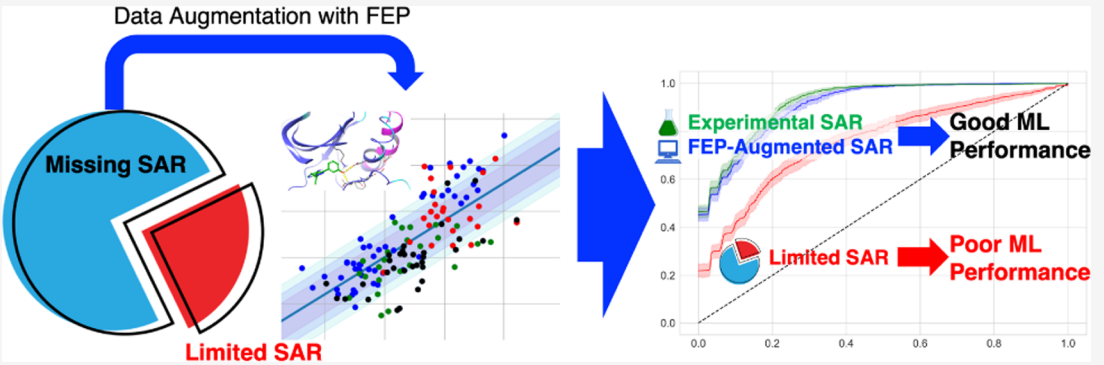
JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

FEP Augmentation as a Means to Solve Data Paucity Problems for Machine Learning in Chemical Biology

Pieter B. Burger,* Xiaohu Hu, Ilya Balabin, Morné Muller, Megan Stanley, Fourie Joubert, and Thomas M. Kaiser*

Cite This: J. Chem. Inf. Model. 2024, 64, 3812−3825 Read Online



ABSTRACT: In the realm of medicinal chemistry, the primary objective is to swiftly optimize a multitude of chemical properties of a set of compounds to yield a clinical candidate poised for clinical trials. In recent years, two computational techniques, machine learning (ML) and physics-based methods, have evolved substantially and are now frequently incorporated into the medicinal chemist's toolbox to enhance the efficiency of both hit optimization and candidate design. Both computational methods come with their own set of limitations, and they are often used independently of each other. ML's capability to screen extensive compound libraries expediently is tempered by its reliance on quality data, which can be scarce especially during early-stage optimization. Contrarily, physics-based approaches like free energy perturbation (FEP) are frequently constrained by low throughput and high cost by comparison; however, physics-based methods are capable of making highly accurate binding affinity predictions. In this study, we harnessed the strength of FEP to overcome data paucity in ML by generating virtual activity data sets which then inform the training of algorithms. Here, we show that ML algorithms trained with an FEP-augmented data set could achieve comparable predictive accuracy to data sets trained on experimental data from biological assays. Throughout the paper, we emphasize key mechanistic considerations that must be taken into account when aiming to augment data sets and lay the groundwork for successful implementation. Ultimately, the study advocates for the synergy of physics-based methods and ML to expedite the lead optimization process. We believe that the physics-based augmentation of ML will significantly benefit drug discovery, as these techniques continue to evolve.

INTRODUCTION

The key task of medicinal chemistry is to expeditiously optimize a wide array of physical, chemical, and biological properties for a chemical library leading to a quality clinical candidate for clinical trials. Medicinal chemistry has been largely unchanged in its approach to this task since Paul Ehrlich, Alfred Bergheim, and Sahachiro Hata first used phenotypic screening across hundreds of compounds to identify Salvarsan, the first effective treatment for syphilis.[1] Empirical and experimental knowledge of the medicinal chemist is used to drive the process of candidate selection in contemporary drug discovery programs, and this optimization approach can be sensitive to serendipity as a consequence (e.g., the magic methyl).[2−5] Target-based drug discovery added to the traditionally empirical methods of medicinal chemistry, but such reductionistic approaches did not displace emergent analysis in the lead optimization and candidate selection phases of preclinical development, where drug metabolism and pharmacokinetics along with pharmacodynamics (DMPK/PD) analysis frequently is not reducible to a single discrete target. This is especially true regarding tissue distribution behaviors and the complex interplay between transport proteins and

Received: January 12, 2024
Revised: April 1, 2024
Accepted: April 2, 2024
Published: April 23, 2024

https://doi.org/10.1021/acs.jcim.4c00071
J. Chem. Inf. Model. 2024, 64, 3812−3825

compound membrane flux.[6−10] However, the field of quantitative structure−activity relationships (QSAR) enabled the generation of mathematical trends which anticipated structural properties responsible for desirable and undesirable DMPK/PD behaviors within a series.[11] Machine learning (ML) was initially explored as a component within QSAR modeling. However, the success of ML across a wide array of properties including absorption, distribution, metabolism, excretion, and toxicity (ADMET), potency, selectivity, and synthetic problems resulted in ML in medicinal chemistry being established as an independent investigational field outside of traditional QSAR.[12−17] While ML can perform a virtual screen of millions to billions of drug-like compounds in a matter of hours to weeks, the learning aspect of ML requires the existence of abundant, high-quality information from which rules are learned.

This need for structure−activity relationship (SAR) training information can profoundly limit the application of ML methods in the lead optimization and candidate selection stages of a medicinal chemistry program, where a novel problem may suddenly arise. This is especially true for problems where limited SAR exists concerning the new impediment. If a team desires a predictive ML algorithm for a late-stage problem, then the acquisition of SAR tables capable of training ML algorithms can be slow and expensive. Therefore, methods that allow ML to work on problems with information paucity are desirable. We envision two approaches for the treatment of this problem: ML methods development (e.g., few-shot learning) or data set augmentation through the use of physics-based calculations. Machine learning methods which have been developed to tackle the lack of large data sets for key medicinal chemistry design concerns include a variety of deep learning architectures like few-shot/one-shot learning and transfer learning.[18,19] Transfer learning research applied to chemical biology seeks to identify methods of learning that can train in domains of information abundance like synthetic organic chemical reactivity and then work in low-information arenas like metabolism prediction.[20] One of the concerns for ML, especially when applied to information-poor property prediction, is that the ML model will have low generalization to the desired task. One of the key interests in few-shot learning is data augmentation to enhance the supervised learning experience.[18] We have been similarly interested in data augmentation approaches to information paucity, but our interest has been in physics-based methods as a means to perform data augmentation with directly relevant information.[21] Physics-based methods for the calculation in binding affinities, relative or absolute, have seen a substantial body of work evaluating free energy perturbation (FEP) methods as well as others like thermodynamic integration (TI).[22−27]

The utility of FEP is especially promising, where there is interest in predicting the potency of interaction between a small molecule and a single biological target. Such calculations can aid the design of compounds where a team desires to tune in potency on a desired target (e.g., to facilitate polypharmacology) or remove potency for an off-target.[28] As an example, kinase selectivity can be a core design task from the very beginning of a discovery program or a design problem that is revealed in the later stages of a drug development program.[29,30] Large-scale FEP initiatives as a means to correct an ML predictor through active learning have previously demonstrated their effectiveness in enriching active compounds during medicinal chemistry campaigns.[31−33] However, the ability to augment a sparse initial structure−activity data set for machine learning with theoretical activity data through FEP has not been well explored. We envisioned that FEP-augmented ML, using virtual data sets composed of hundreds of compounds' theoretical IC_{50} values calculated through FEP, could generate highly accurate predictor algorithms via ML in a much shorter time frame than wet lab experimental work. The ML algorithm trained on those hundreds of theoretical data points would then enable an ML exploration of tens of millions of possible drug-like synthetic targets quickly. The reason for such a hybrid approach is that the current computational expense of FEP precludes the exploration of millions of synthetic candidates within a given scaffold by using FEP alone. FEP has a well-established error profile,[22] and we explored the introduction of classification error that we would expect from a typical data set produced by FEP across several biological targets.[21] We found that a Naïve Bayes Network and a Random Forest (RF) could both be trained to produce viable predictive algorithms when using training data sets with an error profile resembling that of FEP. However, we desired a real test case of using FEP to generate theoretical IC_{50}'s for chemical structures and then evaluating the benefit of such a virtual data set for the training of ML algorithms. We therefore explored the cSrc series as reported by Apsel et al.[34] as a useful retrospective test case for FEP-augmented ML.

METHODS

Protein Preparation. The protein structures utilized in this study (PDBid: 3EN4, 3EN5, 3EN6, and 3EN7) were obtained from the Protein Data Bank (PDB). Subsequently, they underwent preparation using the Protein Preparation Wizard within the Maestro Molecular Modeling Suite (Schrödinger Release 2022-3: Maestro, Schrödinger, LLC, New York, NY, 2021). The protein structures were subjected to several preparation steps. First, hydrogen atoms were added and disulfide bonds were created. The protonation states of heteroatoms were then generated using *Epik*, with a pH value set to 7.0 and a pH tolerance $\Delta pH \pm 2.0$, corresponding to a charge state population cutoff of $p \sim 1\%$ ($p = \exp(-\ln(10)\,\Delta pH)$). Water molecules located beyond 5 Å of heteroatoms were removed. Additionally, any missing side chains were added using *Prime*. Special attention was dedicated to residues within the active site during the preparation process. Specifically, residue D404 of PDBid 3EN4 was resolved by considering two equally populated average occupancy positions. Following a thorough visual inspection, the first position was selected for further analysis. To optimize the protonation state and orientation of side chains at pH 7.0, the Interactive H-Bond Optimizer, which used *PROPKA*, was employed. Visual inspection was used to do the final protonation and side-chain orientation assignment with special focus given to histidines, charged amino acids, and residues within the active site. Restrained minimization of hydrogen atoms only was followed by a full atom minimization with a convergence of heavy atoms to root mean square deviation (RMSD) 0.3 Å using the OPLS4 force field.

Loop Generation. The missing loops in the PDB structures (PDBid 3EN4 and 3EN5) were generated by using the homology modeling module within the Maestro Molecular Modeling Suite. The crystallized domain sequences of PDBid 3EN4 and 3EN5 were extracted from PDBid 3EN7, serving as references for generating the homology models.

https://doi.org/10.1021/acs.jcim.4c00071
J. Chem. Inf. Model. 2024, 64, 3812−3825

Utilizing PDBid 3EN4 and 3EN5 as templates, models of the 3EN4 and 3EN5 structures, including their respective ligands, were generated. The model settings utilized in this study involved minimizing all nontemplate residues to optimize their positions and conformations within the model. Rotamers were preserved to retain the original orientations of conserved residues, while side chains (excluding conserved residues) were optimized to refine their orientations. Additionally, insertions were limited to a maximum of 20 residues, controlling the length of the added segments during the model generation process. Using the energy-based method, a single model was generated while leaving all other parameters at their default values. The resulting models were then prepared for simulations by following the procedures outlined in the protein preparation section.

Molecular Dynamics (MD). We prepared the protein structures for molecular dynamic (MD) simulations utilizing the System Builder module found in Desmond (Schrödinger Release 2022-3: Desmond Molecular Dynamics System, D. E. Shaw Research, New York, NY, 2022. Maestro-Desmond Interoperability Tools, Schrödinger, New York, NY, 2022). An orthorhombic box shape was used to solvate the system using an SPC water model with a buffer size of 10 Å. Neutralizing atoms were added, and a 0.15 M Na^+/Cl^- concentration was applied. We chose the OPLS4 force field for the simulations. The protein−ligand complex, once solvated, was imported into Desmond's Molecular Dynamics module. The parameters for the simulation of the molecule's dynamics were then defined as follows: The simulations were conducted under the NPT ensemble class at 300 K and 1.01325 bar. Prior to the simulation, the model was relaxed using the relaxation protocol provided by Desmond. Simulations were executed for the duration of 10 or 500 ns. The 10 ns simulation was followed by an evaluation and then capture of the protein structure at the 1.2 ns mark. The remaining parameters remained unaltered and are documented in the Desmond user guide. The results of the simulations are presented in the simulation report, which can be found in the Supporting Information section under MD reports.

Ligand Preparation. All ligands were prepared using the *Ligprep*. The OPLS4 force field was selected, compounds were desalted, the specified chiralities were retained, and the protonation states were determined at pH 7.0 and ΔpH of ±2.0 using *Epik*. In the case where the chiral centers were not specified all possible stereoisomers associated with the chiral centers were generated manually and added to the compound list. All other parameters were unchanged. All structures used in FEP+ calculations can be found in the Supporting Information (Compound List).

Molecular Docking. All molecular docking in this study was performed using Glide Schrödinger Release 2022-2: Glide, Schrödinger, LLC, New York, NY, 2022. Two general methodologies were used during docking, constrained and unconstrained docking. The methods for both methods are described below:

Unconstrained Docking. The Receptor Grid Generator panel was used to generate the Grid for docking. The geometric center of the reference ligand within the protein−ligand complex was selected and used as the center of the docking grid. Aromatic hydrogens were included as hydrogen bond donors, and halogens were included as acceptors. No other changes were made to the standard Grid generation protocol (Schrödinger Documentation). The Standard Precision (SP) docking scoring function was used during docking, with ligands being treated flexibly. Depending on the aim of the docking study, between 1 and 5 poses were written out after postdocking minimization (between 1 and 5 poses were subjected to minimization). All other parameters were unchanged from those in the standard Glide docking protocol.

Constrained Docking. The Receptor Grid Generator panel was used to generate the grid for docking. The geometric center of the reference ligand within the protein−ligand complex was selected and used as the center of the docking grid. Aromatic hydrogens were included as hydrogen bond donors and halogens as acceptors. The van der Waals radii of the receptor atoms were scaled to 0.8 to soften the potential of the nonpolar part of the receptor. No other changes were made to the standard grid generation protocol (Schrödinger Documentation). The SP docking scoring function was used during docking, with ligands being treated flexibly. Depending on the aim of docking study, between 1 and 5 poses were written out after postdocking minimization (between 1 and 5 poses were subjected to minimization). Core constraints were placed on ligands being docked by selecting the ligand within the protein−ligand complex as a reference. A core pattern comparison was applied to ligands being docked, and the overall atom positions were restricted between 0.4 and 0.7 Å of the reference compound's position. The core atoms were determined by applying the maximum common substructure (MCS) algorithm. Compounds that failed to dock using the above parameter were aligned using the Ligand Alignment protocol. In this case, the ligand from the protein−ligand complex was selected as a reference structure and aligned using MCS as a method of alignment. This was followed by a visual inspection of the protein−ligand complex to identify gross overlap of atoms within the pocket that would allow for failures during simulations. All other parameters were unchanged from the standard Glide docking protocol.

Positioning of the Reference Structures. For Scaffold 1, the starting point for the MD simulation was determined by docking compound PP56 to the 3EN4 structure. Figure 1C depicts the reference structures of all 3 scaffolds used in the FEP calculations. The docking process utilized the MCS as the constraint criteria, aligning PP56 with compound PP121 co-resolved within this structure. In the case of the type 1 1/2 kinase inhibitor structure PDBid 3EN5, compound PP494 was used as the starting point for the MD simulation. For Scaffold 2, the resolved conformation of PP121 was used in the simulations involving the PDBid 3EN4 structure, whereas constraint molecular docking, as described above, was used to dock PP121 to the 3EN5 structure. For Scaffold 3, a similar approach was employed to position compound PP102 in the docking reference structure in both the 3EN4 and 3EN5 structures. During the docking process, we permitted the generation of two to four docked poses for each compound. From these generated poses, we selected two representative poses that captured the dual binding pose observed in some compounds. In cases where a compound could only adopt a single binding mode, we chose the pose with the lowest docking energy i.e., a compound like PP56 that has a para-substituted phenol. Next, for compounds capable of assuming two different binding poses, we selected the lowest-energy docking pose for each binding mode (similar to S1_a and S1_b). However, in cases where compounds did not produce both of the expected binding poses during the docking process, we conducted further investigation into their potential binding

https://doi.org/10.1021/acs.jcim.4c00071
J. Chem. Inf. Model. 2024, 64, 3812−3825

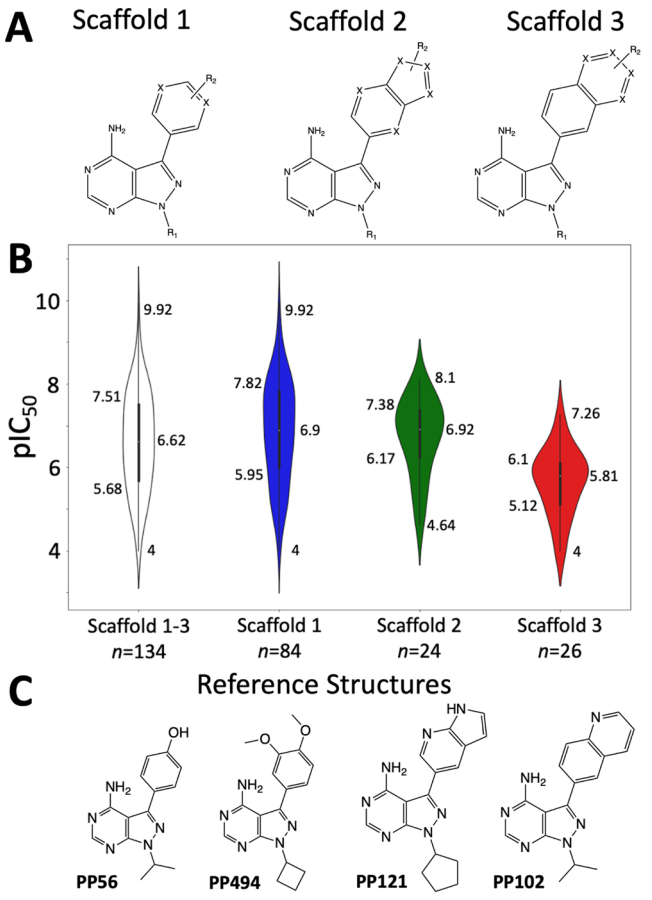


Figure 1. (A) Illustration depicting the generic framework of Scaffolds 1−3, where R_1 and R_2 exhibit potential for multiple substitutions. Moreover, the ring systems commonly undergo substitutions with heteroatoms. The complete list of structures is provided in the Supporting Information (Compound List). (B) Violin plots representing the IC_{50} values of the curated cSrc data set. (C) Depiction of the reference structures used in FEP calculations for Scaffolds 1−3.

modes. If a docked compound pose resulted in significant clashes with the protein, then only one pose was included in the analysis. However, if a second binding pose was not obtained due to minor clashes, then we manually generated a second binding pose. To ensure consistency, all compounds were aligned either using ligand alignment tools or via manual intervention. This alignment process aimed to ensure that the maximum common core between the target and reference compounds fell within a range of 0.1−0.5 Å, producing a suitable input for FEP. The Supporting Information includes all of the input conformations used in this study; see Input Structure Data.

Boltzmann Distribution of Multiple Binding Poses. We used the Boltzmann distribution function to calculate the contribution of each binding mode/state to the overall $\Delta\Delta G$. Most compounds in Scaffold 1 exhibited two binding poses, similar to those of S1_a and S1_b. However, the number of modes or states could increase to 4 or 8 for some compounds depending on the chirality of the molecules. The exact means of determining the contribution can be found in the KNIME workflows provided. Additionally, the KNIME workflows also encompass the conversion process from $\Delta\Delta G$ to ΔG, followed by the conversion into K_i and IC_{50} values.

FEP+ Calculations. The relative binding free energy (RBFE) calculation method as implemented in FEP+ was used to calculate the $\Delta\Delta G$ between the selected reference structures and the compounds of interest. The protein−ligand complex used resulted from the methods described in the MD and docking sections above. The protein−ligand complex and docked compounds were imported into the FEP+ panel. We used a single edge map topology by generating a star map with the reference structure being PP56, PP494, PP121, and PP102 for Scaffolds 1 (type 1 inhibitor), 1 (type 1 1/2 inhibitors), 2, and 3, respectively. The OPLS4 Force Field was used, and missing parameters for ligands were generated using the Force Field Builder module (if not present in the default OPLS4 Force Field). The default FEP protocol settings were used, which included the μVT ensemble with a 5 ns simulation time, 12 lambda windows for neutral, 16 lambda windows for core-change, and 24 lambda windows for charge change perturbations. The system was built using a 5 Å buffer radius, and the relative solvation free energies were calculated. For perturbations that did not have good conversion, we increased the simulation time to 10 or 20 ns depending on if convergence was reached or not. Energy convergence is determined over the last nanosecond of the simulation. Two criteria are used to establish if energy convergence was reached, a global convergence which is calculated using the $\Delta\Delta G$ change divided by the time span (last nanoseconds of the simulation) and a local convergence which is calculated by maximum $\Delta\Delta G$ change divided by the time span (last nanoseconds of the simulation). Simulations that showed a global and local variation value less than 0.3 kcal mol^{-1} ns^{-1} for both legs of the simulation were considered to have reached energy convergence. Similarly, simulations that had a global variation for both legs less than 0.3 kcal mol^{-1} ns^{-1} and a local variation larger than 0.3 kcal mol^{-1} ns^{-1} for one or both legs were also deemed to have reached energy convergence. Additional information can be found in the Schrodinger Documentation.

Absolute Binding Free Energy (ABFE). The absolute binding free energy (ABFE) calculation method as implemented in FEP+ was used to calculate the ΔG values of the compounds of interest. The protein−ligand complex used resulted from the methods followed as described in the MD and docking sections above. The OPLS4 Force Field was used and missing parameters for ligands were generated using the Force Field Builder module (if not present in the default OPLS4 Force Field). The default parameter settings were used, which included the μVT ensemble method with a 5 ns simulation time and a 1 ns MD simulation time (Random Seed 2007).

Machine Learning. The ML methods section is outlined in five specific parts: initial data processing and molecular representation, partitioning of data into training and testing sets, creation of the RF algorithm, and the parameters used to produce the 100 distinct training and test sets. Workflows and input data are provided in the Supporting Information. All cheminformatics workflows and analyses, along with the ML processes, were conducted by using KNIME version 4.3.1.

Data Preprocessing and Molecular Representation. The cSrc data set was sourced from the ChEMBL database. The data were organized in ascending order according to the experimentally determined IC_{50} values and then filtered to retain only those compounds with explicitly stated IC_{50} values measured in nanomolar (nM) units. Duplicates were then

https://doi.org/10.1021/acs.jcim.4c00071
J. Chem. Inf. Model. 2024, 64, 3812−3825

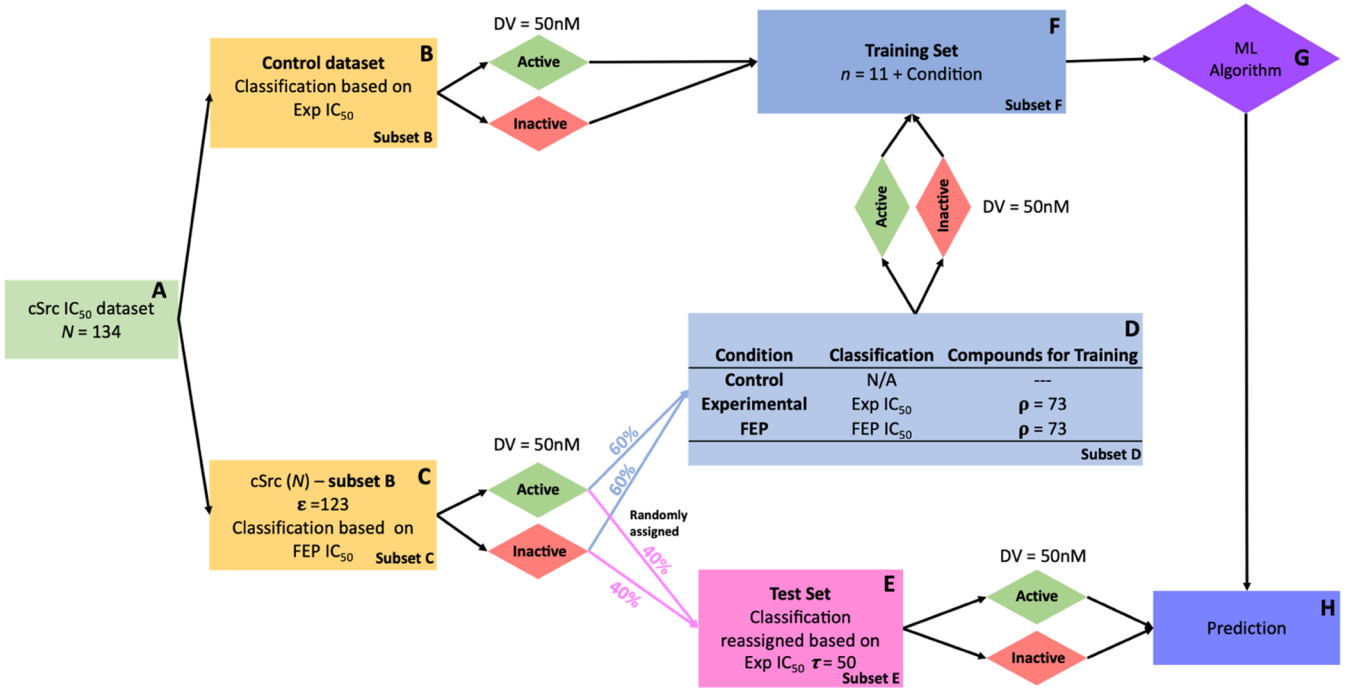


Figure 2. Flow diagram presenting the three experimental conditions employed to assess the efficacy of ML when enhancing activity data sets with FEP-derived activity. (A) The input data set for ML consists of Scaffolds 1−3 (N = 134). This data set was divided into two distinct subsets, labeled as subsets B and C. (B) Subset B is represented here, by the Control data set (C) Subset C, which includes the remaining 123 compounds, is presented here. This subset is further divided into subsets D and E. The classification of active or inactive is made within Subset C based on FEP IC_{50} values, using a decision value (DV) of 50 nM. By maintaining a balanced 60/40% split, we generate Subset D (ρ = 73) and the test set, Subset E (τ = 50). (D) Subset D, comprising 73 compounds, is categorized based on the three experimental conditions outlined in this paper. Depending on the condition and the classification method, either 73 or 0 compounds are added to Subset F. (E) Subset E, consisting of 50 compounds, constitutes the test set. The compounds in Subset E are reclassified as either active or inactive based on their experimentally determined IC_{50} values. (F) Subset F comprises either 84 or 11 compounds. The 11 control compounds, from subset B, are combined with Subset D. (G) ML is performed as described in the Methods section. (H) Predictions are made based on experimentally determined IC_{50} values.

removed and only compounds belonging to the pyrazolopyrimidine scaffold were retained. Compounds with formal changes were also removed. The final data set consisted of 134 compounds. The Indigo canonical smiles node was used to generate canonical smiles, and the RDkit fingerprint module was employed to calculate the Morgan 2 fingerprints.

Training and Test Set Generation. Three ML experiments were designed to evaluate the FEP augmentation approach using three distinct training sets, the Control, the Experimental-Augmented, and the FEP-Augmented data set. A schematic representation of the ML experiments is given in Figure 2. It is important to note that the Control data set was incorporated into each of the three ML experiments and was consistently generated to ensure comparability across the experiments. Below are the specifics for creating each data set and the corresponding ML experiments:

Control Data Set. We selected 11 compounds, based on the assumption that having at least one reference compound and 10 control compounds would suffice for validating the system's FEP predictability. The 11 compounds chosen to represent the initial SAR consisted of four reference compounds, namely, PP56, PP494, PP121, and PP102. These four compounds were consistently included as a subset of the 11 compounds. The remaining seven compounds were selected randomly, with five compounds chosen from Scaffold 1 and one compound each selected from Scaffolds 2 and 3. This partitioning of the 11 compounds was conducted to reflect the diversity of the different scaffolds within the data set. By ensuring a balanced representation of the scaffolds, we facilitated training the algorithms on a representative set of compounds that reflected the populations of the three scaffolds. The Control data set was then classified into the class Active and Inactive using a decision value (DV) of 50 nM determined by *in vitro* experiments. A compound with an IC_{50} value below or equal to the DV was classified as being Active, whereas a value above the DV was deemed Inactive.

Experimentally Augmented Data Set. Following the generation of the Control data set, the remaining 123 compounds were classified into Active and Inactive, using a 50 nM DV based on the FEP-derived IC_{50} values. The balance for the testing and training sets was set using these FEP activity classes. The reasoning for using FEP-calculated values is that we wanted to compare the performance of a data set constructed through FEP (and its associated activity population distribution) with how those molecules would actually inform the ML algorithm if the experimental activity values were determined. It was therefore decided that the FEP-derived IC_{50} values would serve to construct the balanced test and train sets as a more representative experiment of a true prospective test using FEP. We also evaluated the ML performance using experimentally determined IC_{50} values to set the balanced training and test set with the results shown in the Supporting Information, Figure S1. Next 60% of compounds in class Active and Inactive were split out randomly resulting in a subset of 73 compounds. These 73 compounds were reclassified into classes Active and Inactive

https://doi.org/10.1021/acs.jcim.4c00071
J. Chem. Inf. Model. 2024, 64, 3812−3825

using a DV of 50 nM based on the Experimentally determined *in vitro* IC_{50} values. Finally, the Experimentally Augmented data set is generated by combining the 73 reclassified compounds with the Control data set.

FEP-Augmented Data Set. Following the generation of the Control data set, the remaining 123 compounds were classified into Active and Inactive, using a 50 nM DV based on the FEP-derived IC_{50} values. Next 60% of compounds in class Active and Inactive are split out randomly, resulting in a subset of 73 compounds. Finally, the FEP-Augmented data set is generated by combining the 73 classified compounds with the Control data set.

Test Data Set. The test set was constructed to be identical across all three training sets. Following the separation of the control data set, and categorization of the remaining 123 compounds into Active and Inactive based on a DV of 50 nM using FEP-derived IC_{50} values, 40% of each class was randomly assigned to the test set. The resulting test set of 50 compounds was then reclassified into Active and Inactive categories based on their experimentally determined IC_{50} values for evaluation.

Random Forest Model Generation. An RF algorithm was generated for each training set using Morgan 2 fingerprints as the independent variable set and the category Active/Inactive as the dependent variable. The RF algorithm was generated using the Random Forest Learner available in KNIME. The RF algorithm was trained using 1000 trees and using the Gini Index as the split criterion. A static random seed, 1602276747675, was used in generating all of the algorithms. The resulting classifier algorithm was fed into the Random Forest Predictor module for the corresponding algorithm, and the performance of the RF was evaluated on the test set by using the ROC Curve (Java Script) and Enrichment Plotter modules.

Parameters to Generate 100 Distinct Training and Test Sets. This process of generating the test and training set for the three ML experiments was repeated 100 times using a different random seed for each iteration. The random seeds were produced by generating unique integers between 0 and 100,000 using a seed number 2022 in the Random Generator Node in KNIME. This ensured that for each ML run the compounds in the Control data set were identical as well as the test set. Moreover, this also ensured that the compounds in the training sets for the Experimental-Augmented and FEP-Augmented data sets were the same. Finally, the sklearn library and seaborn package were utilized within Python to generate an average ROC curve, accompanied by a 95% confidence interval (CI).

RESULTS AND DISCUSSION

In this work, we built upon our previous findings demonstrating that ML algorithms maintain predictive power when an FEP error profile is incorporated into an activity data set.[21] To demonstrate a practical application of our methodology, we retrospectively examined a completed lead optimization program to generate a real-world scenario experimentally testing FEP data set augmentation. The overall aim of our study was to establish minimum criteria as guidelines for data augmentation in a practical medicinal chemistry setting.

We envisioned a hypothetical early-stage hit-to-lead program and identified what would be required to perform a data augmentation program. We identified two broad criteria that would ensure successful implementation. First, we stipulated that the FEP-ML augmentation experiment be performed on a series that contained sufficient SAR for an initial control set and a retrospective testing set. Taken from this initial series, the starting control set of compounds that resembled a hit series with a restricted SAR would be on the order of 10−20 compounds. The reason for this number is 2-fold. A compound library of 10−20 compounds represents a very early-stage medicinal chemistry optimization program, and having at least 10−20 compounds is essential for carrying out the required controls and adhering to good RBFE practices[26] (Schrödinger Release 2022: FEP+, Schrödinger, LLC, New York, NY, 2022). Second, we expected a medicinal chemistry program at the beginning of a hit-to-lead campaign to have high-quality X-ray structures co-resolved with ligands that are representative of the SAR. This will enable the execution of accurate FEP calculations. Given these two broad constraints, we defined a set of minimum criteria necessary for our retrospective validation project and its selection, which are detailed below:

1. The series must consist of more than 100 compounds with each compound having a defined activity against the target of interest.
2. All compounds should belong to a congeneric series and maintain a similar charge profile.
3. The activity range should span across multiple orders of magnitude and the activity values should be recorded under the same conditions.
4. The performance of ML on the total activity data set should show robustness.
5. At least one X-ray structure should be available, featuring representative structures of the scaffold of interest.
6. The target protein should be of a manageable size to curtail computational time and expense.

We chose to evaluate a set of cSrc inhibitors based on the study from the Shokat lab by Apsel et al.[34] because it met all of our established criteria and offered a well-defined set of compounds. This enabled us to construct a retrospective, hypothetical, real-world test scenario. cSrc is a nonreceptor tyrosine-protein kinase that plays a multifaceted role in numerous cellular pathways, regulating cell survival, migration, and proliferation.[35] The cSrc protein is a small, well-characterized cytoplasmic kinase and serves as an ideal test bed for target-based drug design methods.

Data Set Refinement. In the study by Apsel et al.,[34] 171 compounds were synthesized and their IC_{50} values were determined. For our study, we refined this data set to meet the above set criteria. We retained only those compounds that belong to the pyrazolopyrimidine scaffold shown in Figure 1A due to our focus on generating methods for lead optimization rather than scaffold hopping. In this study, we removed compounds with formal charges to minimize the increased error associated with changes in the net charges of molecules during RBFE calculations. We want to note that, in the later versions of the FEP+ implementations, the error associated with charged compounds is significantly reduced by applying the pK_a correction for different charged states of the same compound.[36] These improvements will allow for the inclusion of compounds with different net charges. Next, we segmented our scaffold into three distinct subscaffolds, labeled as Scaffolds 1−3 (Figure 1). This partitioning was carried out to reduce the number of ring opening and closing events in order to limit the primary source of the errors to R-group changes instead of

https://doi.org/10.1021/acs.jcim.4c00071
J. Chem. Inf. Model. 2024, 64, 3812−3825

mixing in additional errors that can result from the changes of the core. The resulting data set is composed of a total of 134 compounds that belonged to the pyrazolopyrimidine scaffold. Assessment of the activity landscape revealed a broad range of IC_{50} activity spanning from 0.12 to 100,000 nM (pIC_{50} 4−9.92; Figure 1B). This enabled us to examine the performance of FEP across a spectrum of activities, mirroring that of a typical lead optimization program.

Experimental Design for Machine Learning. To show the enhancement of accuracy afforded by augmentation through experimental or calculated IC_{50} values, we assessed the performance of ML using a classification method on the newly generated data set to ensure the SAR was solvable by ML (*vide infra*). The methods were similar to the approach taken as described in our earlier work as well as in the Methods section.[21] We designed three distinct ML experiments, each defined by a unique data set. The outlines and objectives of these experiments and their data sets are outlined as follows:

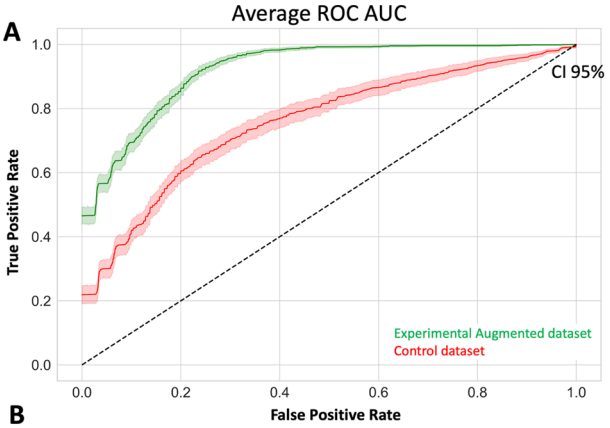
(1) Control data set: A small data set representative of a limited starting point for medicinal chemistry where IC_{50} values are obtained from *in vitro* studies for a small number of compounds (n = 11).
Aim: Generate ML classifiers using only *in vitro* derived IC_{50} values from the initial limited SAR.
Objective: Assess the contribution of the initial SAR data on the predictive performance of the other classifiers produced.

(2) Experimental-augmented data set: A data set where all IC_{50} values for all compounds are obtained from *in vitro* studies. (n = 134).
Aim: Generate ML classifiers using the full set of IC_{50} values determined from *in vitro* experimental studies as training data.
Objective: Assess the performance of ML classifiers trained by using experimental IC_{50} values, which serve as a benchmark for comparison to the other two ML model conditions.

(3) FEP-augmented data set: A data set consisting of the Control data set (IC_{50} values derived from *in vitro* studies) that is augmented with the remaining compounds in the series using IC_{50} values determined by FEP calculations.
Aim: Generate ML classifiers using an FEP-augmented data set as training data.
Objective: Assess and compare the performance of ML classifiers generated from a FEP-Augmented data set and compare it to the algorithms generated from the Experimental and the Control data sets.

A schematic representation of the workflow employed to meet our stated goals is delineated step-by-step in Figure 2. The initial activity data set, A (N = 134), is partitioned into two distinct subsets. Subset B comprises the Control data set of 11 compounds with their experimentally determined IC_{50} values while subset C encompasses the remaining 123 compounds. It is subset C that will be used to create the hidden test set as well as the augmenting sets of compounds. This subset is split into two groups using a 60/40 partition. 60% of the data will be used to augment the training set of 11 compounds with either the experimental IC_{50} or the FEP-calculated IC_{50}. However, the control data set experiment (n = 11 training compounds) will have this 60% Subset C discarded. The remaining 40% of subset C in Figure 2 will be used as the same test set across all three experiments. Compounds will be classified using a DV of 50 nM where an $IC_{50} \leq 50$ nM indicates an active compound. A random selection process was used to partition the training and testing compounds, and this same random process was employed across the three experiments.

To achieve statistical significance, the above process was repeated 100 times for each experiment. For each repetition of an experiment, a different random seed number was utilized to ensure a random assignment. Figure 3A displays the average



B

Dataset	ROC AUC	Enrichment at 10%	Enrichment at 20%
	Mean ± Std Dev	Mean ± Std Dev	Mean ± Std Dev
Experimental Augmented	0.921 ± 0.034	3.18 ± 0.48	2.73 ± 0.43
Control	0.794 ± 0.101	2.57 ± 0.78	2.26 ± 0.46

Figure 3. (A) Average ROC curve for all 100 iterations ran under the Experimental-Augmented (green) and Control data set (red) conditions are shown with a 95% confidence interval (CI). The Control data set (red) algorithms were trained on only 11 compounds, whereas the Experimental-Augmented (green) algorithms were trained on 84 compounds. All activity values used for training and testing were experimentally derived IC_{50} values. The test set consisted of 50 compounds for both the Control and Experimental-Augmented algorithms. (B) Mean ROC AUC and standard deviation, together with the Enrichment Factors at 10 and 20% are provided for a DV of 50 nM.

receiver operating characteristic area under the curve (ROC AUC) with a 95% confidence interval. The unaugmented Control data set of 11 compounds exhibited a mean ROC AUC of 0.794 when predicting the activity class of the hidden training data set with a standard deviation of ±0.101. With this baseline value in hand, we could turn our attention to how the ROC AUC, sensitivity, and specificity of the algorithm improved through augmentation by either the experimentally determined IC_{50} data or the theoretical IC_{50} data as calculated by FEP.

Protein Structure Evaluation. To generate the FEP IC_{50} values for the augmented set, we needed to prepare the crystal structures that were needed for the calculations. An evaluation of the suitability of cSrc crystal structures revealed four crystal structures of the cSrc kinase domain, with PDBid's 3EN4, 3EN5, 3EN6, and 3EN7—that were resolved with ligands representing Scaffolds 1−3, as illustrated in Figure 4. The protein structures had moderate quality, with their resolutions ranging between 2.39 and 2.81 Å (Figure 4). A number of unresolved loops in the *N*-terminal kinase domains of structures 3EN6 and 3EN7 impose significant difficulties in RBFE calculations. As a result, these structures were excluded

https://doi.org/10.1021/acs.jcim.4c00071
J. Chem. Inf. Model. 2024, 64, 3812−3825

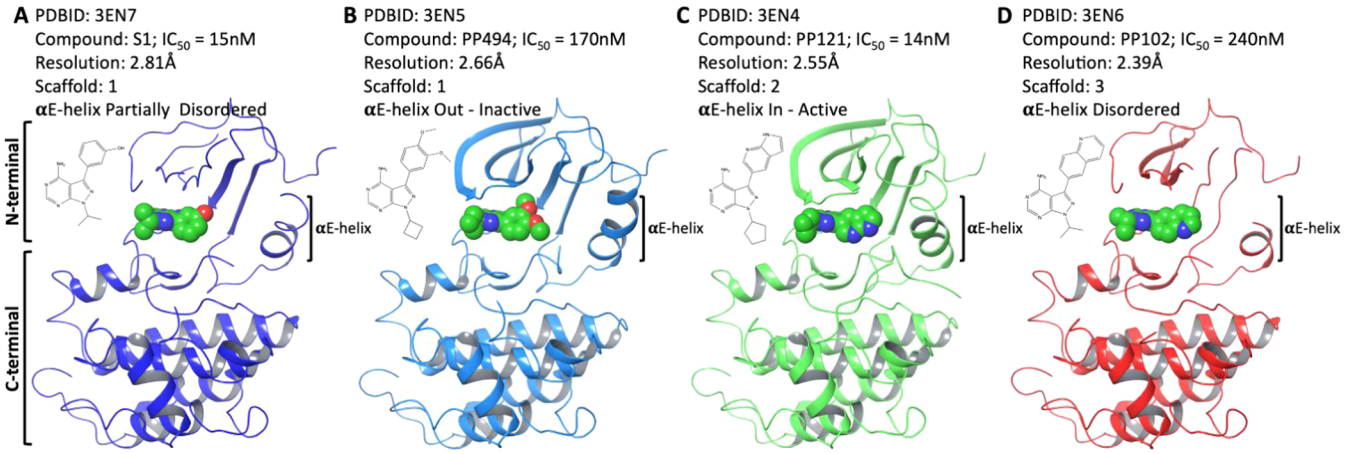


Figure 4. (A–D) Schematic illustrations of the four cSrc structures with their resolved inhibitors are presented with the inhibitors depicted in green. A two-dimensional (2D) structure of each inhibitor is provided, with the N- and C-terminals, as well as the αE-helix, highlighted.

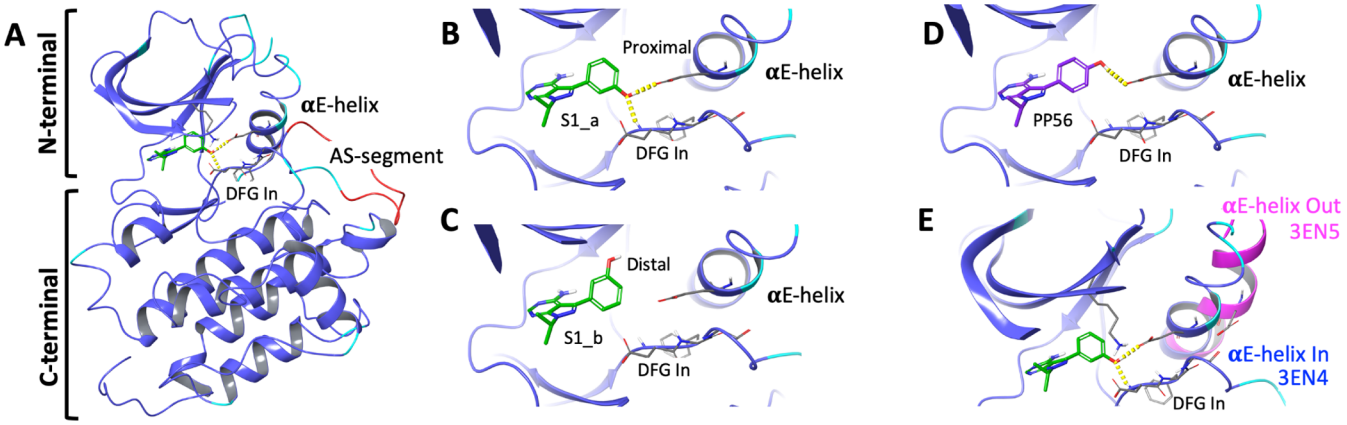
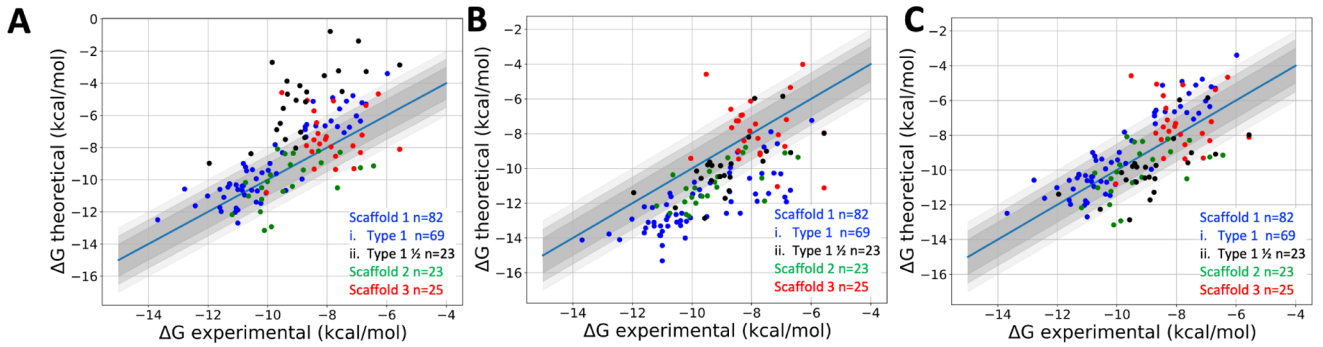


Figure 5. (A) Cartoon representation of the kinase domain of cSrc bound with compound S1. The red loop represents the computationally derived part of the activation segment. (B) Compound S1 (colored green), labeled as S1_a in the alternative binding pose. The hydroxyl group of the phenol ring is oriented toward (proximal) the DFG motif. (C) Compound S1 (colored green), labeled as S1_b and depicted as resolved in structure 3EN4. The hydroxyl group of the phenol ring is oriented away (distal) from the DFG motif. (D) The selected reference structure compound PP56 (colored purple), with only one possible binding pose with regard to its phenol moiety. (E) Overlay of compound S1 in the proximal binding pose (S1_a) within the 3EN4 (type 1; blue) and 3EN5 (type 1 1/2; purple) structure. Note the movement of the α-E helix and the loss of a hydrogen bond with E310.

from use in calculations in this study (Figure 4). None of the structures were resolved with intact activation segments, and each structure was processed using Schrödinger software (see the Methods section for details). We observed that two distinct types of kinase inhibitors were resolved within cSrc giving rise to two unique protein–ligand kinase conformational states. PDBid 3EN4 was co-resolved with a type 1 inhibitor with the αE-helix in the inward and active kinase state (Figure 4C). In contrast, the PDBid 3EN5 structure contained a type 1 1/2 inhibitor resolved in the active site, with the αE-helix in the outward and inactive kinase states (Figure 4B). The PDBid structures 3EN6 and 3EN7 had either a partially resolved or an unresolved αE-helix, making the determination of the inhibitor type uncertain (Figure 4A,D). Finally, we observed ambiguities during the evaluation of the binding poses of ligands S1 and PP494, specifically regarding the uncertainty in the binding orientation or pose of the phenol and 3,4-dimethoxyphenol rings, respectively (Figure 4). The above evaluation satisfied our criteria, 5 and 6, for proceeding with FEP calculations, despite some challenges that will be addressed in the relevant sections below.

FEP Preparation and Calculations. The selected structures 3EN4 and 3EN5 were prepared for free energy calculations as described in the Methods section. Since FEP calculations ideally prefer structures that are well resolved with clear binding poses of the reference ligands, we examined the binding poses of each of the four resolved compounds. It was found that compounds representing Scaffold 1, namely, compounds S1 and PP494, had ambiguous orientations of the phenol and 3,4-dimethoxyphenol rings, respectively. To identify the lowest binding free energy pose for each of these compounds, we used ABFE calculations through FEP+. We calculated the ΔG for both conformations of compound S1 within the 3EN4 structure and found that compound S1 had significantly lower free energy when the hydroxyl was oriented proximally (S1_a) compared to when it pointed distally to the DGF region (S1_b; Figure 5). The S1_a pose exhibited a ΔG of -12.47 ± 0.12 kcal/mol, in contrast to the S1_b pose, which had a ΔG of -9.97 ± 0.02 kcal/mol. This 2.5 kcal/mol difference in the ΔG between the S1_a and S1_b binding poses suggests inadequate ring orientation sampling during the ABFE calculations. If sufficient sampling of both the S1_a and

https://doi.org/10.1021/acs.jcim.4c00071
J. Chem. Inf. Model. 2024, 64, 3812–3825



D

	3EN4: Type 1					3EN5: Type 1 ½					Mixed Type Inhibition				
	SC1 ALL	SC1 Type 1	SC1 Type 1½	SC2 ALL	SC3 ALL	SC1 ALL	SC1 Type 1	SC1 Type 1½	SC2 ALL	SC3 ALL	SC1 ALL	SC1 Type 1	SC1 Type 1½	SC2 ALL	SC3 ALL
MUE*	1.87	1.14	3.72	1.20	1.51	2.09	2.30	1.41	1.72	1.61	1.22	1.14	1.41	1.20	1.51
$R^{2\#}$	0.67	0.82	0.48	0.30	0.00	0.56	0.59	0.46	0.46	0.01	0.60	0.82	0.46	0.30	0.00
RMSE*	2.44	1.36	4.05	1.53	1.89	2.36	2.58	1.66	1.96	2.10	1.45	1.36	1.66	1.53	1.89

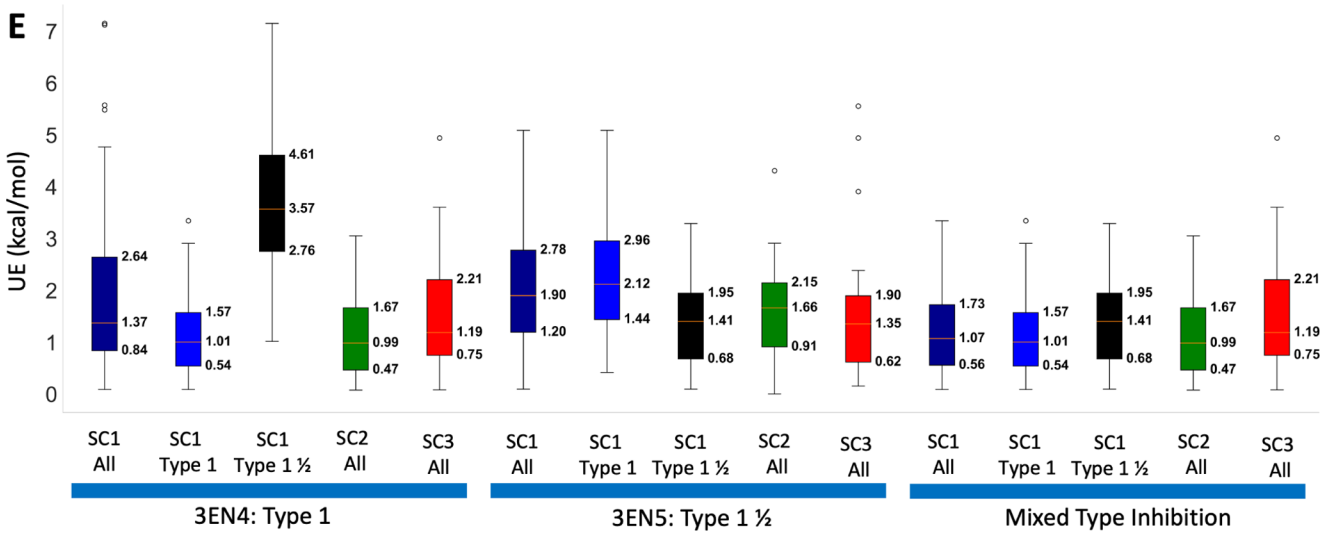


Figure 6. (A–C) Plots of the ΔG derived from experimentally determined IC_{50} values vs ΔG theoretically determined values. (A) The data presented are for ΔG values, which are derived from the 3EN4 structure representing type 1 inhibitors. The blue dots symbolize compounds from scaffold 1, which belong to the type 1 inhibitor class. The black dots symbolize compounds from scaffold 1, which belong to the type 1 1/2 inhibitor class. The green and red dots symbolize compounds from scaffolds 2 and 3, respectively. (B) The data presented is for ΔG values, which are derived from the 3EN5 structure representing type 1 1/2 inhibitors. (B) follows the same color scheme as (A). (C) Data plotted are for ΔG values, derived from a combination of both type 1 and type 1 1/2 structures. (C) follows a similar color scheme to (A). (D) Statistical analysis of the FEP predictions for each set of compounds. Generally, the color scheme for (D) aligns with that of (A). (E) Boxplot displaying the ΔG UE associated with scaffolds 1, 2, and 3 for the data represented in (A–C). Dark blue boxplots signify the unsigned error (UE) when both type 1 and 1 1/2 are combined for scaffold 1. Generally, the color scheme for (E) aligns with that of (A). The MUE and RMSE are denoted in * kcal/mol. # R^2 is the coefficient of determination.

S1_b pose was achieved, the resulting ΔG would converge. Moreover, the 2.5 kcal/mol difference in the ΔG of S1_a and S1_b also suggests that other compounds in the Scaffold 1 series could display dual binding modes, and both ring orientations as initial binding poses should be evaluated. Therefore, we operated under the assumption that compounds within Scaffold 1 could potentially adopt two binding poses, similar to that observed for compounds S1_a and S1_b, though this is not universally true (i.e., large substitution on the benzyl ring can cause significant steric clashes and only allow one binding pose). Due to the significant increase in the computational cost of having to calculate $\Delta\Delta G$ values for two binding poses, we further investigated ring orientation sampling in both ABFE and FEP calculations under divergent set of conditions (refer to the provided Supporting Information for more details). We found no conditions under which sufficient ring sampling was achieved and accordingly decided to calculate $\Delta\Delta G$ values for both possible poses for each of the compounds belonging to Scaffold 1. To further mitigate the uncertainty around the ambiguous binding of compound S1 as the reference structure, we revisited the SAR. Our objective was to identify a reference structure with only one binding pose. We identified compound PP56, which had an IC_{50} value of 75 nM, as a suitable reference structure (Figure 5D). Notably, compound PP56 varies from S1 by featuring a para-substituted phenol rather than a meta-substituted phenol (Figure 5B–D). This modification limits the binding pose possibilities, allowing compound PP56 to assume only one binding pose with respect to its phenol ring, unlike compound S1. We thus elected to proceed with

https://doi.org/10.1021/acs.jcim.4c00071
J. Chem. Inf. Model. 2024, 64, 3812–3825

compound PP56 as one of the reference compounds for Scaffold 1 in this study.

Next, we assessed the dual binding dynamics of the reference compound, PP494, which represented the type 1 1/2 inhibitor receptor complex (PDBid 3EN5, Figure 4B). Interestingly, we observed no substantial difference in the ΔG values when the dimethoxy group was oriented distal or proximal to the DFG moiety. Specifically, PP494_a and PP494_b exhibited ΔG values of -6.5 ± 0.09 kcal/mol and -6.4 ± 0.10 kcal/mol, respectively. These findings suggest that the specific orientation of the dimethoxy group did not significantly affect the overall ΔG of the compound. In this case, we decided to use the crystal structure binding pose of compound PP494, representing the type 1 1/2 kinase inhibitors class complex (PDBid 3EN5) as the reference structure for conducting FEP calculations. Finally, the resolved protein−ligand complexes for ligands from scaffolds 2 and 3 exhibited clear and unambiguous binding modes, thus eliminating the need for any ABFE calculations. Here we used the resolved binding poses of compounds PP121 and PP102 (PDBid's 3EN4 and 3EN6, respectively) as reference structures.

In the end, a total of six protein-reference structure complexes were generated for use in docking and subsequent FEP calculations. These consisted of three protein−ligand complexes between the 3EN4 type 1 inhibitor protein structure and the ligands PP56, PP121, and PP102. The other three protein-reference structure complexes consisted of the 3EN5 type 1 1/2 inhibitor protein conformation and ligands PP494, PP121, and PP102. PP56 and PP494 both belong to Scaffold 1 and represent the two inhibitor types 1 and 1 1/2, respectively. PP121 and PP102 served as reference structures of Scaffolds 2 and 3, respectively.

As the last step before initiating FEP calculations, a relaxation and evaluation process was carried out on all six protein-reference ligand complexes used in this study (details are provided in the Supporting Information). In general, FEP calculations require that target compounds are well aligned with the reference structure ensuring a consistent and accurate comparison of ligand energetics between different states or binding modes. To align the target compounds with the reference compound, we used constrained docking. The use of constrained docking allowed us to position the target compounds on the reference compound by enforcing alignment of only the MCS while allowing flexibility in the remaining parts of the molecules. The Methods section provides an in-depth explanation of our alignment and positioning techniques.

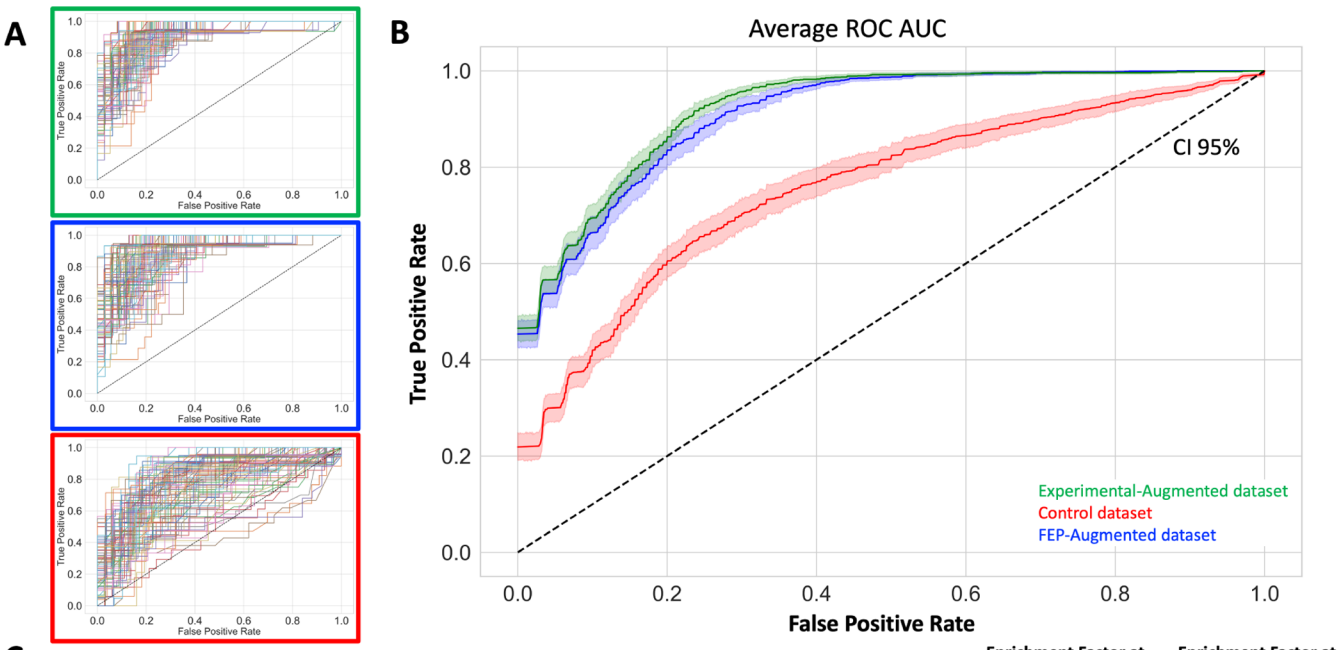
Finally, FEP calculations were carried out according to the methods described in the Methods section, with one exception. If upon analysis the RBFE calculation exhibited inadequate energy convergence, ligand RMSD, or REST exchange, further steps were taken. In such cases, the simulation time was extended until the inadequacies were resolved, first to 10 ns, and if the poor quality persisted, then the simulation time was increased to 20 ns. The Supporting Information includes all $\Delta\Delta G$'s calculated and associated data tables provided by FEP+ (FEP Results). All FEP calculations were conducted using a star topology map due to a significant increase in computational calculation cost when using an optimal topology map employing hysteresis.

FEP Analysis and Processing of Dual Binding Compounds. The presence of dual binding modes in compounds within Scaffold 1 necessitated the conversion of the $\Delta\Delta G$ values associated with these two energy states to a single energy value. To achieve this, we employed the Boltzmann distribution function to calculate the contribution of each binding mode to the overall $\Delta\Delta G$. Alternative methods have also been explored, such as the arithmetic average of the $\Delta\Delta G$ of multiple binding poses/states, with one such approach demonstrated in a study focusing on two binding modes/states of inhibitors binding to c-Jun N-terminal kinase-1.[37]

As outlined in our earlier work,[21] we employ the Cheng−Prusoff, equation to convert the relative difference in Gibbs free energy of the compound into an IC_{50}. A substrate concentration (S) of 10 μM for adenosine triphosphate (ATP) was used as stipulated in the assay conditions.[34] We utilized a K_m value for cSrc that has been previously reported in the literature (75 μM).[38] For an in-depth explanation of how to relate $\Delta\Delta G$, ΔG, K_i, and IC_{50} values, refer to the 2019 publication by Kaiser and Burger for the principles of conversion.[21]

Our initial step in the analysis of the FEP calculation results was to determine the ΔG mean unsigned error (MUE) for the calculations performed using each protein structure (Figure 6). The MUE for the 3EN4 structure (type 1 inhibition class protein structure) was 1.837, 1.198, and 1.513 kcal/mol for each of Scaffolds 1−3, respectively (Figure 6A,D,E). The ΔG MUE for the calculations using the 3EN5 structure (type 1 1/2 inhibition class protein structure) was 2.087, 1.724, and 1.608 kcal/mol for Scaffolds 1 and 3, respectively (Figure 6B,D,E). These values were higher than the expected 1−1.2 kcal/mol MUE[22] and could in part be explained by the relatively low quality of protein structures available and potentially by the mixed inhibitor classes that exist in this series. For the latter case, we investigated the different types of possible mechanisms of action for each of the small molecule ligands. We considered it crucial to examine the impact of accurately categorizing the inhibitors into their respective types, as this would enable us to create a more refined augmented data set for subsequent ML applications. In the original research carried out by Apsel et al., the authors acknowledged the existence of two different inhibitor types within the series, but it was unknown exactly which compounds belonged to either the type 1 or type 1 1/2 inhibition categories.[34] Our clarification of the mechanism for each compound relies on an initial observation that a key hydrogen bond forms between E310 of the αE-helix and the type 1 inhibitors locking the helix in place. We hypothesize that if this hydrogen bond were unable to form due to a change in the inhibitor, the αE-helix repositions itself, resulting in a switch of the type 1 1/2 conformation as illustrated in Figure 5E. Indeed, this was observed for compound PP494, which is captured in its 3EN5 structure (representing the type 1 1/2 inhibitor state; Figure 4B). This compound differs from compound S1 (type 1 inhibitor) by having a methoxy substituent in place of the hydroxyl group. We speculate that this loss of a hydrogen bond would lead to the inability of similar compounds to lock the αE-helix in the type 1 inhibitor class and result in a type 1 1/2 inhibitor when binding to cSrc. We used the above insight into classify Scaffold 1 into the two inhibitor classes based on their molecular structure [see the Supporting Information for which compounds were classified into type 1 and type 1 1/2 inhibitors (Compound List)]. This classification is by no means comprehensive but would reflect a real-life scenario if we had a limited initial SAR. After dividing Scaffold 1 into the two distinct inhibitor categories, we reevaluate the ΔG MUE.

https://doi.org/10.1021/acs.jcim.4c00071
J. Chem. Inf. Model. 2024, 64, 3812−3825



C

Decision Value		ROC AUC	Sensitivity	Precision	Enrichment Factor at 10%	Enrichment Factor at 20%
Dataset	DV	Mean ± Std Dev	Mean ± Std Dev	Mean ± Std Dev	Mean ± Std Dev	Mean ± Std Dev
Experimental-Augmented	50	0.921 ± 0.034	0.656 ± 0.137	0.764 ± 0.116	3.18 ± 0.48	2.73 ± 0.43
Control	50	0.794 ± 0.100	0.307 ± 0.251	0.768 ± 0.203	2.57 ± 0.78	2.26 ± 0.46
FEP-Augmented	50	0.910 ± 0.045	0.737 ± 0.144	0.709 ± 0.097	3.14 ± 0.49	2.64 ± 0.42

Figure 7. (A) ROC curves corresponding to all 100 iterations across the three distinct experimental scenarios using a DV of 50 nM. The curve representing the Experimental data set is color-coded in green, the Control data set is depicted in red, and the FEP-Augmented data set is illustrated in blue. It is important to note that the test set used to gauge the ML algorithms' performance consisted of 50 compounds not previously seen by the algorithms (Figure 2; subset E). (B) Mean ROC curve, aggregated from all 100 iterations, graphed for each experimental condition and is accompanied by a 95% confidence interval (shaded areas). (C) Table comparing the average statistical performance of the three ML experiments, each evaluated over 100 iterations and at a DV of 50 nM.

In Figure 6, we color-coded the following subdivisions: Scaffold 1 type 1 inhibitors are highlighted in bright blue, Scaffold 1 type 1 1/2 inhibitors are marked in black, Scaffold 2 compounds are green, and Scaffold 3 are red. Compounds from Scaffolds 2 and 3 were treated as if they belonged to the type 1 inhibitor class due to the insufficient resolution of the crystal structures. However, we did calculate $\Delta\Delta G$ for Scaffolds 2 and 3 in both the 3EN4 and 3EN5 structures in an effort to probe the error profile in each context (Figure 6).

As can be seen in Figure 6 E, the error distribution of the unsigned error (UE) for Scaffold 1 using the 3EN4 scaffold is much higher than the UE distributions for Scaffolds 2 and 3. When we split Scaffold 1 into putative type 1 inhibitors (light blue dots in Figure 6A−C and light blue bar in Figure 6D) and type 1 1/2 (black dots in Figure 6A−C and black bar in Figure 6D) based on each ligand's molecular structure, we could see that the type 1 1/2 compounds had very high error which was significantly greater than the typical 1−1.2 kcal/mol usually observed, with a ΔG MUE of 3.723 kcal/mol (n = 23). Of note is that the ΔG MUE for the type 1 inhibitors interacting with the 3EN4 type 1 protein structure dropped to 1.138 kcal/mol (n = 69) from 1.873 kcal/mol (n = 82) when the type 1 1/2 inhibitors were excluded. Likewise, when focusing on the type 1 1/2 inhibitors interacting with the 3EN5 protein structure, we observed that the ΔG MUE dropped to 1.409 kcal/mol (n = 23) from 2.087 kcal/mol (n = 82) when the type 1 inhibitors were removed (as shown in Figure 6B, represented by black dots, n = 23). In this case, the type 1 inhibitors MUE increase to 2.299 kcal/mol (n = 69) from 2.087 kcal/mol (Figure 6B, blue dots n = 82). This led us to the conclusion that we needed to use the 3EN5 structure for FEP calculations on the putative type 1 1/2 compounds and the 3EN4 Structure for the type 1 compounds given the change in protein conformational dynamics as a function of the inhibition mechanism.

Ultimately, we assembled the mixed inhibitor FEP-augmented data set by merging the $\Delta\Delta G$ data collected for type 1 inhibitors from the 3EN4 structure and type 1 1/2 inhibitors from the 3EN5 structure, along with data for Scaffolds 2 and 3, also derived from the 3EN4 structure (Figure 6C). The ΔG MUE for each of the scaffolds and inhibitor types are shown in Figure 6D. The final ΔG MUE for Scaffold 1 was 1.224 (n = 82), and 1.198 (n = 24) and 1.513 kcal/mol (n = 25) for Scaffolds 2 and 3. The data represented in Figure 6C represent the FEP-derived data set used in ML. The final data set comprises a total of 134 compounds, with 84 compounds belonging to Scaffold 1, 24 compounds belonging to Scaffold 2, and 25 compounds belonging to Scaffold 3 (Supporting Information; Compound List provided as a CSV file).

Evaluation of the Effect of FEP-Augmented Data Set on Machine Learning. Upon completion of the FEP calculations, we obtained the needed data set to be used as input for ML. We selected a classification approach to our ML methodology due to classification having an increased tolerance for activity noise in a data set.[39,40] We chose a threshold that would identify potent compounds as required by a lead optimization campaign (a typical campaign seeks to identify compounds with an IC_{50} < 100 nM for an initial

https://doi.org/10.1021/acs.jcim.4c00071
J. Chem. Inf. Model. 2024, 64, 3812−3825

optimization program) while ensuring a sufficient amount of data points remained in the test set for a robust evaluation of the ML algorithms generated in this approach. The classification threshold for defining an active compound was set at 50 nM, effectively dividing the data set into 30% active and 70% inactive. While the primary focus of this paper is on the 50 nM decision threshold (Figure 7), additional data using different active compound classification thresholds corresponding to 31 nM (Q1) and 239 nM (Q3) cutoffs are provided in the Supporting Information, Table S1.

ML Evaluation. With the FEP data set in hand, we could evaluate the improvement in prediction when we added either the FEP-calculated IC_{50} values or the experimentally derived IC_{50} values to the 11 "initial" compounds. Given our previous work, we selected a random forest as the ML approach, and algorithms were generated following the ML workflow outlined in Figure 2. A systematic assessment of the ML performance for each of the three ML experiments is shown in Figure 7. Figure 7A depicts the individual ROC curves resulting from each of the 100 runs conducted at a decision of 50 nM for the three augmentation experiments, whereas Figure 7B depicts the average ROC curves. The average behavior of the random forest generated on the Control data set (red), which was trained on only 11 selected compounds, yielded a ROC AUC value of 0.794 ± 0.100. The ROC AUC was found to be 0.921 ± 0.034 for the augmentation of the training set with experimental IC_{50} values (11 + 73 augmenting compounds; green). Gratifyingly the training set augmented with IC_{50} values as calculated by FEP (blue) afforded a ROC AUC of 0.910 ± 0.045. Interestingly, the ROC AUC distributions for the Experimental and FEP-Augmented data sets are considerably more compact than for the much sparser Control data set (Figure 7A), and we speculate that this reduced variance in the distribution of individual ROC curves for each experiment represents greater fidelity in generalization to the test set by both the Experimental IC_{50} augmented data set and the FEP-calculated IC_{50} augmented data set as compared to the unaugmented Control experiment. To visually assess the performance of our ML conditions, we plotted the average ROC curve with a 95% confidence interval in Figure 7B. As mentioned above, the mean ROC AUC value and its standard deviation were found to be very similar for the Experimental (green) and FEP-Augmented data set (blue). Moreover, the overall curve of the Experimental and FEP-Augmented data set follows a very similar trajectory indicating significant improvements in sensitivity via both means of augmentation. The findings are also mirrored in the average sensitivity and precision metrics computed for each algorithm; the mean sensitivity for the algorithm generated from training on the initial 11 compounds stands at a mere 0.307 compared to 0.656 and 0.737 for the algorithms augmented with Experimental and FEP-calculated IC_{50} values, respectively (as seen in Figure 7C). A notable discovery is that the Enrichment Factor at 10 and 20% shows no discernible difference between the algorithms developed using augmenting Experimental data and the ones employing FEP to generate the augmented data set.

The ML outcomes here underscore the remarkable similarity between the results derived from the FEP-Augmented data set and those from the original experimentally determined data set. These results suggest that data augmentation using FEP calculations can expedite the lead optimization process by saving substantial research time by synthesizing and testing a wide breadth of novel compounds once a small core SAR is generated with a small set of crystal structures.

CONCLUSIONS

In this paper, we establish the practical groundwork governing an accelerated lead optimization campaign employing physics-based methods to augment activity data sets for use in training ML algorithms. This augmentation affords a data set of sufficient information quality so that ML can quickly query millions of lead-like compounds and identify promising leads for drug development. Without this approach, the only other means of generating the requisite data is to resort to the labor- and time-intensive traditional make-test-analyze design cycles of contemporary medicinal chemistry. Our present guidelines for the necessary FEP calculations for data set augmentation demonstrate that an initial series of 10−20 related compounds, accompanied by 3D structures co-resolved a small set of ligands, can serve as a foundation for rapidly expanding the understanding of the activity landscape for a medicinal chemistry series at a very early stage using only physics-based methods. Additionally, we emphasized the significance of determining the relevant lowest-energy protein−ligand complex when multiple ligand binding poses or states are theoretically possible, as well as how to sample that energy landscape. Moreover, we highlighted the importance of understanding target conformational dynamics, which are essential for highly accurate calculations, and show that these dynamics can be revealed early in an optimization design cycle. Finally, we illustrated that using an FEP-augmented activity data set yielded predictive ML algorithms with similar predictive power as compared to those generated from experimentally derived activity data sets. Ultimately, we foresee the widespread adoption of augmenting activity data sets using physics-based methods to accelerate the acquisition of ML algorithms, which will enhance the lead optimization phase of drug discovery especially around engineering potency and selectivity against a potential biological target.

ASSOCIATED CONTENT

Data Availability Statement

All software generated for this paper is available in the Supporting Information. The KNIME analytics platform can be downloaded for free at https://www.knime.com/. All KNIME workflows are provided within the Supporting Information. All necessary data to replicate the study can be found in the public domain or within the provided Supporting Information.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c00071.

Comprehensive description of the methodologies and parameters employed; list of the chemicals involved in this research; outcomes for each FEP calculation; MD reports; workflow of the ML experiments, including the corresponding initial data; and ML performance at two additional categorical cutoff values (PDF)

MD reports (ZIP)

Input structure data (ZIP)

FEPML workflows (ZIP)

FEPML results (ZIP)

Compound list (ZIP)

SMILES (CSV)

https://doi.org/10.1021/acs.jcim.4c00071
J. Chem. Inf. Model. 2024, 64, 3812−3825

Processing Data Workflow (ZIP)

AUTHOR INFORMATION

Corresponding Authors

Pieter B. Burger − *Avicenna Biosciences Inc, Durham, North Carolina 27001, United States;* orcid.org/0000-0003-0272-1111; Email: pburger@avicenna-bio.com

Thomas M. Kaiser − *Avicenna Biosciences Inc, Durham, North Carolina 27001, United States;* orcid.org/0000-0001-5174-9183; Email: tkaiser@avicenna-bio.com

Authors

Xiaohu Hu − *Schrödinger, Inc., New York, New York 10036, United States*

Ilya Balabin − *Avicenna Biosciences Inc, Durham, North Carolina 27001, United States*

Morné Muller − *Avicenna Biosciences Inc, Durham, North Carolina 27001, United States*

Megan Stanley − *Microsoft Research AI4Science, Cambridge CB1 2FB, U.K.*

Fourie Joubert − *Centre for Bioinformatics and Computational Biology, Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0001, South Africa*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.4c00071

Author Contributions

The manuscript was written through the contributions of all authors. All authors have approved the final version of the manuscript.

Funding

This study was funded by Avicenna Biosciences, Inc.

Notes

The authors declare no competing financial interest.

ABBREVIATIONS

ABFE, absolute binding free energy; DMPK/PD, drug metabolism and pharmacokinetics/pharmacodynamics; DV, decision value; FEP, free energy perturbation; MCS, maximum common substructure; MD, molecular dynamics; ML, machine learning; MUE, mean unsigned error; PDBid, Protein Data Bank identification number; QSAR, quantitative structure−activity relationship; RBFE, relative binding free energy; RF, random forest; ROC AUC, receiver operating characteristic area under the curve; SAR, structure−activity relationship; SC1, scaffold 1; SC2, scaffold 2; SC3, scaffold 3; UE, unsigned error

REFERENCES

(1) Ehrlich, P. Address in Pathology, on Chemotherapy. *Br. Med. J.* **1913**, *2*, 353−359.

(2) Feng, K.; Quevedo, R. E.; Kohrt, J. T.; Oderinde, M. S.; Reilly, U.; White, M. C. Late-stage oxidative C(sp3)−H methylation. *Nature* **2020**, *580*, 621−627.

(3) Barreiro, E. J.; Kümmerle, A. E.; Fraga, C. A. M. The Methylation Effect in Medicinal Chemistry. *Chem. Rev.* **2011**, *111* (9), 5215−5246.

(4) Hargrave-Thomas, E.; Yu, B.; Reynisson, J. Serendipity in anticancer drug discovery. *World J. Clin. Oncol.* **2012**, *3*, 1−6.

(5) Ban, T. A. The role of serendipity in drug discovery. *Dialogues Clin. Neurosci.* **2006**, *8*, 335−344.

(6) Sams-Dodd, F. Target-based drug discovery: is something wrong? *Drug Discovery Today* **2005**, *10*, 139−147.

(7) Brown, D. Unfinished business: target-based drug discovery. *Drug Discovery Today* **2007**, *12*, 1007−1012.

(8) Ahn, A. C.; Tewari, M.; Poon, C.-S.; Phillips, R. S. The Limits of Reductionism in Medicine: Could Systems Biology Offer an Alternative? *PLoS Med.* **2006**, *3*, No. e208.

(9) Scannell, J. W.; Bosley, J. When Quality Beats Quantity: Decision Theory, Drug Discovery, and the Reproducibility Crisis. *PLoS One* **2016**, *11*, No. e0147215.

(10) Cantrill, C.; Chaturvedi, P.; Rynn, C.; Schaffland, J. P.; Walter, I.; Wittwer, M. B. Fundamental aspects of DMPK optimization of targeted protein degraders. *Drug Discovery Today* **2020**, *25*, 969−982.

(11) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977−5010.

(12) Keith, J. A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Müller, K.-R.; Tkatchenko, A. Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. *Chem. Rev.* **2021**, *121*, 9816−9872.

(13) Tsou, L. K.; Yeh, S.-H.; Ueng, S.-H.; Chang, C.-P.; Song, J.-S.; Wu, M.-H.; Chang, H.-F.; Chen, S.-R.; Shih, C.; Chen, C.-T.; Ke, Y.-Y. Comparative study between deep learning and QSAR classifications for TNBC inhibitors and novel GPCR agonist discovery. *Sci. Rep.* **2020**, *10*, No. 16771, DOI: 10.1038/s41598-020-73681-1.

(14) Ekins, S.; Puhl, A. C.; Zorn, K. M.; Lane, T. R.; Russo, D. P.; Klein, J. J.; Hickey, A. J.; Clark, A. M. Exploiting machine learning for end-to-end drug discovery and development. *Nat. Mater.* **2019**, *18*, 435−441.

(15) Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today* **2018**, *23*, 1538−1546.

(16) Struble, T. J.; Alvarez, J. C.; Brown, S. P.; Chytil, M.; Cisar, J.; DesJarlais, R. L.; Engkvist, O.; Frank, S. A.; Greve, D. R.; Griffin, D. J.; Hou, X.; Johannes, J. W.; Kreatsoulas, C.; Lahue, B.; Mathea, M.; Mogk, G.; Nicolaou, C. A.; Palmer, A. D.; Price, D. J.; Robinson, R. I.; Salentin, S.; Xing, L.; Xing, L.; Jaakkola, T.; Green, W. H.; Barzilay, R.; Coley, C. W.; Jensen, K. F. Current and Future Roles of Artificial Intelligence in Medicinal Chemistry Synthesis. *J. Med. Chem.* **2020**, *63*, 8667−8682.

(17) Ching, T.; Himmelstein, D. S.; Beaulieu-Jones, B. K.; Kalinin, A. A.; Do, B. T.; Way, G. P.; Ferrero, E.; Agapow, P.-M.; Zietz, M.; Hoffman, M. M.; Xie, W.; Rosen, G. L.; Lengerich, B. J.; Israeli, J.; Lanchantin, J.; Woloszynek, S.; Carpenter, A. E.; Shrikumar, A.; Xu, J.; Cofer, E. M.; Lavender, C. A.; Turaga, S. C.; Alexandari, A. M.; Lu, Z.; Harris, D. J.; DeCaprio, D.; Qi, Y.; Kundaje, A.; Peng, Y.; Wiley, L. K.; Segler, M. H. S.; Boca, S. M.; Swamidass, S. J.; Huang, A.; Gitter, A.; Greene, C. S. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **2018**, *15*, No. 20170387.

(18) Wang, Y.; Yao, Q.; Kwok, J. T.; Ni, L. M. Generalizing from a Few Examples: A Survey on Few-shot Learning. *ACM Comput. Surv.* **2021**, *53*, 1−34.

(19) Weiss, K.; Khoshgoftaar, T. M.; Wang, D. A survey of transfer learning. *J. Big Data* **2016**, *3*, 1−40, DOI: 10.1186/s40537-016-0043-6.

(20) Choudhury, S.; Moret, M.; Salvy, P.; Weilandt, D.; Hatzimanikatis, V.; Miskovic, L. Reconstructing Kinetic Models for Dynamical Studies of Metabolism using Generative Adversarial Networks. *Nat. Mach. Intell.* **2022**, *4*, 710−719, DOI: 10.1038/s42256-022-00519-y.

(21) Kaiser, T. M.; Burger, P. B. Error Tolerance of Machine Learning Algorithms across Contemporary Biological Targets. *Molecules* **2019**, *24*, No. 2115, DOI: 10.3390/molecules24112115.

(22) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm,

W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015**, *137*, 2695−2703.
(23) Pérez-Benito, L.; Keränen, H.; Vlijmen, H.; Tresadern, G. Predicting Binding Free Energies of PDE2 Inhibitors. The Difficulties of Protein Conformation. *Sci. Rep.* **2018**, *8*, No. 4883.
(24) Williams-Noonan, B. J.; Yuriev, E.; Chalmers, D. K. Free Energy Methods in Drug Design: Prospects of "Alchemical Perturbation" in Medicinal Chemistry. *J. Med. Chem.* **2018**, *61*, 638−649.
(25) Hansen, N.; Gunsteren, W. F. v. Practical Aspects of Free-Energy Calculations: A Review. *J. Chem. Theory Comput.* **2014**, *10*, 2631−2647.
(26) Cournia, Z.; Allen, B.; Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model.* **2017**, *57*, 2911−2937.
(27) Abel, R.; Wang, L.; Harder, E. D.; Berne, B. J.; Friesner, R. A. Advancing Drug Discovery through Enhanced Free Energy Calculations. *Acc. Chem. Res.* **2017**, *50*, 1625−1632.
(28) Moraca, F.; Negri, A.; Oliveira, Cd.; Abel, R. Application of Free Energy Perturbation (FEP+) to Understanding Ligand Selectivity: A Case Study to Assess Selectivity Between Pairs of Phosphodiesterases (PDE's). *J. Chem. Inf. Model.* **2019**, *59*, 2729−2740.
(29) Gagic, Z.; Ruzic, D.; Djokovic, N.; Djikic, T.; Nikolic, K. In silico Methods for Design of Kinase Inhibitors as Anticancer Drugs. *Front. Chem.* **2020**, *7*, 873.
(30) Huggins, D. J.; Sherman, W.; Tidor, B. Rational Approaches to Improving Selectivity in Drug Design. *J. Med. Chem.* **2012**, *55*, 1424−1444.
(31) Thompson, J.; Walters, W. P.; Feng, J. A.; Pabon, N. A.; Xu, H.; Maser, M.; Goldman, B. B.; Moustakas, D.; Schmidt, M.; York, F. Optimizing active learning for free energy calculations. *Artif. Intell. Life Sci.* **2022**, *2*, No. 100050.
(32) Konze, K. D.; Bos, P. H.; Dahlgren, M. K.; Leswing, K.; Tubert-Brohman, I.; Bortolato, A.; Robbason, B.; Abel, R.; Bhat, S. Reaction-Based Enumeration, Active Learning, and Free Energy Calculations To Rapidly Explore Synthetically Tractable Chemical Space and Optimize Potency of Cyclin-Dependent Kinase 2 Inhibitors. *J. Chem. Inf. Model.* **2019**, *59* (9), 3782−3793.
(33) Khalak, Y.; Tresadern, G.; Hahn, D. F.; de Groot, B. L.; Gapsys, V. Chemical Space Exploration with Active Learning and Alchemical Free Energies. *J. Chem. Theory Comput* **2022**, *18* (10), 6259−6270.
(34) Apsel, B.; Blair, J. A.; Gonzalez, B.; Nazif, T. M.; Feldman, M. E.; Aizenstein, B.; Hoffman, R.; Williams, R. L.; Shokat, K. M.; Knight, Z. A. Targeted polypharmacology: discovery of dual inhibitors of tyrosine and phosphoinositide kinases. *Nat. Chem. Biol.* **2008**, *4* (11), 691−699.
(35) Belli, S.; Esposito, D.; Servetto, A.; Pesapane, A.; Formisano, L.; Bianco, R. c-Src and EGFR Inhibition in Molecular Cancer Therapy: What Else Can. We Improve? *Cancers* **2020**, *12* (6), No. 1489.
(36) de Oliveira, C.; Yu, H. S.; Chen, W.; Abel, R.; Wang, L. Rigorous Free Energy Perturbation Approach to Estimating Relative Binding Affinities between Ligands with Multiple Protonation and Tautomeric States. *J. Chem. Theory Comput.* **2019**, *15* (1), 424−435.
(37) Kaus, J. W.; Harder, E.; Lin, T.; Abel, R.; McCammon, J. A.; Wang, L. How to deal with multiple binding poses in alchemical relative protein-ligand binding free energy calculations. *J. Chem. Theory Comput.* **2015**, *11* (6), 2670−2679.
(38) Kemble, D. J.; Wang, Y. H.; Sun, G. Bacterial expression and characterization of catalytic loop mutants of SRC protein tyrosine kinase. *Biochemistry* **2006**, *45* (49), 14749−14754.
(39) Tsimring, L. S. Noise in biology. *Rep. Prog. Phys.* **2014**, *77* (2), No. 026601.
(40) Eling, N.; Morgan, M. D.; Marioni, J. C. Challenges in measuring and understanding biological noise. *Nat. Rev. Genet.* **2019**, *20* (9), 536−548.

https://doi.org/10.1021/acs.jcim.4c00071
J. Chem. Inf. Model. 2024, 64, 3812−3825