capabilities of AI and quantum mechanics (QM) to streamline and accelerate the process of lead identification and optimization within the drug development cycle. SmartCADD combines a diverse spectrum of screening methodologies into a unified workflow, sketched in Figure 1. SmartCADD uniquely integrates deep learning-driven screening with both classical and quantum-informed screening approaches, forming a seamless end-to-end filtration pipeline. This pipeline is modular as depicted in Figure 2, consisting of independent units that can function as isolated filters or be incorporated into a sequential screening strategy. Utilizing SmartCADD enables scientists to pinpoint promising leads more efficiently, reducing the complexity of the advanced stages of drug development.

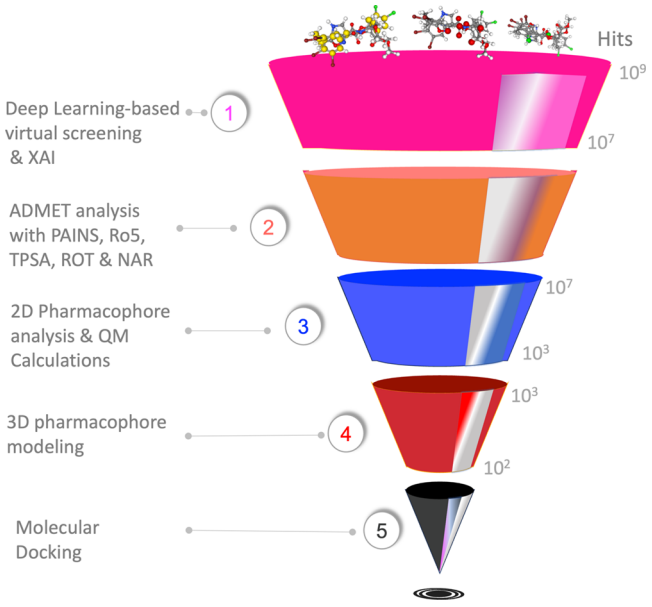


Figure 1. SmartCADD filter design flow.

METHODOLOGY

Architecture Design. The SmartCADD framework is designed using the Bridge Pattern,[25] facilitating a modular and adaptable approach to virtual drug screening. The flexibility of the Bridge Pattern allows users to customize the implementations of filters to fit their needs while maintaining the functionality of the pipeline as a whole. For example, a filter that performs molecular docking can have an implementation for both the Smina[26] and Autodock Vina[27] methods without changing the higher-level pipeline's code. This architecture enhances system flexibility by allowing customization of drug screening pipelines to accommodate various use cases, meeting the diverse needs in drug discovery. SmartCADD is implemented as an open-source Python package with installation and usage instructions available at the SMU's Computational and Theoretical Chemistry Group's (CATCO) GitHub repository.[28]

SmartCADD Interfaces. SmartCADD implements two distinct interfaces: the Pipeline Interface and the Filter Interface, Figure 2A. The Pipeline interface defines the required pipeline-specific functions that every implementation of a pipeline should implement, such as *getdata* and *runfilters*. The Filter Interface defines the required filter-specific functions, such as *filter* and *preprocess*. To put them together, a specific pipeline for a data set is made by adding filters that implement the Filter Interface to a Pipeline, as seen in Figure 2B. The flexibility of the bridge design allows users to create custom implementations of the Pipeline or Filter classes that for example handle different data types, machine learning frameworks, or multiprocessing schemes.

SmartCADD Data Readers. Simplifying the data reading process, SmartCADD provides data set loader classes, such as the *IterableData set* class that loads compounds from SMILES files and batches them into a universal *Compound* data structure. SmartCADD's practical *Compound* data structure represents a compound and its chemical and nonchemical properties, such as its SMILES string, RDKit *mol* object,[29] graph representation, ring system descriptors, and more. This data structure is used to pass compounds between filters when running the pipeline and performs any data conversions necessary for custom filter implementations. Datareader modules employ a template-based technique where the relevant reference SMILES string is loaded and used as a

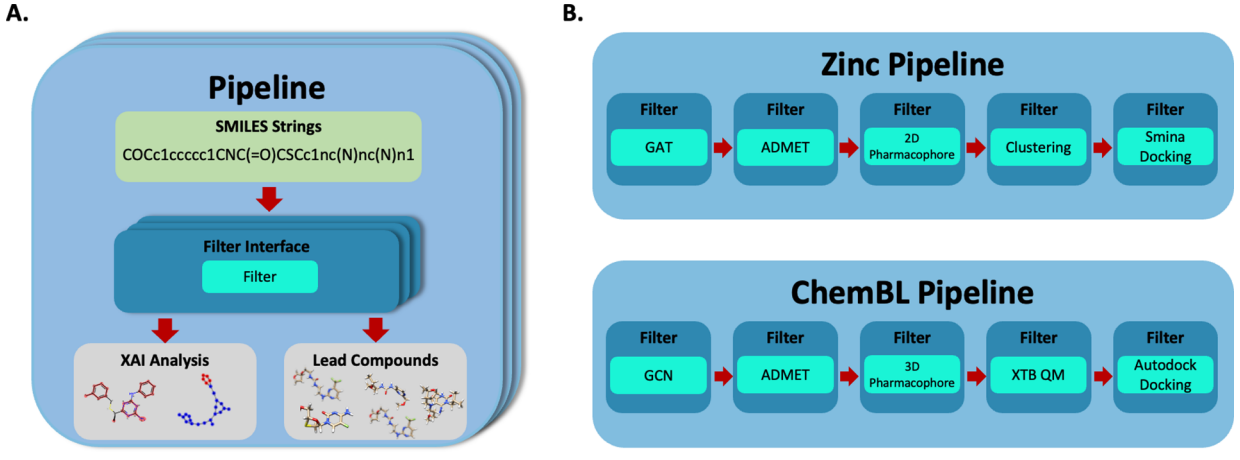


Figure 2. (A) Shows flow of SmartCADD starting with SMILES strings, filtering, and ending with final lead compounds. (B) Shows an example of two pipeline configurations designed with different filters for a specific data set.

https://doi.org/10.1021/acs.jcim.4c00720
J. Chem. Inf. Model. 2024, 64, 6799−6813

template for comparison with the loaded coordinate data (ex: PDB and XYZ). This comparison ensures that any discrepancies between the SMILES data and the coordinate data will generate an error.

SmartCADD Transform Modules. SmartCADD data packages transform and analyze functionality into a simple Module Interface. The Module Interface is intended to wrap functions that do not filter input compounds but rather perform data transforms or analysis, such as a SMILE to PDB conversion, geometry optimization, or explainable AI analysis. Some notable modules that come prepackaged in SmartCADD are the *SMILETo3D*, *XTBOptimization*, and *ExplainableAI* Modules, some of which wrap data transform capabilities from RDkit[29] and OpenBabel tools.[30]

Overview of SmartCADD Capabilities. SmartCADD integrates an extensive suite of preconfigured filters tailored for every stage of the virtual screening workflow, as illustrated in Figure 1. This comprehensive array encompasses Deep Learning-based screening filters, tools powered by ADMET filters, both 2D and 3D Pharmacophore filters, QM filters, and molecular docking filters. The following sections provide a comprehensive description of each filter set, detailing its design, operational principles, and application within the SmartCADD framework. In addition, Tautomer and protomer packages are described in Supporting Information under the section called "Tautomers and Protomers".

Deep Learning-Based Virtual Screening Filter. The Deep Learning (DL) filter is a versatile tool in cheminformatics, enabling the screening of compounds for specific target activities. By leveraging any pretrained deep learning model trained on a comprehensive proxy data set, such as MoleculeNet.[31] This filter is used to identify promising candidates from larger molecule databases such as ZINC,[32] ChEMBL[33] and PubChem.[34] As with all filters in SmartCADD, the DL filter accepts a list of *Compound* objects that contain graph representations and chemical descriptors. Implementations of this filter cast each *Compound*'s graph representation into the data type required by the wrapped deep learning model, such as a Pytorch Geometric[35] or DeepChem's[36] respective *GraphData* objects, which include initial atomic and bond features are described in Table 1. The DL filter is designed for flexibility, allowing users to wrap their preferred deep learning model using SmartCADD's ModelWrapper interface. This interface defines the *predict*, *featurize*, and *load* functions that perform predictions, data preparation, and trained weights loading from a model from any deep learning framework.

Table 1. Initial Atomic and Bond Features Included in a Compound Object

feature	description	size
atom type	type of atom (C, N, O, etc., or metal), one-hot	10
chirality	chirality (R, S, or none), one-hot	2
formal charge	integer charge	1
partial charge	computed charge	1
degree	atom's connectivity (0−5), one-hot	6
# of hydrogens	hydrogens bonded (0−4), one-hot	5
hybridization	hybridization state (sp, sp2, sp3, or none), one-hot	3
hydrogen bonding	hydrogen bond donor/acceptor, one-hot	2
bond type	bond type (single, double, triple, aromatic, or none), one-hot	4
conjugated	if bond is conjugated, one-hot	2
stereo	stereo configuration of bond, one-hot	2
same ring	if atoms are in the same ring	2
total number of features:		40

Explainable AI Analysis Module. Unlike traditional black-box AI models that lack transparency in their decision-making processes, explainable AI (XAI) offers valuable insights into the model's inner workings. This transparency allows scientists to trace the path from input data to output predictions, fostering a deeper understanding of the "why" and "how" behind predictions. In scientific research, where comprehension is as crucial as the results themselves, XAI proves invaluable.[21] While graph neural networks (GNNs) are still evolving in their ability to generate practical explainability descriptors for molecules, many advancements including GNNExplainer,[37] PGM-Explainer,[38] and SubgraphX[39] provide molecule-level descriptors, while XGNN[40] and XInsight[23] provide high-level concept descriptors.

SmartCADD integrates these XAI algorithms to address this need, enabling researchers to understand the factors driving a deep learning model's decisions. The XAI modules in SmartCADD identify and visualize specific substructures within compounds that significantly influence the model's predictions. These insights, termed explanations, allow researchers to assess the validity of their model's learning process, particularly in distinguishing genuine biochemical relationships from spurious correlations in the data set. For example, previous works[20,23,40] have used XAI to identify functional groups, such as aromatic rings, related to chemical properties like mutagenicity, highlighting the use of XAI for knowledge discovery. This capability enhances the reliability and transparency of deep learning applications in drug discovery.

ADMET Analysis Filter. In addition to high potency and selectivity, a favorable ADMET profile is crucial, ensuring safe and effective drug exposure. A drug should be effectively absorbed, distributed to target tissues, metabolized without rapid inactivation, and eliminated appropriately.[41] A drug-like molecule typically shares similar physicochemical properties with orally active drugs, a concept that guides the drug design process to ensure proper efficacy. The first standard in drug design, known as the "*Rule of 5*" (Ro5), was introduced in 1997 by Lipinski.[42] The Ro5 criteria for orally active drugs "…'*the rule of 5' predicts that poor absorption or permeation is more likely when there are more than 5 H-bond donors, 10 H-bond acceptors, the molecular weight (MWT) is greater than 500 and the calculated LogP is greater than 5.*…" based on a library of 2245 FDA approved compounds or at least in phase II clinical studies. However, the relevance of the Ro5 has been questioned by many researchers.[43−45] According to the report published by Hartung et al.,[45] which examines FDA-approved drugs from 2018 to August 31st, 2022, a notable trend emerges. Specifically, 93% of drugs have molecular weights exceeding 500 Da, with the 90th percentile reaching 588 Da. Moreover, 42% of drugs exhibit HBA violations, with the 90th percentile standing at 11. Conversely, HBD and *clogP* violations were observed in 12 and 19% of drugs, respectively.

Moreover, in a study by Veber,[46] a researcher at GlaxoSmithKline, the drug-like physicochemical space was expanded to include additional parameters such as the number of rotatable bonds and topological polar surface area. These descriptors are relevant to the drug's ability to cross cell membranes from the gastrointestinal tract, as highly flexible

https://doi.org/10.1021/acs.jcim.4c00720
J. Chem. Inf. Model. 2024, 64, 6799−6813

and highly polar molecules are typically less permeable.[47] Additionally, Ritchie and Macdonald[48] suggest that lead compounds with no more than 3 aromatic ring counts are most favored for advancement in Phase 1, 2, and Proof-of-Concept (POC) stages of the GSK pipeline.

To assess these properties, SmartCADD employs an ADMET filter that analyzes compounds based on seven key chemical attributes: molecular weight (M_w), partition coefficient (log*P*), number of hydrogen bond donors (HBD), number of hydrogen bond acceptors (HBA), topological polar surface area (TPSA), number of aromatic rings (NAR), and rotational degree of freedom (ROT). Furthermore, SmartCADD's ADMET filter incorporates a step to exclude pan-assay interference compounds (PAINS) as outlined by Baell and Holloway[49] and refined by Walters.[50] SmartCADD's ADMET feature is designed with flexibility in mind, providing users the ability to tailor ADMET parameters and choose filters that best align with their specific research requirements. SmartCADD's ADMET filter parameters are described in Table 2 and visualized in Figure 3.

Table 2. ADMET Parameters Used in ADMET Filter

ADMET parameter	range
M_w (Ro5)	[0, 600]
Log*p* (Ro5)	[−2, 5]
HBD (Ro5)	[0, 5]
HBA (Ro5)	[0, 12]
TPSA	[0, 140]
NAR	[0, 4]
ROT	[3, 12]

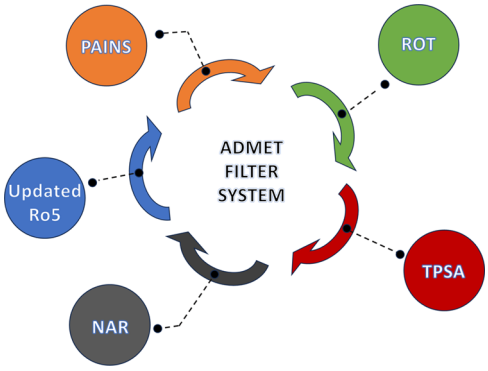


Figure 3. ADMET filter complex described in SmartCADD.

2D Pharmacophore Analysis Filter. The concept of a pharmacophore was introduced by Paul Ehrlich in the early 1900s. The term itself, however, was coined later, defining a pharmacophore as a collection of molecular features necessary for a drug's biological activity.[51] Molecular pharmacophore patterns, encompassing HBD, HBA, positive ionizable groups, negative ionizable groups, aromatic rings, and hydrophobic regions, influence a drug molecule's biological activity.[52,53] These features play a key role in many commercial and noncommercial pharmacophore modeling applications, including HipHop, HypoGen, Pharmer, PHASE, GASP, PharmaGist, PharmMapper, MOE, and LigandScout.[54] While 2D molecular descriptors like MACCS keys,[55,56] Morgan fingerprints[57,58] and Daylight fingerprint[59] are widely used as a 2D screening tools in database searches, a significant drawback is their inability to account for 3D conformation.[60] Therefore, to enhance the capabilities of 2D pharmacophore search methods, additional pharmacophore features that capture intricate details of molecular structures should be considered.[61] Our approach involves incorporating specific ring features through a four-step process. First, pharmacophore features were identified excluding any ring structures. Next, rings were categorized as aromatic or aliphatic[62,63] and separated according to the number of carbon atoms[64] they contained. Then, heterocyclic rings and their types were determined.[65] Finally, 2D pharmacophore data set was generated containing all the information per each molecule. This strategy can provide information about a molecule's shape, complementing the limitations of purely 2D descriptors.[66] A simplified visual representation of 2D pharmacophore process is presented as a four-step model in Figure 4.

3D Pharmacophore Analysis Filter with Quantum Mechanical Calculations. Modern computational drug discovery methods frequently utilize molecular docking simulations to identify potential lead compounds.[67,68] Docking simulations, despite their widespread use, have limitations. Docking scores (DS) often show inconsistencies in correlation with experimental data,[69] and these simulations typically assume a rigid protein structure. Furthermore, docking typically imprecise the critical influence of water molecules and solvation effects on ligand-protein interactions.[70] Rather than discarding compounds solely based on docking analysis, we propose utilizing a combination of scoring functions commonly used in 3D pharmacophore studies such as shape Tanimoto distances (STD),[71] shape protrude distance (SPD)[72] and align score (AS).[73] In addition to the four scoring functions, we introduce a novel scoring function that assesses compound similarity by aligning pharmacophore coordinates (HBA, HBD and center of the rings) and measuring distances between pharmacophores of the target and lead compounds. This 3D model is designed to capture five distinct distances within pharmacophores, ensuring that potential compounds are not prematurely excluded before undergoing lead optimization. Each calculation utilized QM-optimized molecules, with each molecule generating 100 different conformations to assess molecular flexibility. As illustrated in Figure 5, steps A and B sample the HBA, HBD, and middle coordinates of the rings for each conformation of the lead and the reference compounds, respectively. Then, the lead compound is aligned with the reference compound using a common structural element (for NRTIs this is the aromatic ring: Figure 5C), and calculated AS, STD and SPD scores.

Finally, To generate distance scores, we first calculate the distances between each pharmacophore feature in the reference compound and its corresponding reference point. This process is repeated for all lead compounds (Figure 5d). We then scan a range of distance cutoffs (±0.2, ±0.4, ±0.6, ±0.8, and ±1.0 Å) around the reference point in the reference compound and identify the pharmacophore features in the lead compounds that fall within these distance ranges. Finally, a scoring function is employed to assess the distribution of the identified pharmacophore features relative to the reference compound. It is important to note that distance scores were not utilized in the screening process. Table 3 documents the comprehensive set of nine 3D computational parameters.

While accurate 3D structures are crucial for virtual screening (VS), traditional QM methods like Hartree−Fock (HF) or Density Functional Theory (DFT) calculations are often too time-consuming for large data sets.[74] Molecular dynamics

https://doi.org/10.1021/acs.jcim.4c00720
J. Chem. Inf. Model. 2024, 64, 6799−6813

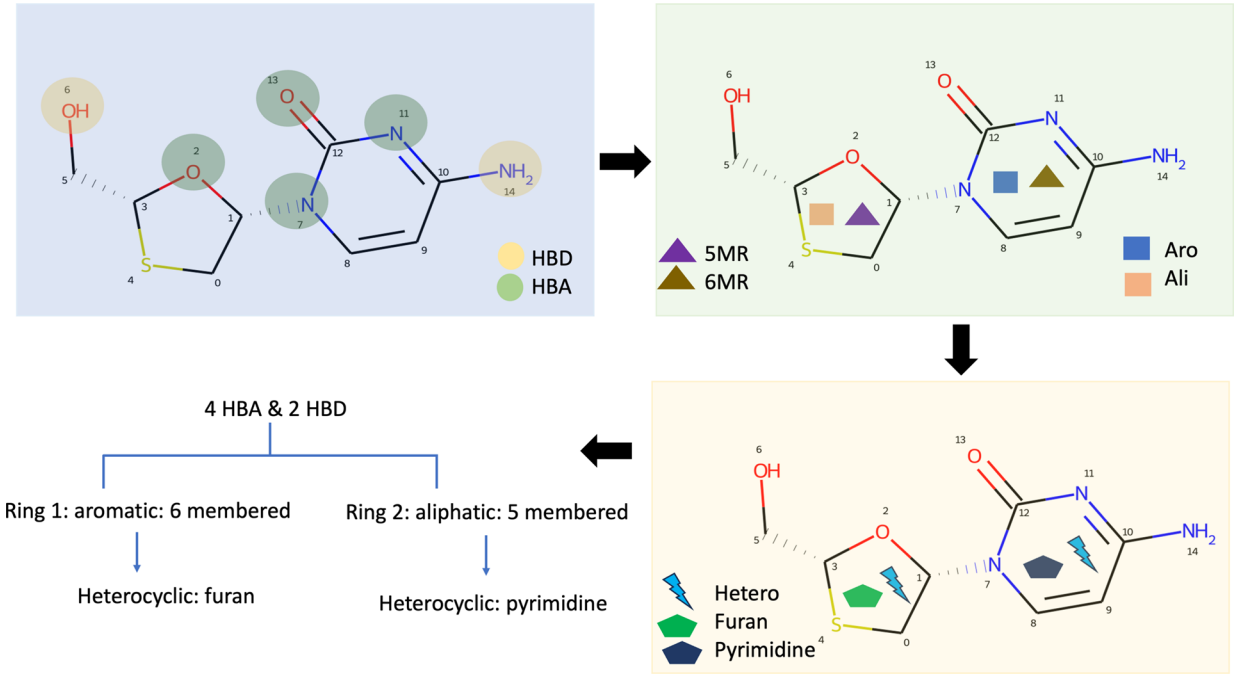


Figure 4. Four-step process of 2D pharmacophore design architecture (Aro: Aromatic ring, Ali: Aliphatic ring, MR: Membered Ring, Hetro: Heterocyclic, Furan: Furan like rings, Pyrimidine: Pyrimidine like rings).

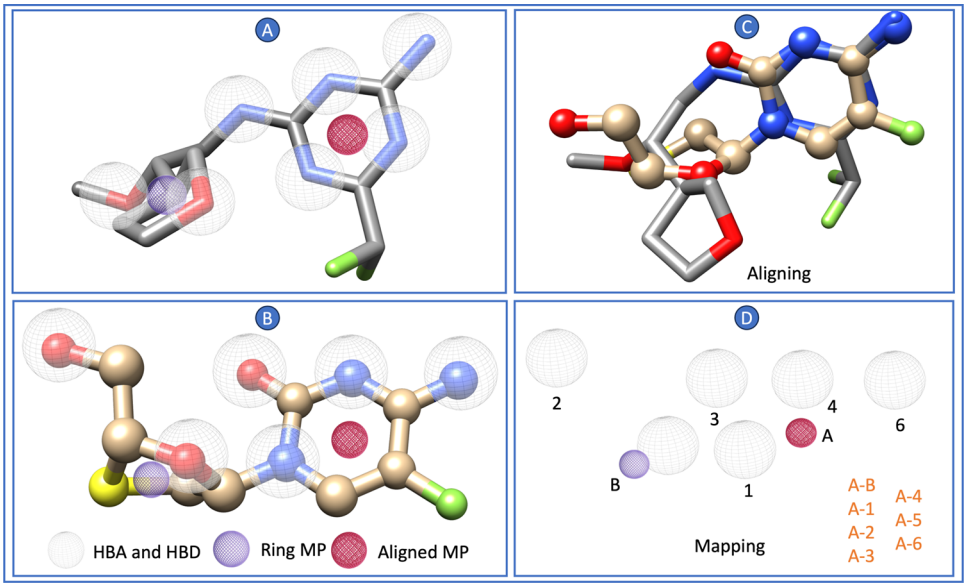


Figure 5. 3D pharmacophore designing. (A) Selected pharmacophore coordinates of ZN000872594074 compound (37th conformation). (B) Target pharmacophore coordinates. (C) Aligning of target and lead compound with pharmacophore coordinates. (D) Mapping pharmacophore distances.

(MD) simulations, though faster, have lower accuracy due to their neglect of electronic effects. A promising solution lies in semiempirical QM methods, which strike a balance between accuracy and computational cost, making them well-suited for VS applications. Over the past few years, significant strides have been made in the field of semiempirical QM calculations with the development of the Extended Tight Binding (XTB)[75] method. XTB has yielded improvements in the accuracy of quantum mechanical calculations in both small molecules[76,77] and macromolecules,[78,79] approaching the level of precision observed in experimental data in numerous cases. SmartCADD introduces a de novo 3D Pharmacophore filter, intricately combined with QM calculations. Utilizing the RDKit library, the platform converts SMILES representations into 3D molecular structures. XTB is used to optimize these molecules at the GFN2-xTB level of theory.[80] This optimization is crucial for accurately determining the coordinates of HBD and HBA, as well as the central coordinates of molecular rings, thereby enhancing the model's ability to pinpoint potential drug candidates with high accuracy.

Docking Analysis Filters. The Docking filter wraps the Smina[26] tools as default. This Filter automates protein−ligand docking, taking a list of SMILES strings and a protein structure as inputs and delivering docking scores as output. The

https://doi.org/10.1021/acs.jcim.4c00720
J. Chem. Inf. Model. 2024, 64, 6799−6813

Table 3. 3D Modeling Techniques Used in SmartCADD

	parameter
1	shape Tanimoto distances (STD)
2	shape protrude distance (SPD)
3	align score (AS)
4	docking score (DS)
5	distance score within 0.2 (0.2d)
6	distance score within 0.4 (0.4d)
7	distance score within 0.6 (0.6d)
8	distance score within 0.8 (0.8d)
9	distance score within 1.0 (1.0d)

complete docking process involves obtaining the protein and ligand structures and preparing them for docking. This preparation includes cleaning unnecessary data from protein stucture and ensuring the correct format. The protein and ligand PDB structures are then converted into a PDBQT format. Next, the binding pocket search space is explored and defined. Upon completion of the docking simulation, the results can be visualized.

The filter offers two options for protein input. Users can provide a protein structure ID from the RCSB Protein Data Bank[81] or users can upload their own protein structure file directly. It is important to note that if the protein does not have a ligand already occupying the binding pocket, the user needs to define the potential binding site by providing its coordinates in XYZ format. The docking simulation runs on AutoDock Vina docking engine. We utilize Vinardo[82] scoring function, known for its accuracy based on experimental results. The exhaustiveness parameter was set to its default value of 8. RDKit and Openbabel were utilized to develop the docking filter, while py3Dmol[83] was employed for visualization purposes.

CASE STUDY: IDENTIFYING THREE TYPES OF HIV INHIBITORS

The human immunodeficiency virus (HIV) infection remains a significant global public health concern, despite the development of life-saving combination antiretroviral treatment (cART).[84] According to reports from the World Health Organization (WHO), as of the end of 2022, an estimated 39.0 million individuals were living with HIV, with a significant majority—25.6 million—residing in the WHO African Region. In 2022, 630,000 individuals lost their lives due to HIV-related causes, while 1.3 million people acquired HIV during that year.[85] This virus exhibits the capability to infect various immune system cells, including CD4+ T cells, dendritic cells, and macrophages. Nonetheless, its primary predilection is toward CD4+ T cells, wherein it causes infection and subsequent cell death. Consequently, this process leads to a depletion of the CD4+ T cell population, resulting in severe immunodeficiency.[86] This, in turn, incapacitates the immune system, rendering the patient vulnerable to opportunistic infections. The structure of the HIV is depicted in Figure S1.

Our case study serves a dual purpose: validating the SmartCADD platform and identifying potential HIV inhibitors for further drug development. The study uses SmartCADD, with a deep learning model, trained on experimental HIV screening data from the MoleculeNet database,[31] and other filters, to screen the 800 million compounds from the ZINC database for potential HIV inhibitors.

Data Set, Preprocessing, Training and Prediction. In this case study, we utilize the HIV data set as described by Wu et al.[31] within the MoleculeNet library, which is derived from the AIDS Antiviral Screen Data set released by the National Cancer Institute (NCI).[87] This data set consists of 43,850 molecules, each annotated with its respective experimental EC_{50} and IC_{50} values. These molecules are classified into three categories based on these values: Confirmed Active (CA), Confirmed Moderately Active (CM), and Confirmed Inactive (CI).

Notably, the MoleculeNet version of the HIV data set introduces modifications to the original NCI data set, primarily by aggregating both CA and CM molecules under the active category, while labeling CI molecules as inactive. The MoleculeNet HIV data set is imbalance, with 1443 active compounds contrasted against 39,684 inactive ones. Random undersampling[88] method was employed to balance the data set with 1443 compounds in each class. Undersampling process was repeated three times to account for the inherent randomness in selecting data points for removal. It is important to note that data cleaning procedures can vary depending on the specific data set. Therefore, we recommend that users provide a cleaned data set for SmartCADD's deep learning filter.

The HIV data set was used to train multiple GNNs with different architectures, with the best being wrapped in the deep learning filter for screening HIV-active compounds. The deep learning filter was then applied to predict activity against the ZINC database, which encompasses nearly 800 million compounds. The 800 million compounds were prioritized based on their predicted activity probabilities, facilitating a more focused analysis.

Selection of Three HIV Target Proteins: NNRTIs, NRTIs and PIs. Several proteins exist within the HIV virus as drug targets, including reverse transcriptase (RT), protease, integrase, envelope proteins, and entry coreceptors.[89] HIV drugs target specific HIV proteins, such as non-nucleoside reverse transcriptase inhibitors (NNRTIs), nucleoside reverse transcriptase inhibitors (NRTIs), protease inhibitors (PIs), integrase inhibitors (IIs) and more. The NIC HIV data set contains a collection of inhibitors, potentially including several different inhibitor types. To identify the inhibitor types present in this data set, we employed our trained model to analyze a set of FDA-approved HIV drugs. The identified NNRTIs, NRTIs, and PIs from the model were utilized as three distinct use cases to validate the SmartCADD platform, as elaborated in detail in the results section.

Case Study Design. We conducted a screening of the ZINC database for potential leads using a custom SmartCADD pipeline. The pipeline for this case study begins with the deep learning filter wrapping a trained GNN, followed by the ADMET filter, 2D pharmacophore filter, 3D pharmacophore and quantum mechanics filter, clustering filter, and finally a Smina docking filter. Additionally, we employed an explainable AI module to analyze the predictive ability of our deep learning filter. Further details on these filters are provided in subsequent sections.

RESULTS AND DISCUSSION

Attentive FP: GNN-based Predictions on ZINC Data Set. The initial filter of the SmartCADD pipeline is a deep learning filter, which wraps a GNN trained to discriminate between active and inactive HIV compounds. For GNN model

https://doi.org/10.1021/acs.jcim.4c00720
J. Chem. Inf. Model. 2024, 64, 6799−6813

selection, we conducted a comprehensive comparison of several state-of-the-art GNN architectures, including Attentive FP,[90] GAT (Graph Attention Networks),[91] GCN (Graph Convolutional Networks),[92] and PAGTN (Position-Aware Graph Neural Networks).[93] The efficacy of these models in distinguishing between HIV active and inactive compounds was quantitatively evaluated using the Receiver Operating Characteristic and Area Under the Curve (ROC-AUC) metric. The evaluation revealed that all tested models achieved ROC-AUC scores ranging from 75 to 85%, signifying their substantial predictive capacities shown in Table 4. Among these, the Attentive FP model distinguished itself by employing an attention mechanism that efficiently captures information from neighboring atoms, outperforming other techniques. Consequently, the Attentive FP model was selected for extracting HIV active compounds from the ZINC database, although any model could easily be used within the SmartCADD pipeline due to its flexible filter design. The Deep Learning-based virtual screening filter was employed to identify the top 10 million potential HIV inhibitors, encompassing a diverse range of classes including NNRTIs, NRTIs, PIs and more, for subsequent pharmacophore analysis.

Table 4. Comparison of Leading-Edge GNN Architectures

model	ROC_AUC
attentive FP	**83.34%**
GCN	81.40%
PAGTN	80.18%
GAT	77.35%

ADMET Analysis. Our ADMET filter system effectively identified and removed 409,997 compounds (4.10%) from the 10 M compounds, ensuring the ZINC data set contains drug-like molecules suitable for further analysis.

HIV Inhibitors from De Novo 2D Pharmacophore Analysis. 2D pharmacophore models operate by analyzing pharmacophore features extracted from target compounds, typically FDA-approved drugs or those in clinical trials. Unlike conventional drug targets, viruses often possess multiple drug targets. Table S2 provides a detailed list of the selected drugs, while Figure 6 offers a visual representation. Our case study aimed to identify three categories of lead compounds: NRTIs, NNRTIs and PIs. Emtricitabine, lamivudine, and zidovudine from NRTIs, nevirapine and rilpivirine from NNRTIs, and atazanavir, darunavir, and fosamprenavir from PIs were selected as target compounds for each respective analysis.

Analysis of NRTIs using the 2D pharmacophore filter yielded 1452 lead compounds. Notably, all these compounds share the basic structure of the target compounds: a six-membered aromatic pyrimidine ring and a five-membered aliphatic furan ring. Similarly, the analysis of NNRTIs identified 2716 lead compounds, all containing three six-membered aromatic rings, one of which is a pyrimidine ring. However, the lead compounds derived from PIs were less effective than those from the other two categories, primarily due to the significant structural differences between the target compounds. Three example compounds from each analysis are depicted in Figure 7.

Explainable AI Analysis. To investigate the functional groups governing compound potency, we strategically selected representative compounds from each use case—NNRTIs, NRTIs, and PIs—as illustrated in Figure 8A. Then we use the

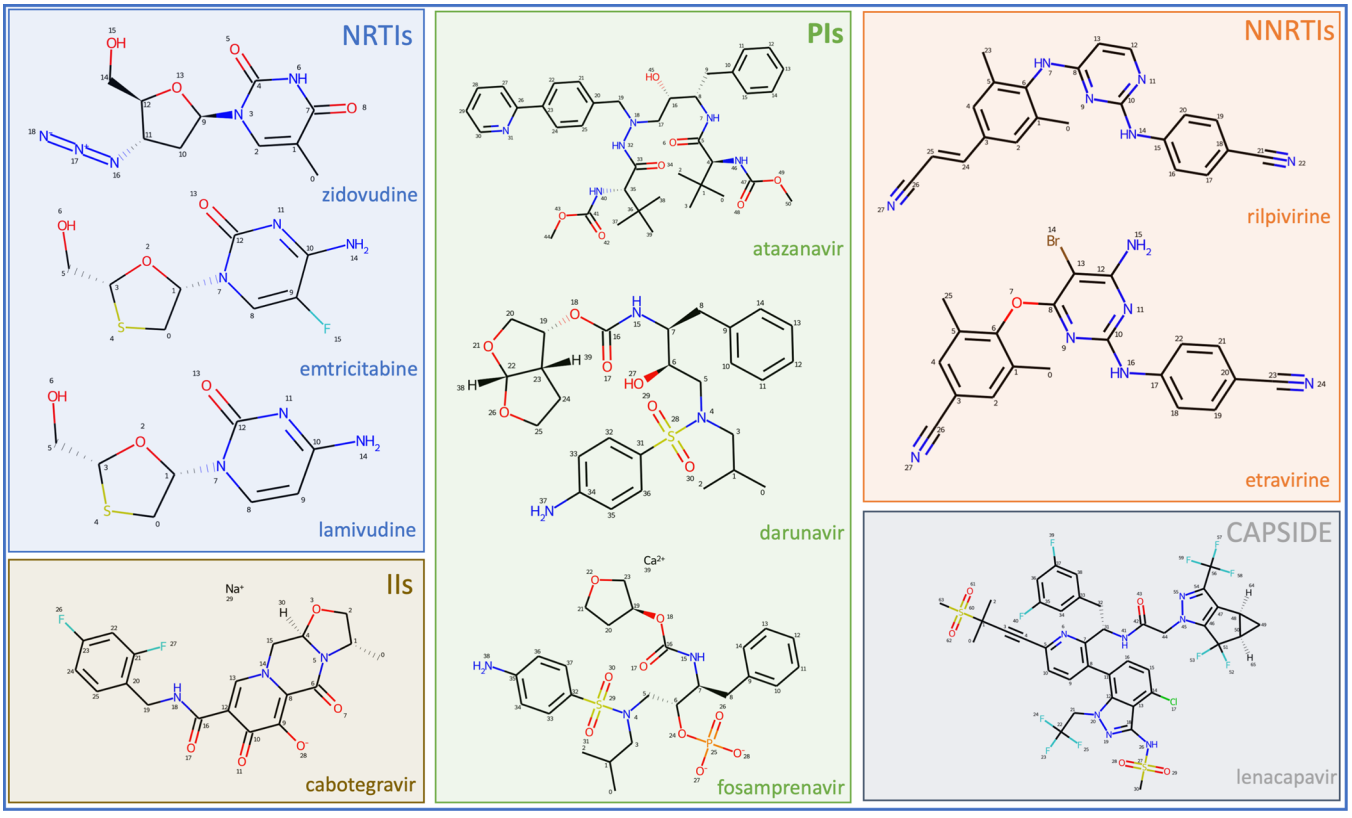


Figure 6. FDA-Approved HIV inhibitors predicted active by GNN.

https://doi.org/10.1021/acs.jcim.4c00720
J. Chem. Inf. Model. 2024, 64, 6799−6813

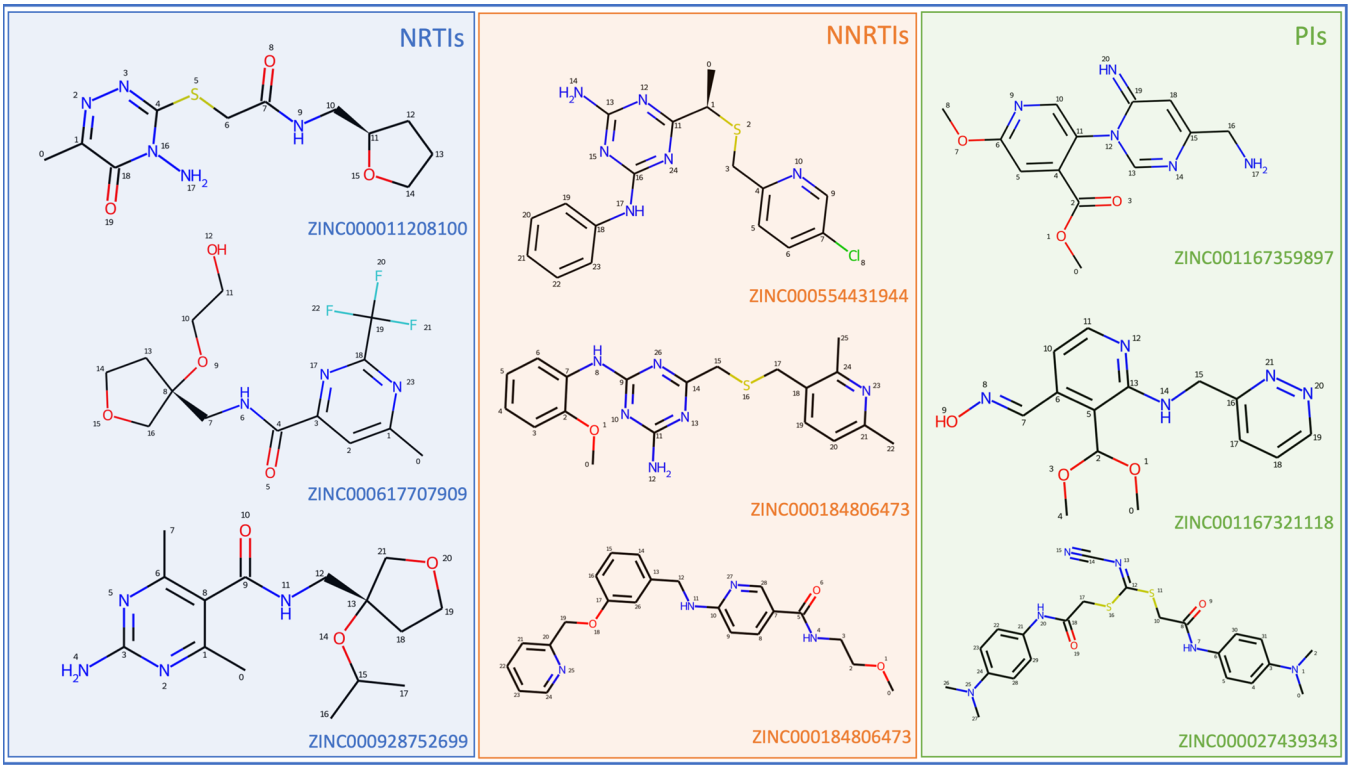


Figure 7. Example compounds obtained by 2D pharmacophore modeling for NRTIs, NNRTIs, and PIs.

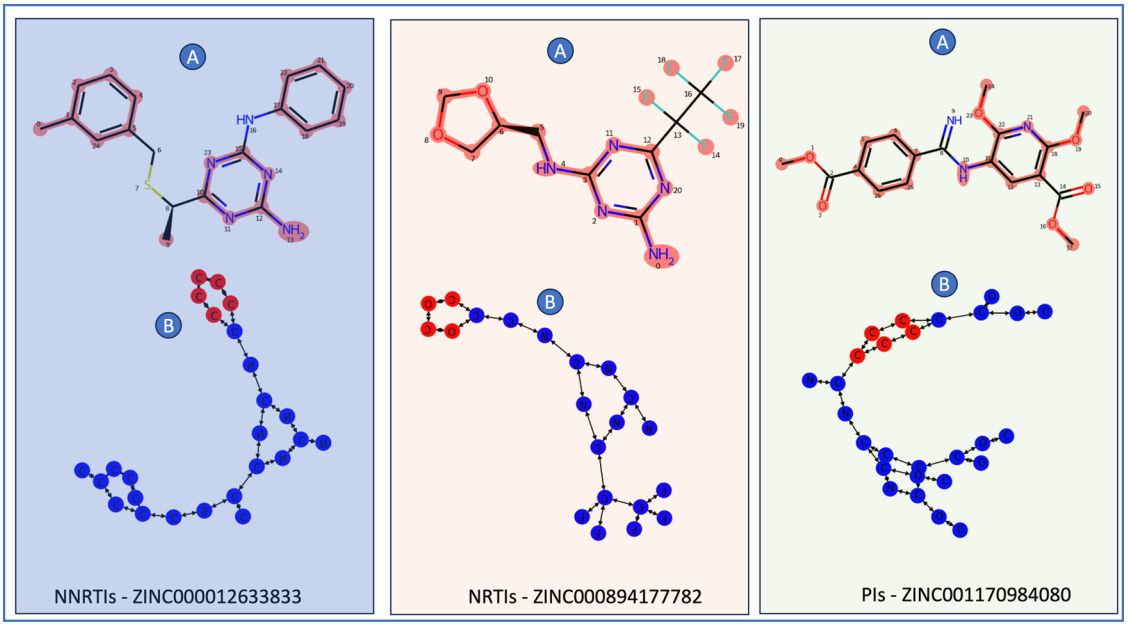


Figure 8. Example XAI analysis from each case study(NNRTIs, NRTIs and PIs). (A) XAI analysis utilizing functional group removal method. (B) SubgraphX method.

ExplainableAI module, wrapping the SubgraphX algorithm,[39] to investigate the functional groups and their importance to the deep learning filter's predictions. The SubgraphX algorithm uses Monte Carlo Tree Search and Shapley Values[94,95] to identify subgraphs within a compound's graph representation that are important to the deep learning models prediction.

A core objective in XAI for drug discovery is to identify functional groups that influence the potency difference between active and inactive compounds. However, current XAI algorithms are under development and often struggle to address more than one specific question at a time.[96−98] In our case, we leveraged XAI to understand the role of aromatic rings[99,100] in compound activity.

To understand the functional groups affecting activity in these molecules, we employed a two-pronged approach. First, a functional group removal method identified potential activity determinants including HBA, HBD, and aromatic rings. Second, SubgraphX algorithm-based XAI analysis provided

https://doi.org/10.1021/acs.jcim.4c00720
J. Chem. Inf. Model. 2024, 64, 6799−6813

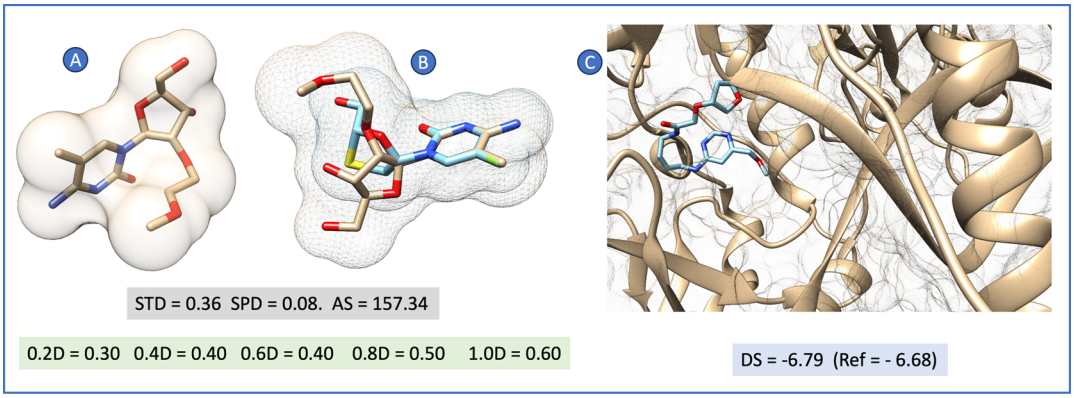


Figure 9. Example 3D analysis results for 26th conformation of ZINC000002583385 taken by NRTIs category. (A) ZINC000002583385 compound, (B) 26th conformation of ZINC000002583385 with emtricitabine, and (C) docked ZINC000002583385 structure with reverse transcriptase protein (PDB ID: 6WPJ).[101] The reference value mentions the docking score of emtricitabine.

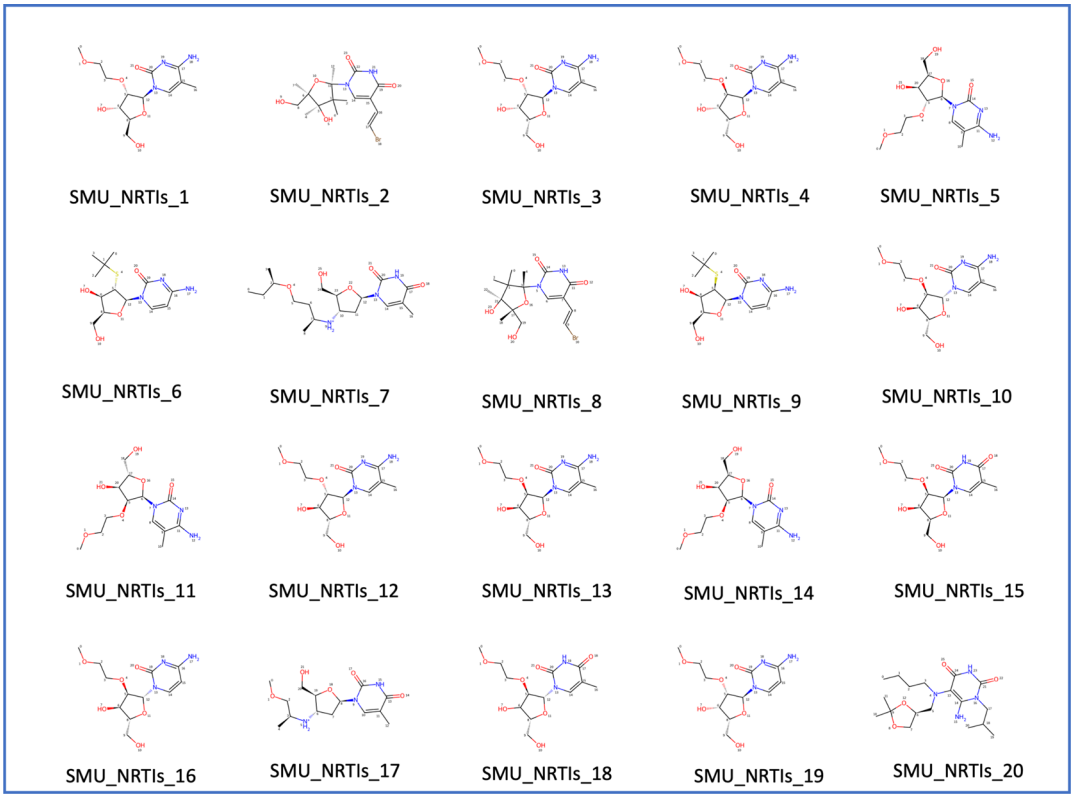


Figure 10. Molecular representations of the top 20 NRTIs from SmartCADD described in Table 5.

deeper insights for specific drug classes. This analysis revealed aromatic rings as the key functional group for predicting activity in NNRTIs and PIs, while for NRTIs, heterocyclic rings[20] emerged as the critical factor.

It is crucial to acknowledge that the chosen method, while insightful, possesses inherent limitations. These limitations include its applicability only to active compounds and its inability to modify ring structures in Figure 8A. SubgraphX effectively identified ring structures critical for predicting compound activity, as depicted by the red atoms in Figure 8B. However, it does not capture the full context by overlooking other functional groups.

3D Pharmacophore and Docking Analysis. The most promising lead compounds exhibit higher AS and DS values, coupled with lower STD and SPD values. This allows users to selectively choose the best compounds based on their preferences. Scores between $0.2d$ and $-1.0d$ indicate the likelihood of finding the maximum number of pharmacophores within a specific distance. Higher scores at any distance suggest a greater chance of compound similarity to the target. An example analysis is showcased in Figure 9. These scores are crucial for the lead optimization process, guiding modifications to enhance the compound's potency and pharmacokinetic features.

Following a 2D study, we extracted 1452 compounds and generated 8 3D parameters for each, considering 100 conformations per compound. Then, we sorted all compounds (1452×100) based on their aligning score (AS). The sorting based on AS yielded a pattern nearly identical to the sorting based on the shape Tanimoto and shape protrude scores. To

https://doi.org/10.1021/acs.jcim.4c00720
J. Chem. Inf. Model. 2024, 64, 6799−6813

Table 5. Top 20 NRTIs from SmartCADD with Their Eight 3D Parameters and Docking Score[a]

lead compound	conf	STD	SPD	AS	DS	0.2*d*	0.4*d*	0.6*d*	0.8*d*	1.0*d*
SMU_NRTIs_1	29	0.36	0.08	157.34	−6.79	0.30	0.40	0.40	0.50	0.60
SMU_NRTIs_2	51	0.42	0.07	140.79	−5.62	0.50	0.50	0.63	0.63	0.88
SMU_NRTIs_3	74	0.41	0.12	140.22	−6.92	0.40	0.40	0.60	0.60	0.70
SMU_NRTIs_4	83	0.42	0.14	139.53	−6.37	0.40	0.40	0.50	0.50	0.60
SMU_NRTIs_5	47	0.41	0.13	137.86	−7.30	0.40	0.40	0.50	0,60	0.70
SMU_NRTIs_6	6	0.45	0.15	137.83	−5.95	0.50	0.50	0.63	0.63	0.75
SMU_NRTIs_7	39	0.52	0.14	137.76	−7.63	0.32	0.44	0.66	0.66	0.77
SMU_NRTIs_8	18	0.43	0.08	137.74	−5.65	0.50	0.50	0.63	0.75	0.88
SMU_NRTIs_9	19	0.46	0.17	137.40	−5.35	0.50	0.50	0.63	0.63	0.88
SMU_NRTIs_10	39	0.44	0.16	136.80	−6.61	0.00	0.30	0.40	0.70	0.80
SMU_NRTIs_11	52	0.40	0.11	136.63	−6.06	0.30	0.40	0.40	0.60	0.60
SMU_NRTIs_12	92	0.43	0.15	136.41	−7.41	0.40	0.40	0.50	0.50	0.60
SMU_NRTIs_13	22	0.45	0.17	136.17	−7.30	0.40	0.40	0.50	0.50	0.70
SMU_NRTIs_14	2	0.40	0.12	135.90	−6.08	0.30	0.40	0.50	0.60	0.60
SMU_NRTIs_15	22	0.41	0.13	135.89	−6.32	0.40	0.40	0.50	0.60	0.70
SMU_NRTIs_16	48	0.44	0.18	134.93	−6.95	0.30	0.40	0.60	0.60	0.70
SMU_NRTIs_17	8	0.41	0.12	134.80	−6.61	0.33	0.44	0.55	0.55	0.66
SMU_NRTIs_18	66	0.41	0.13	134.40	−6.78	0.40	0.50	0.60	0.60	0.60
SMU_NRTIs_19	83	0.44	0.18	134.32	−7.35	0.30	0.40	0.40	0.50	0.60
SMU_NRTIs_20	77	0.50	0.10	133.65	−5.59	0.00	0.11	0.22	0.55	0.55

[a]Compounds were sorted according to the AS value, and the pattern was validated using docking simulation. The docking simulation was carried out for the HIV NRT protein (PDB ID: 6WPJ). Emtricitabine was taken as the reference docking score with a docking score of −6.68 kJ/mol. (Conf, molecular conformation; STD, shape tanimoto distances; SPD, shape protrude distance; AS, align score; DS, docking score; *d*, distance score).

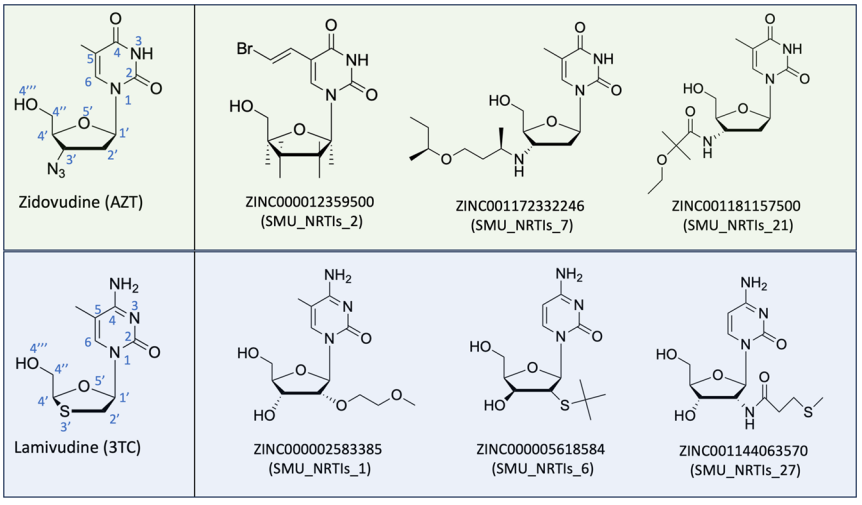


Figure 11. Structural alignment of NRTI derivatives from SmartCADD with zidovudine (AZT) and lamivudine (3TC).

further validate the observed pattern, we performed docking simulations of the compounds with their corresponding specific receptors (PDB ID: 6WPJ). Interestingly, docking scores for top-selected compounds consistent with the identified pattern were either higher than or close to the reference value, suggesting strong agreement between the pattern and docking results. It is important to note that while docking simulation is a valuable tool have limitations in precisely capturing protein−ligand interactions due to potential variations in docking scores of up to ±1 kJ/mol. The top 20 lead compounds for NRTIs are depicted in Figure 10, accompanied by their respective nine parameters listed in Table 5. Similarly, the lead compounds for NNRTIs and PIs are illustrated in Figures S2 and S3, respectively, along with their corresponding data provided in Tables S1 and S2.

Literature consistently highlights the use of FDA-approved drug derivatives as foundational structures for drug discovery campaigns, aiming to enhance properties like ADMET,[102] affinity,[103] and selectivity.[104] Similarly, derivatives of HIV NRTI drugs, such as Zidovudine (AZT) and Lamivudine (3TC), have demonstrated potential for drug development.[105−108] We selected the top 30 NRTI compounds from the SmartCADD and divided them into two clusters based on structural similarity to the FDA-approved drugs, AZT and 3TC, as detailed in Figure 11. SMU_NRTIs_2, 7, and 21 exhibit high structural similarity to AZT. Compounds 2 and 7 are particularly similar, differing only in the substituent at the 3′ position of the five-membered ring. On the other hand, SMU_NRTIs_1, 6, and 27 displayed structural similarities to Lamivudine, with primary differences observed in the substituents at the 3′ and 2′ positions of the five-membered

https://doi.org/10.1021/acs.jcim.4c00720
J. Chem. Inf. Model. 2024, 64, 6799−6813

Table 6. SmartCADD Feature Comparison with the AIDDISON Package

	AIDDISON	SmartCADD
DL-based initial Screening	NA	deep learning filter - Users have the flexibility to bring their pretrained models or choose from a variety of model options
ADMET calculation	NA	updated Ro5 filters; TPSA, ROT and NAR filters; PAINS filters
2D pharmacophore	**FTrees**[112] topological discriptors capture rings, chains and pharmacophore attributes.	**2D pharmacophore filter** capture both aromatic and aliphatic rings, pharmacophore attributes and specific heterocyclic rings (ex. furan, pyrimidine and etc.)
3D pharmacophore	**Cresset's flare**[113] scores -3D alignment score with the target energy minimization - XTB (GFN2-xTB) ligand flexibility - not mentioned extra - asymmetric Tversky index subfield or superfield search to filter out molecules	**3D pharmacophore filter** scores -3D alignment score (AS) with the target shape Tanimoto distances (STD) shape protrude distance (SPD) energy minimization - XTB (GFN2-xTB) ligand flexibility - use 100 conformations extra - distance score (0.2−1.0) determine the number of pharmacophores can be found in a specific distance.
molecular docking	**flare docking from Cresset**[114] PDB ID or user can directly upload a protein structure.	**Smina Docking** PDB ID or user can directly upload a protein structure.
generative techniques	REINVENT 2.0[115] and synthetic accessibility via SYNTHIA[116]	NA
explainable AI	NA	SubgraphX module
type of application	web-based tool	Python package
type of workflow	isolated calculations	modular pipeline flow
availability	commercially available	open-source

ring. Furthermore, the ZINC database offers multiple conformations for compounds, each assigned a unique ZINC ID. SmartCADD can also identify top lead conformations. For example, among the top 30 NRTI compounds, SMU_NRTIs_1, 3, 4, 5, 10, 11, 12, 13, 14, 15, 16, and 18 were identified as different conformations of the same molecule. However, SMU_NRTIs_20 and 26 exhibited significant structural differences compared to the FDA-approved compounds. We have included detailed information regarding the SmartCADD validation protocol in the Supporting Information.

Summary: Case Study. We demonstrated the SmartCADD platform through three case studies targeting HIV NRTIs, NNRTIs, and PIs. The GNN model was trained on a data set of approximately 1500 active and 1500 inactive compounds. The model achieved an accuracy of approximately 85% and was further validated with FDA-approved compounds. Subsequently, the trained GNN model was wrapped in the deep learning filter to identify the top 10 million potentially active HIV compounds from the ZINC data set. Next, the ADMET filter ensured the retention of only drug-like candidates. Following this initial filtering step, the 2D pharmacophore filter further refined the pool, resulting in 1452 NRTIs, 2716 NNRTIs, and 871 PIs as promising lead candidates. Next, the 3D pharmacophore filter was applied to these QM-optimized compounds identified from the 2D analysis. This 3D model generated nine key 3D parameters, with 100 conformations analyzed for each molecule to account for structural flexibility. The top 20 compounds were selected based on a careful evaluation of the nine parameters. In addition, the explainable AI model pinpointed functional groups crucial for potency determination, providing valuable insights into GNN decision-making. Computational time and power for the case study are described in Supporting Information (Table S1).

While this specific example showcases SmartCADD's capabilities, it is important to note that the software's full potential extends beyond this scenario. One key aspect of SmartCADD is its deep learning filter. This filter requires training data with both active and inactive compounds, these types of Boiassay data is available on the PubChem BioAssay database[109] which includes over one million records. SmartCADD is versatile. It can function as a complete screening pipeline, but its individual filters can also be used independently. Furthermore, 2D and 3D pharmacophore analysis require reference compounds for a specific target. While FDA-approved drugs are ideal, Users can use reference compounds with experimental results, such as those obtained from in vitro or clinical trials. Also, if the reference compounds are structurally different, consider using reference compounds with higher alignment scores for better accuracy for 2D pharmacophore analysis. SmartCADD's complete filter process can be applied to uncover potential lead compounds for a variety of targets, including the Formylpeptide Receptor, Rho kinase 2, and the sphingolipid G-protein-coupled receptor and more.[110]

SmartCADD Feature Comparison with AIDDISON. AI and CADD-empowered drug discovery platforms are not extensively documented in the literature. A recent publication from Merck Healthcare detailed the AIDDISON platform,[111] highlighting its utilization of AI and CADD methodologies. AIDDISON uses advanced 2D/3D pharmacophore models and docking analysis to identify promising drug candidates with generative techniques. By comparatively analyzing the functionalities of SmartCADD with those of AIDDISON (as presented in Table 6), we aim to illuminate the specific advantages offered by our SmartCADD platform.

CONCLUSIONS

We introduce SmartCADD, a user-friendly virtual screening platform providing researchers with a highly integrated and flexible framework for building drug discovery pipelines. Built as a Python package, it seamlessly integrates a wide range of functionalities, including AI/ML algorithms, explainable AI descriptors, ADMET property calculations, de novo 2D/3D pharmacophore analysis, molecular docking, and even QM calculations. When applying SmartCADD to screen the ZINC database for HIV-inhibiting compounds, we successfully identified promising drug candidates, including 1452 NRTIs, 2716 NNRTIs, and 871 PIs. These candidates were further refined using QM-optimized 3D parameters generated for all compounds by SmartCADD. Notably, these nine 3D parameters hold significant value for following lead optimization and development processes. SmartCADD's ability to efficiently screen billions of compounds daily significantly reduces discovery timelines and expedites the identification of promising leads. This unique combination of flexibility,

https://doi.org/10.1021/acs.jcim.4c00720
J. Chem. Inf. Model. 2024, 64, 6799−6813

cutting-edge technology, and efficiency positions SmartCADD at the forefront of drug discovery, empowering researchers to make groundbreaking advancements in the field.

■ ASSOCIATED CONTENT

Data Availability Statement

The SmartCADD platform is freely available on GitHub at https://github.com/SMU-CATCO/SmartCADD.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c00720.

> Review of explainable AI: tools, methods, applications, and limitations; description of SmartCADD's tautomers and protomers models; evaluating computational cost of SmartCADD modules; graphical representation of HIV virus structure; validation of the GNN filter with FDA-approved HIV drugs; case study 2 results: HIV NNRTIs from SmartCAD; case study 3 results: HIV PIs from SmartCADD; SmartCADD validation; references (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Elfi Kraka − *Department of Chemistry, Southern Methodist University, Dallas, Texas 75205, United States;* orcid.org/0000-0002-9658-5626; Email: amahamadakalapuwage@smu.edu, ejlaird@smu.edu

Authors

Ayesh Madushanka − *Department of Chemistry, Southern Methodist University, Dallas, Texas 75205, United States*

Eli Laird − *Department of Computer Science, Southern Methodist University, Dallas, Texas 75205, United States;* orcid.org/0000-0002-0668-8745

Corey Clark − *Department of Computer Science, Southern Methodist University, Dallas, Texas 75205, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.4c00720

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was financially supported by the National Science Foundation (grant number CHE 2102461) and the DSF Charitable Foundation. We thank SMU's O'Donnell Data Science and Research Computing Institute for a generous allotment of computer time.

■ REFERENCES

(1) Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackermann, Z.; et al. A deep learning approach to antibiotic discovery. *Cell* **2020**, *180*, 688−702.

(2) Cortes, J.; Perez-García, J. M.; Llombart-Cussac, A.; Curigliano, G.; El Saghir, N. S.; Cardoso, F.; Barrios, C. H.; Wagle, S.; Roman, J.; Harbeck, N.; et al. Enhancing global access to cancer medicines. *CA Cancer J. Clin.* **2020**, *70*, 105−124.

(3) Bianculli, R. H.; Mase, J. D.; Schulz, M. D. Antiviral polymers: past approaches and future possibilities. *Macromolecules* **2020**, *53*, 9158−9186.

(4) Walters, W. P.; Barzilay, R. Applications of deep learning in molecule generation and molecular property prediction. *Acc. Chem. Res.* **2021**, *54*, 263−270.

(5) Sousa, T.; Correia, J.; Pereira, V.; Rocha, M. Generative deep learning for targeted compound design. *J. Chem. Inf. Model.* **2021**, *61*, 5343−5361.

(6) Makoś, M. Z.; Verma, N.; Larsson, E. C.; Freindorf, M.; Kraka, E. Generative adversarial networks for transition state geometry prediction. *J. Chem. Phys.* **2021**, *155*, No. 024116.

(7) Wei, Y.; Li, S.; Li, Z.; Wan, Z.; Lin, J. Interpretable-ADMET: a web service for ADMET prediction and optimization based on deep neural representation. *Bioinformatics* **2022**, *38*, 2863−2871.

(8) Tian, H.; Ketkar, R.; Tao, P. ADMETboost: a web server for accurate ADMET prediction. *J. Mol. Model.* **2022**, *28*, 408.

(9) Maia, E. H. B.; Assis, L. C.; De Oliveira, T. A.; Da Silva, A. M.; Taranto, A. G. Structure-based virtual screening: from classical to artificial intelligence. *Front. Chem.* **2020**, *8*, 343.

(10) Yang, Y.; Zhu, Z.; Wang, X.; Zhang, X.; Mu, K.; Shi, Y.; Peng, C.; Xu, Z.; Zhu, W. Ligand-based approach for predicting drug targets and for virtual screening against COVID-19. *Briefings in Bioinformatics* **2021**, *22*, 1053−1064.

(11) Verma, N.; Qu, X.; Trozzi, F.; Elsaied, M.; Karki, N.; Tao, Y.; Zoltowski, B.; Larson, E.; Kraka, E. SSnet: A Deep Learning Approach for Protein−Ligand Interaction Prediction. *Int. J. Mol. Sci.* **2021**, *22*, 1392.

(12) Coderc, G.; de Lacam, E.; Roux, B.; Chipot, C. Classifying Protein−Protein Binding Affinity with Free-Energy Calculations and Machine Learning Approaches. *J. Chem. Inf. Model.* **2024**, *64*, 1081−1091.

(13) Jiang, Y.; Yu, Y.; Kong, M.; Mei, Y.; Yuan, L.; Huang, Z.; Kuang, K.; Wang, Z.; Yao, H.; Zou, J.; et al. Artificial intelligence for retrosynthesis prediction. *Engineering* **2023**, *25*, 32−50.

(14) Ishida, S.; Terayama, K.; Kojima, R.; Takasu, K.; Okuno, Y. Ai-driven synthetic route design incorporated with retrosynthesis knowledge. *J. Chem. Inf. Model.* **2022**, *62*, 1357−1367.

(15) Schneider, P.; Walters, W. P.; Plowright, A. T.; Sieroka, N.; Listgarten, J.; Goodnow, R. A., Jr; Fisher, J.; Jansen, J. M.; Duca, J. S.; Rush, T. S.; et al. Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discovery* **2020**, *19*, 353−364.

(16) Bender, A.; Cortés-Ciriano, I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. *Drug Discovery Today* **2021**, *26*, 511−524.

(17) Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2021**, *23*, 18.

(18) Laird, E.; Madushanka, A.; Kraka, E.; Clark, C. XInsight: Revealing Model Insights for GNNs with Flow-Based Explanations. *Explainable Artificial Intelligence*; Springer: Cham, 2023; 303−320.

(19) Saeed, W.; Omlin, C. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems* **2023**, *263*, No. 110273.

(20) Harren, T.; Matter, H.; Hessler, G.; Rarey, M.; Grebner, C. Interpretation of structure−activity relationships in real-world drug design data sets using explainable artificial intelligence. *J. Chem. Inf. Model.* **2022**, *62*, 447−462.

(21) Xu, F.; Uszkoreit, H.; Du, Y.; Fan, W.; Zhao, D.; Zhu, J. Explainable AI: A brief survey on history, research areas, approaches and challenges. *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9−14, 2019, Proceedings, Part II 8*. 2019; 563−574.

(22) Ponzoni, I.; Páez Prosper, J. A.; Campillo, N. E. Explainable artificial intelligence: A taxonomy and guidelines for its application to drug discovery. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2023**, *13*, No. e1681.

(23) Laird, E.; Madushanka, A.; Kraka, E.; Clark, C. XInsight: Revealing Model Insights for GNNs with Flow-Based Explanations. *World Conference on Explainable Artificial Intelligence*; Springer: Cham, 2023; 303−320.

(24) Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* **2020**, *2*, 573−584.
(25) Gamma, E.; Helm, R.; Johnson, R.; Vlissides, J. M. *Design Patterns: Elements of Reusable Object-Oriented Software*, 1st ed.; Addison-Wesley Professional, 1994; 1−366.
(26) Koes, D. R.; Baumgartner, M. P.; Camacho, C. J. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J. Chem. Inf. Model.* **2013**, *53*, 1893−1904.
(27) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **2009**, *30*, 2785−2791.
(28) Laird, E.; Madushanka, A. *SmartCADD: An AI-Integrated Drug Designing Platform*, 2024 https://github.com/SMU-CATCO/SmartCADD; urldate: (June 08 2024).
(29) RDKit, *RDKit: Open-source cheminformatics*. 2023; http://www.rdkit.org, urldate: (Aug 18 2023).
(30) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, *3*, 33.
(31) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513−530.
(32) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757−1768.
(33) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids res.* **2012**, *40*, D1100−D1107.
(34) Li, Q.; Cheng, T.; Wang, Y.; Bryant, S. H. PubChem as a public resource for drug discovery. *Drug Discovery Today* **2010**, *15*, 1052−1057.
(35) Fey, M.; Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. *ICLR Workshop on Representation Learning on Graphs and Manifolds*. 2019; 1−9.
(36) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V.; Leswing, K.; Wu, Z. *Deep Learning for the Life Sciences*; O'Reilly Media, 2019; https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837.
(37) Ying, Z.; Bourgeois, D.; You, J.; Zitnik, M.; Leskovec, J. Gnnexplainer: Generating explanations for graph neural networks. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 9240−9251.
(38) Vu, M.; Thai, M. T. Pgm-explainer: Probabilistic graphical model explanations for graph neural networks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12225−12235.
(39) Yuan, H.; Yu, H.; Wang, J.; Li, K.; Ji, S.On explainability of graph neural networks via subgraph explorations. *International conference on machine learning*, 2021, 1224112252 .
(40) Yuan, H.; Tang, J.; Hu, X.; Ji, S. XGNN: Towards Model-Level Explanations of Graph NeuralNetworks. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; Virtual Event: CA USA, 2020; 430−438.
(41) Xiong, G.; Wu, Z.; Yi, J.; Fu, L.; Yang, Z.; Hsieh, C.; Yin, M.; Zeng, X.; Wu, C.; Lu, A.; et al. ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. *Nucleic Acids Res.* **2021**, *49*, W5−W14.
(42) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(43) Zhang, M.-Q.; Wilkinson, B. Drug discovery beyond the 'rule-of-five'. *Curr. Opin. Biotechnol.* **2007**, *18*, 478−488.
(44) Lipinski, C. A. Rule of five in 2015 and beyond: Target and ligand structural limitations, ligand chemistry structure and drug discovery project decisions. *Adv. Drug Delivery Rev.* **2016**, *101*, 34−41.
(45) Hartung, I. V.; Huck, B. R.; Crespo, A. Rules were made to be broken. *Nat. Rev. Chem.* **2023**, *7*, 3−4.
(46) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **2002**, *45*, 2615−2623.
(47) Protti, Í. F.; Rodrigues, D. R.; Fonseca, S. K.; Alves, R. J.; de Oliveira, R. B.; Maltarollo, V. G. Do Drug-likeness Rules Apply to Oral Prodrugs? *ChemMedChem.* **2021**, *16*, 1446−1456.
(48) Ritchie, T. J.; Macdonald, S. J. The impact of aromatic ring count on compound developability−are too many aromatic rings a liability in drug design? *Drug Discovery Today* **2009**, *14*, 1011−1020.
(49) Baell, J. B.; Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **2010**, *53*, 2719−2740.
(50) Walters, P. *rd filters*, 2017 https://github.com/PatWalters/rd_filters; urldate: (May 12 2024).
(51) Guuner, O. F.; Bowen, J. P. Setting the record straight: The origin of the pharmacophore concept. *J. Chem. Inf. Model.* **2014**, *54*, 1269−1283.
(52) Horvath, D.; Mao, B.; Gozalbes, R.; Barbosa, F.; Rogalski, S. L. Strengths and Limitations of Pharmacophore-Based Virtual Screening. *Chemoinformatics in Drug DiscoVery* **2005**, 117−137.
(53) Stromgaard, K.; Krogsgaard-Larsen, P.; Madsen, U. *Textbook of drug design and discovery*; CRC press, 2009; 1−222.
(54) Muhammed, M. T.; Akı-yalcın, E. Pharmacophore modeling in drug discovery: methodology and current status. *Journal of the Turkish Chemical Society Section A: Chemistry* **2021**, *8*, 749−762.
(55) Fernández-de Gortari, E.; García-Jacas, C. R.; Martinez-Mayorga, K.; Medina-Franco, J. L. Database fingerprint (DFP): an approach to represent molecular databases. *J. Cheminform.* **2017**, *9*, 9.
(56) Duan, J.; Dixon, S. L.; Lowrie, J. F.; Sherman, W. Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. *J. Mol. Graph. Model* **2010**, *29*, 157−170.
(57) Pattanaik, L.; Coley, C. W. Molecular representation: going long on fingerprints. *Chem.* **2020**, *6*, 1204−1207.
(58) Zhong, S.; Guan, X. Count-based morgan fingerprint: A more efficient and interpretable molecular representation in developing machine learning-based predictive regression models for water contaminants' activities and properties. *Environ. Sci. Technol.* **2023**, *57*, 18193−18202.
(59) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58−63.
(60) Stiefl, N.; Watson, I. A.; Baumann, K.; Zaliani, A. ErG: 2D pharmacophore descriptions for scaffold hopping. *J. Chem. Inf. Model.* **2006**, *46*, 208−220.
(61) Abu Hammad, A. M.; Taha, M. O. Pharmacophore modeling, quantitative structure- activity relationship analysis, and shape-complemented in silico screening allow access to novel influenza neuraminidase inhibitors. *J. Chem. Inf. Model.* **2009**, *49*, 978−996.
(62) Ward, S. E.; Beswick, P. What does the aromatic ring number mean for drug design? *Expert Opin. Drug Discovery* **2014**, *9*, 995−1003.
(63) Ritchie, T. J.; Macdonald, S. J. Physicochemical descriptors of aromatic character and their use in drug discovery: miniperspective. *J. Med. Chem.* **2014**, *57*, 7206−7215.
(64) Yan, M.; Xu, L.; Wang, Y.; Wan, J.; Liu, T.; Liu, W.; Wan, Y.; Zhang, B.; Wang, R.; Li, Q. Opportunities and challenges of using five-membered ring compounds as promising antitubercular agents. *Drug Dev. Res.* **2020**, *81*, 402−418.
(65) Taylor, A. P.; Robinson, R. P.; Fobian, Y. M.; Blakemore, D. C.; Jones, L. H.; Fadeyi, O. Modern advances in heterocyclic chemistry in drug discovery. *Org. Biomol. Chem.* **2016**, *14*, 6611−6637.
(66) Li, J. J. *Heterocyclic chemistry in drug discovery*; John Wiley & Sons, 2013; 1−116.
(67) Kontoyianni, M. Docking and virtual screening in drug discovery. *Proteomics for drug discovery: Methods and protocols* **2017**, *1647*, 255−266.

https://doi.org/10.1021/acs.jcim.4c00720
J. Chem. Inf. Model. 2024, 64, 6799−6813

(68) Forli, S.; Huey, R.; Pique, M. E.; Sanner, M. F.; Goodsell, D. S.; Olson, A. J. Computational protein−ligand docking and virtual drug screening with the AutoDock suite. *Nat. Protoc.* **2016**, *11*, 905−919.
(69) Gupta, M.; Sharma, R.; Kumar, A. Docking techniques in pharmacology: How much promising? *Comput. Biol. Chem.* **2018**, *76*, 210−217.
(70) Chen, Y.-C. Beware of docking! *Trends Pharmacol. Sci.* **2015**, *36*, 78−95.
(71) Bolcato, G.; Heid, E.; Bostroom, J. On the value of using 3D shape and electrostatic similarities in deep generative methods. *J. Chem. Inf. Model.* **2022**, *62*, 1388−1398.
(72) Hua, Y.; Huang, D.; Liang, L.; Qian, X.; Dai, X.; Xu, Y.; Qiu, H.; Lu, T.; Liu, H.; Chen, Y.; Zhang, Y.; et al. FSDscore: An Effective Target-focused Scoring Criterion for Virtual Screening. *Mol. Inform.* **2023**, *42*, No. 2200039.
(73) Tosco, P.; Balle, T.; Shiri, F. Open3DALIGN: an open-source software aimed at unsupervised ligand alignment. *J. Comput. Aided Mol. Des.* **2011**, *25*, 777−783.
(74) Baraque de Freitas Rodrigues, S.; Santos Aquino de Araújo, R.; Dantas de Mendonça, T. R.; Bezerra Mendonça-Júnior, F. J.; Zhan, P.; Ferreira da Silva-Júnior, E.; Santos Nascimento, I. J. d.; de Freitas Rodrigues, S. B.; de Araújo, R. S. A.; de Mendonça, T. R. D.; Mendonça-Júnior, F. J. B.; Zhan, P.; da Silva-Júnior, E. F.Quantum Chemistry in Drug Design: Density Function Theory (DFT) and Other Quantum Mechanics (QM)-related Approaches. *Applied Computer-Aided Drug Design: Models and Methods*, 2023, 258309 .
(75) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. Extended tight-binding quantum chemistry methods. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2021**, *11*, No. e1493.
(76) Rasmussen, M. H.; Jensen, J. H. Fast and automatic estimation of transition state structures using tight binding quantum chemical calculations. *PeerJ. Physical Chemistry* **2020**, *2*, No. e15.
(77) Grimme, S.; Bohle, F.; Hansen, A.; Pracht, P.; Spicher, S.; Stahn, M. Efficient quantum chemical calculation of structure ensembles and free energies for nonrigid molecules. *J. Phys. Chem. A* **2021**, *125*, 4039−4054.
(78) Schmitz, S.; Seibert, J.; Ostermeir, K.; Hansen, A.; Goller, A. H.; Grimme, S. Quantum chemical calculation of molecular and periodic peptide and protein structures. *J. Phys. Chem. B* **2020**, *124*, 3636−3646.
(79) Chen, Y.-Q.; Sheng, Y.-J.; Ma, Y.-Q.; Ding, H.-M. Efficient calculation of protein−ligand binding free energy using GFN methods: The power of the cluster model. *Phys. Chem. Chem. Phys.* **2022**, *24*, 14339−14347.
(80) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652−1671.
(81) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(82) Quiroga, R.; Villarreal, M. A. Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening. *PloS one* **2016**, *11*, No. e0155183.
(83) Rego, N.; Koes, D. 3Dmol. js: molecular visualization with WebGL. *Bioinformatics* **2015**, *31*, 1322−1324.
(84) Maenza, J.; Flexner, C. Combination antiretroviral therapy for HIV infection. *Am. Fam. Physician* **1998**, *57*, 2789−2798.
(85) WHO *HIV and AIDS*. https://www.who.int/news-room/fact-sheets/detail/hiv-aids?%gclid=CjwKCAjwseSoBhBXEiwA9iZtxvOx6OMF3C61jJMcs4ckrV8tg-%jmR5WER9TVwjpc67X9LaypfbJrORoCi_wQAvD_BwE, Last updated: (July 13 2023); urldate: (Oct 01 2023).
(86) Monaco, C. L.; Gootenberg, D. B.; Zhao, G.; Handley, S. A.; Ghebremichael, M. S.; Lim, E. S.; Lankowski, A.; Baldridge, M. T.; Wilen, C. B.; Flagg, M.; et al. Altered virome and bacterial microbiome in human immunodeficiency virus-associated acquired immunodeficiency syndrome. *Cell Host. Microbe.* **2016**, *19*, 311−322.
(87) Institute, N. C. *AIDS Antiviral Screen Data*. https://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Data, Last updated: (May 2004); urldate: (Oct 01 2023).
(88) Paula, B.; Torgo, L.; Ribeiro, R. A survey of predictive modelling under imbalanced distributions. *arXiv preprint arXiv* **2015**, *1505*.
(89) Pan, X.; Baldauf, H.-M.; Keppler, O. T.; Fackler, O. T. Restrictions to HIV-1 replication in resting CD4+ T lymphocytes. *Cell Res.* **2013**, *23*, 876−885.
(90) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J. Med. Chem.* **2020**, *63*, 8749−8760.
(91) Brody, S.; Alon, U.; Yahav, E. How attentive are graph attention networks?. *arXiv preprint arXiv:2105.14491* **2021**.
(92) Chen, M.; Wei, Z.; Huang, Z.; Ding, B.; Li, Y. Simple and deep graph convolutional networks. *International conference on machine learning*. 2020; 1725−1735.
(93) You, J.; Ying, R.; Leskovec, J. Position-aware Graph Neural Networks. *International Conference on Machine Learning*. 2019; 7134−7143.
(94) Świechowski, M.; Godlewski, K.; Sawicki, B.; Mańdziuk, J. Monte Carlo tree search: A review of recent modifications and applications. *Artificial Intelligence Review* **2023**, *56*, 2497−2562.
(95) Shapley, L. S. In *Contributions to the Theory of Games II*; Kuhn, H. W.; Tucker, A. W., Eds.; Princeton University Press: Princeton, 1953; 307−317.
(96) Kanehira, A.; Harada, T. Learning to explain with complemental examples. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019; 8603−8611.
(97) Sheridan, R. P. Interpretation of QSAR models by coloring atoms according to changes in predicted activity: how robust is it? *J. Chem. Inf. Model.* **2019**, *59*, 1324−1337.
(98) Russell, C. Efficient search for diverse coherent explanations. *Proceedings of the conference on fairness, accountability, and transparency*. 2019; 20−28.
(99) Wellawatte, G. P.; Seshadri, A.; White, A. D. Model Agnostic Generation of Counterfactual Explanations for Molecules. *Chemical Science* **2022**, *13*, 3697−3705.
(100) Wellawatte, G. P.; Gandhi, H. A.; Seshadri, A.; White, A. D. A Perspective on Explanations of Molecular Prediction Models. *J. Chem. Theory Comput.* **2023**, *19*, 2149−2160.
(101) Bertoletti, N.; Chan, A. H.; Schinazi, R. F.; Anderson, K. S. Post-Catalytic Complexes with Emtricitabine or Stavudine and HIV-1 Reverse Transcriptase Reveal New Mechanistic Insights for Nucleotide Incorporation and Drug Resistance. *Molecules* **2020**, *25*, 4868.
(102) Zhou, P.; Chen, G.; Gao, M.; Wu, J. Design, synthesis and evaluation of the osimertinib analogue (C-005) as potent EGFR inhibitor against NSCLC. *Bioorg. Med. Chem.* **2018**, *26*, 6135−6145.
(103) Mishiro, K.; Nishii, R.; Sawazaki, I.; Sofuku, T.; Fuchigami, T.; Sudo, H.; Effendi, N.; Makino, A.; Kiyono, Y.; Shiba, K.; et al. Development of radiohalogenated osimertinib derivatives as imaging probes for companion diagnostics of osimertinib. *J. Med. Chem.* **2022**, *65*, 1835−1847.
(104) Gao, H.; Yang, Z.; Yang, X.; Rao, Y. Synthesis and evaluation of osimertinib derivatives as potent EGFR inhibitors. *Bioorg. Med. Chem.* **2017**, *25*, 4553−4559.
(105) da Silva, F. D. C.; de Souza, M. C. B.; Frugulhetti, I. I.; Castro, H. C.; Souza, S. L. D. O.; de Souza, T. M. L.; Rodrigues, D. Q.; Souza, A. M.; Abreu, P. A.; Passamani, F.; et al. Synthesis, HIV-RT inhibitory activity and SAR of 1-benzyl-1H-1, 2, 3-triazole derivatives of carbohydrates. *Eur. J. Med. Chem.* **2009**, *44*, 373−383.
(106) Ravetti, S.; Gualdesi, M. S.; Trinchero-Hernandez, J. S.; Turk, G.; Brinon, M. C. Synthesis and anti-HIV activity of novel 2', 3'-dideoxy-3'-thiacytidine prodrugs. *Bioorg. Med. Chem.* **2009**, *17*, 6407−6413.

https://doi.org/10.1021/acs.jcim.4c00720
J. Chem. Inf. Model. 2024, 64, 6799−6813

(107) Len, C.; Mackenzie, G. Synthesis of 2', 3′-didehydro-2', 3′-dideoxynucleosides having variations at either or both of the 2'-and 3′-positions. *Tetrahedron* **2006**, *62*, 9085−9107.
(108) Turk, G.; Moroni, G.; Pampuro, S.; Brinon, M. C.; Salomon, H. Antiretroviral activity and cytotoxicity of novel zidovudine (AZT) derivatives and the relation to their chemical structure. *Int. J. Antimicrob. Agents* **2002**, *20*, 282−288.
(109) Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B. A.; Thiessen, P. A.; He, S.; Zhang, J. Pubchem bioassay: 2017 update. *Nucleic acids res.* **2017**, *45*, D955−D963.
(110) Schierz, A. *PubChem Bioassay Data*, 2011 https://archive.ics.uci.edu/dataset/209/pubchem+bioassay+data; urldate: (June 12 2024).
(111) Rusinko, A.; Rezaei, M.; Friedrich, L.; Buchstaller, H.-P.; Kuhn, D.; Ghogare, A. AIDDISON: Empowering Drug Discovery with AI/ML and CADD Tools in a Secure, Web-Based SaaS Platform. *J. Chem. Inf. Model.* **2024**, *64*, 3−8.
(112) Rarey, M.; Dixon, J. S. Feature trees: a new molecular similarity measure based on tree matching. *J. Comput. Aided Mol. Des.* **1998**, *12*, 471−490.
(113) Software, C. C. *Cresset Flare*, 2022. https://www.cresset-group.com/software/flare/; urldate: (Apr 12 2024).
(114) Software, C. C. *Cresset Flare Docking*, 2022 https://www.cresset-group.com/software/flare-docking/; urldate: (Apr 16 2024).
(115) Blaschke, T.; Arús-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. REINVENT 2.0: an AI tool for de novo drug design. *J. Chem. Inf. Model.* **2020**, *60*, 5918−5922.
(116) Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P.; et al. Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. *Chem* **2018**, *4*, 522−532.

https://doi.org/10.1021/acs.jcim.4c00720
J. Chem. Inf. Model. 2024, 64, 6799−6813