

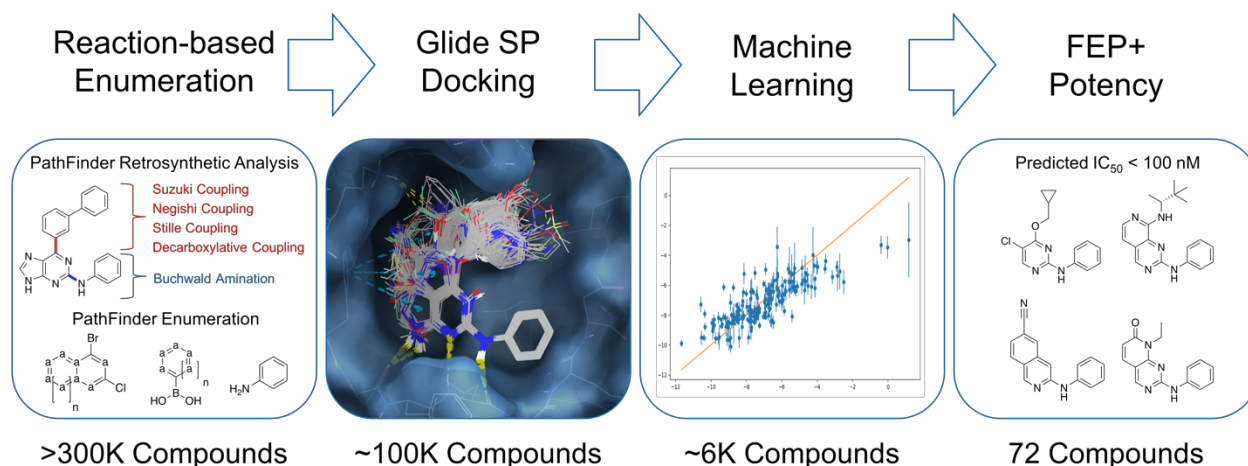
Reaction-based Enumeration, Active Learning, and Free Energy Calculations to Rapidly Explore Synthetically Tractable Chemical Space and Optimize Potency of Cyclin Dependent Kinase 2 Inhibitors

Kyle D. Konze[‡], Pieter H. Bos[‡], Markus K. Dahlgren, Karl Leswing, Ivan Tubert-Brohman, Andrea Bortolato, Braxton Robbason, Robert Abel, Sathesh Bhat^{*}

[‡]These authors contributed equally to this work

^{*} sathesh.bhat@schrodinger.com

Schrödinger Inc, 120 West 45th St, 17th floor, New York, New York, 10036



Abstract

The hit-to-lead and lead optimization processes usually involve the design, synthesis, and profiling of thousands of analogs prior to clinical candidate nomination. A drug discovery campaign may begin with a virtual screening hit-finding campaign that explores millions of compounds; if not more. Interestingly, this scale of computational profiling is not frequently performed in the hit-to-lead or lead optimization phases of drug discovery. This is likely due to the lack of appropriate computational tools to generate synthetically tractable compounds within a given lead-like space, a lack of computational methods to accurately profile compounds prospectively on a large scale, and the ability of these approaches to deliver results within a suitable time frame to impact real-world project decisions. Increases in computational power and the development of new computational

methods provide the ability to profile much larger libraries of ligands, which in turn has created a clear need for technologies capable of rapidly generating relevant design ideas for drug discovery projects. We here report a new computational technique, referred to herein as ‘PathFinder’, that uses retrosynthetic analysis followed by combinatorial synthesis to generate novel compounds in synthetically accessible chemical space. Coupling the *in silico* generation of lead matter with active learning and cloud-based free energy calculations allows for large-scale potency predictions of compound libraries on a timescale that can impact real-world drug discovery projects. The process is further accelerated by using a combination of population-based statistics and active learning techniques. Here, we present a workflow that integrates PathFinder enumeration, cloud-based FEP simulations, and active learning to rapidly optimize R-groups and generate new cores for inhibitors of cyclin-dependent kinase 2 (CDK2). Using this approach, we explored greater than 300 thousand ideas and identified 35 ligands with diverse commercially available R-groups and a predicted $IC_{50} < 100$ nM, and four unique cores with a predicted $IC_{50} < 100$ nM. The rapid turnaround time, and scale of chemical exploration, suggests that this is a useful approach to accelerate the discovery of novel chemical matter in drug discovery campaigns.

Introduction

Computational chemistry has had a profound impact on early-stage drug development. For example, the utilization of virtual screening as a hit-finding strategy in early drug discovery campaigns is now very common.¹⁻⁴ Virtual screening is a rapid, inexpensive method that allows for the *in silico* evaluation of large libraries of small molecules from both commercial vendors and/or in-house compound collections. These methods often also provide an initial hypothesis of the binding mode of a compound, and after experimental validation in the relevant biological assays, the most promising ‘hits’ can be used as starting points for lead generation.

The size of *in silico* libraries for virtual screening has steadily increased over time⁴, but the same paradigm is surprisingly underutilized in the lead optimization stage. This is likely due to the lack of appropriate computational tools to create synthetically tractable compounds within a given lead-like space^{5,6}, and a lack of computational methods to accurately profile compounds to optimize multiple parameters, prospectively. The process of optimizing lead compounds to improve their potency, selectivity, and ADMET properties is often challenging, expensive, and time-consuming, and on average accounts for more than half, or \$414 million, of the total capitalized costs in the preclinical

phase.⁷ Therefore, new tools to rapidly ideate synthetically tractable ligands based on the structure of a lead compound coupled with accurate ADMET, potency and selectivity modeling could greatly accelerate the hit-to-lead and lead-optimization phases of drug discovery.

Traditionally, the lead-optimization phase relies heavily on empirical knowledge derived from structure-activity relationship (SAR) studies, in which close analogs of the lead compound are designed, synthesized, and evaluated for potency against the target. Generally, several of these cycles are required to obtain compounds that have the desired potency, and this approach comes with a number of drawbacks and limitations. First, the iterative nature of this process, in which generally only one property is optimized at a time, can lead to suboptimal compounds (*e.g.* excellent potency, but poor physicochemical properties), and ‘activity cliffs’ (small structural changes that have a dramatic effect on potency) can be missed due to limited exploration. Second, the combination of time and resource constraints in traditional SAR-studies curbs the exploration of the total available chemical space, which is estimated to be between 10^{20} - 10^{24} molecules⁸, and if combinations of known fragments are considered, up to as many as 10^{60} molecules^{9,10} that satisfy at least two of the four criteria defined in Lipinski’s ‘rule of five’.¹¹ Both of these limitations can benefit greatly from the inclusion of computational screening techniques in the lead optimization phase of drug development projects.

Recently, the use of computational approaches in the *de novo* design of ligands has gained considerable momentum in both academia and the pharmaceutical industry^{5,12-14}, and a number of strategies have been published. Examples of strategies that have been used are: (1) to start from a database of available molecules and simulate organic synthesis steps¹⁵; (2) *De novo* design using reaction vectors¹⁶; (3) Rule based molecular fragmentation in combination with a pharmacophore fingerprint-based fragment replacement algorithm¹⁷; (4) *De novo* design using Reaction-MQL encoded reactions¹⁸; and, recently, (5) The emergence of machine learning approaches in the *de novo* design process.¹⁹⁻²⁴

Additionally, there are a number of tools available that aim to aid in synthetic route design. Computer aided synthesis design (CASD) was pioneered by E. J. Corey in the 1960’s²⁵⁻²⁷, and a plethora of tools exist that use a variety of strategies ranging from network algorithms searching the wealth of literature reported reactions²⁸⁻³⁰, to automated retrosynthetic rule generation based on the recognition of common patterns³¹, curated sets of expert-coded transformations^{28,32}, and neural-symbolic machine learning approaches to retrosynthesis and reaction planning and prediction.^{33,34}

In this manuscript we will discuss the development and application of PathFinder, a reaction-based enumeration tool that enables the rapid exploration of synthetically tractable ligands. The combination of retrosynthetic analysis, reaction-based enumeration, and robust filtering in an easy to

use graphical user interface (GUI) is what differentiates PathFinder from other available tools. PathFinder, coupled with multi-parameter optimization (MPO), docking, machine learning, and cloud-based FEP simulations, provides a streamlined approach for rapidly creating and evaluating large sets of synthetically tractable, lead-like, potent ligands that are of significant interest in drug discovery campaigns.

Methods

The process of creating a large library of synthetically tractable ligands *in silico* begins with the conversion of the chemical reactions, as depicted in textbooks, to their computational counterparts. The use of SMiles ARbitrary Target Specification (SMARTS) is a common and efficient method for the translation and interpretation of chemical structures into interpretable patterns.³⁵ Reaction SMARTS are robust, allowing the user to encode chemical, and alchemical, transformations in a selective manner (Fig 1). The use of reaction SMARTS to identify bonds in a structure that can be broken and recreated using known chemical reactions is an integral part of PathFinder. In recent years, a number of reactions have been encoded into their SMARTS counterparts and made available to the scientific community.³² We sought to improve these patterns and expand the set by adding the most common reactions in the medicinal chemist's toolkit.³⁶⁻³⁸ Currently, PathFinder incorporates > 100 reactions ranging from carbon-carbon bond forming reactions (*e.g.* Suzuki coupling, Sonogashira coupling) to alkylation reactions, ether formation, and amide coupling reactions (See Supporting Information; Table S4). Due to the importance of chemical rings in molecular scaffolds and drug discovery^{38,39}, a large number of the reactions contained within PathFinder focus on the formation of heterocyclic systems (*e.g.* imidazoles, triazoles, oxazoles, indoles, pyridines, etc.).

PathFinder is implemented in Python using the RDKit cheminformatics toolkit.⁴⁰ We chose RDKit because of its fast and robust implementation of reaction SMARTS and its support for computing a large number of molecular descriptors which can be used for filtering. The retrosynthetic analysis is done in two steps. First a retrosynthetic tree is created by applying all possible retrosynthetic transforms to the target recursively down to a user-specified depth. Retrosynthetic trees are directed rooted trees with two kinds of nodes: molecules and reactions. Each molecule can be connected to any number of reactions, which are the reactions that can make the molecule in one step; each reaction is connected to one or more molecules, which are the precursors of the target of the reaction (Figure 1a).

Next, the retrosynthetic tree is converted into multiple route trees, by traversing all possible paths from the target to the starting materials. A route tree is similar to a retrosynthesis tree, with the additional restriction that each molecule can be connected to one reaction at most: the specific reaction used to make the molecule in that route. Examples of both trees are shown in Figure 1.

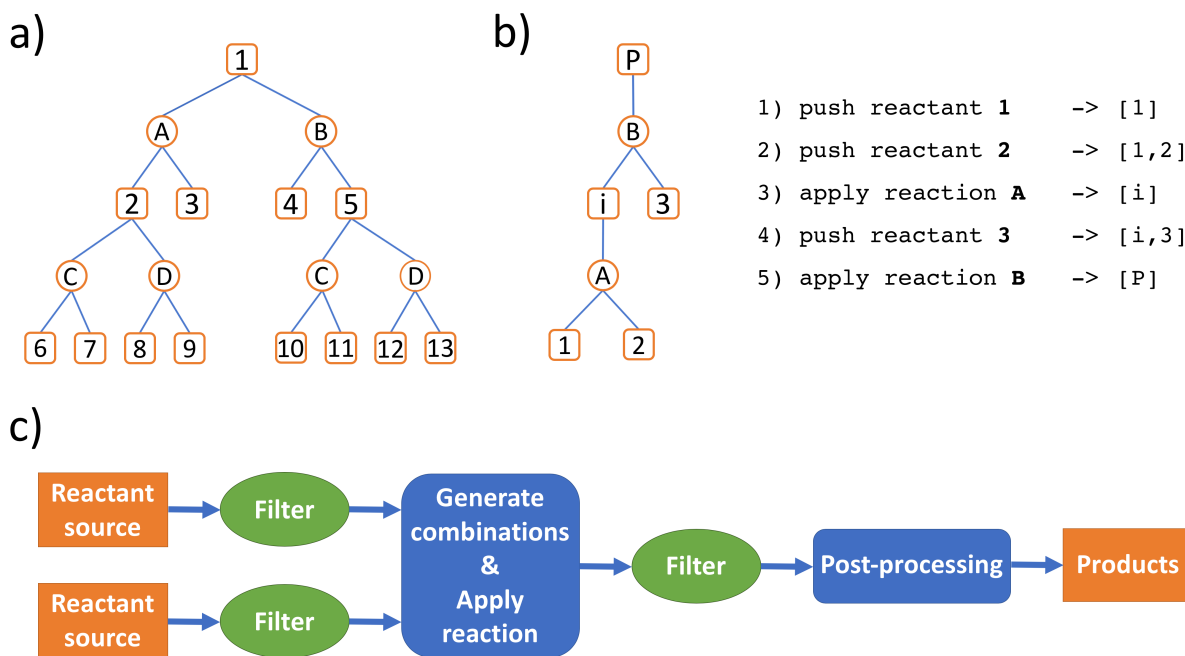


Figure 1. a) Retrosynthesis tree of depth 2. Squares represent molecules; circles represent reactions. The target molecule is at the top (1). From this retrosynthesis tree four two-step route trees can be extracted. b) A route tree is converted into instructions for a stack-based machine. The values in brackets represent the contents of the stack after each instruction. Although 1, 2, and 3 each represent one structure, i and P may represent multiple structures if the reactions generate multiple products. c) Flowchart for a combinatorial synthesis workflow using two reactant sources.

To perform a virtual synthesis, the route is first converted into a program for a simple stack-based machine, as shown in Figure 1b. Each value on the stack represents one or more molecules, and the machine supports two instructions: *push* (a molecule) and *apply* (a reaction). The *apply* instruction pops one entry from the stack for each reactant needed by the reaction (e.g. two for a typical coupling reaction), applies the reaction, and then pushes the product(s) to the stack. Given that each stack entry may contain more than one molecule, the reaction may be attempted multiple times, using the Cartesian product of the reactant lists. When the program runs to completion, the entry at the top of the stack will hold the product(s) of the synthesis.

The virtual synthesis step is part of an enumeration workflow which supports filtering of both reactants and products, as shown in Figure 1c. In addition to the filtering, post-processing of the products includes removal of duplicates, optional generation of 3D coordinates, and computation of molecular descriptors.

The central step of Figure 1c, *Generate combinations*, constitutes the main loop of the combinatorial synthesis program and may operate in two modes: random or systematic. Systematic mode tries all possible combinations of reactants but is impractical when more than one reactant is being varied and/or the reactant libraries are large. Given that our largest reactant libraries contain approximately 10^6 compounds, an enumeration varying two starting materials can generate up to 10^{12} products. For this reason, the typical enumeration mode is random sampling: for each iteration of the main loop, a random reactant is picked from each of the reactant sources, and if the reaction can be applied successfully, the product is moved to the filtering and post-processing stages. This cycle is repeated until the desired number of products (or unsuccessful attempts) has been reached.

In addition to reactions that combine two reactants and generate a product ($A + B \rightarrow C$; *e.g.*, a traditional amide coupling reaction), a number of functional group transformations ($D \rightarrow E$; *e.g.*, Miyaura borylation) are also included to allow for additional chemical diversity of be considered. Functional group transformations have to be used judiciously, especially if this transformation results in a chemical feature that is present in a lot of molecules (*e.g.* the reduction of a ketone to form an alcohol), so these transformations are only considered when no more reactions can be applied to the products from the previous retrosynthetic step. For example, when performing an amide coupling, one would want to consider all commercially available carboxylic acids, but primary alcohols can be oxidized to carboxylic acids, so allowing one additional transformation (primary alcohol to carboxylic acid) opens up a completely different set of reactants to consider as R-groups.

Our general workflow illustrated by the design of novel potent CDK2 inhibitors uses the PathFinder tool, combined with a number of computational techniques (see Figure 2), and consists of six steps, that can be summarized as follows: Step (1): PathFinder analysis of the lead molecule provides a number of synthetic routes that can be used as starting points in the ligand enumeration step. Step (2): After selection of the desired route, two main enumeration strategies can be carried out: (a) R-group enumeration or (b) core hop enumeration. R-group enumeration most often involves a one-step route that will focus on exploration of a specific region of the compound, whereas a core hop enumeration will generally involve at least two steps in order to couple the original R-group(s) onto the new cores. Step (3): The enumerated library is pared down in an MPO-filtering step based

on project-specific physicochemical properties. Step (4): Fast computational docking techniques are employed to rapidly identify and eliminate molecules that are incompatible with the binding site. Step (5): We performed FEP simulations using FEP+ to accurately predict binding affinity.⁴¹ However, it is possible that at step 5 the enumerated library may still be too large (> 10K ligands) for computationally expensive methods such as FEP simulations. To avoid arbitrarily filtering the remaining ligands, we report an active learning approach using machine learning (ML) models trained on successive rounds of FEP+ predictions to triage the ideas. We will discuss the use of ML to enrich the dataset in this work. Step (6): Compounds predicted to be potent, and possess the desired properties, are suggested for synthesis and evaluation in the relevant assays. Based on the resulting structure-activity relationship data, the reaction-based enumeration workflow can be repeated to generate additional libraries if necessary, and the additional experimental data used to feed the model should improve model accuracy with each iteration.

The first step, PathFinder analysis of the lead compound, uses a library of highly curated reactions, encoded as SMARTS patterns, to break down molecules into their respective starting materials. This process is repeated on the products as many times as the user specifies (reaction depth) and/or until the building blocks cannot be broken down further. Currently, over 100 reactions are incorporated in PathFinder (See Supporting Information; Table S4). These reactions were selected based on the frequency of their occurrence in the medicinal chemistry literature, general applicability, and robustness. Each reaction encoded in the database contains a SMARTS pattern for the breakdown reaction (the retrosynthetic step) and a pattern for the buildup reaction (the enumeration step). Having separate SMARTS patterns is essential, because small differences in how molecules are formed synthetically, or broken down during a retrosynthesis, can lead to issues if the patterns for both the products and starting materials are simply inverted. Most of our SMARTS patterns include the use of recursive SMARTS which define the chemical environments that are compatible, or incompatible, with the corresponding reaction it encodes.

In the second step, the enumeration step, the desired synthetic route is selected, and the reaction-based enumeration tool generates the enumerated library using the SMARTS pattern(s) together with the appropriate reactant libraries. For the enumeration step, a number of concepts are of key importance: (1) curation of the building blocks/reactants; (2) automatic retrieval of the appropriate building blocks; and (3) flexibility to use custom reactant files and custom reactions. The standard set of reactant libraries included with PathFinder is curated from the eMolecules building block database,⁴² and the flexibility of PathFinder allows for the use of additional sets of reactant

libraries based on preferred vendor catalogues, or in-house reactant collections. The eMolecules library is created by filtering the entire set of building blocks with a collection of proprietary SMARTS patterns that filter out structures based on incorrect valence, drawing errors, and reactivity beyond that required in PathFinder. This set of filtered building blocks is then further filtered based on physicochemical properties, which can be customized depending on the desired outputs. For the purposes of this paper, the following criteria are used: MW < 350, RB < 10, HBA < 10, HBD < 6, PSA < 160. The resulting filtered building block set is then split into the various sets of reactants compatible with the reactions contained within PathFinder.

Currently, the enumeration tool contains 83 different reactant library sets, ranging in size from 10^1 - 10^6 reactants. These libraries are generated using a custom Python script and RDKit. The script searches the filtered eMolecules building block library using SMARTS pattern matching and generates a CSV file containing the SMILES string of the building block together with the eMolecules ID. To prevent selectivity issues caused by reactants matching the SMARTS pattern more than once, only reactants with a single match are allowed. Additional filtering rules based on SMARTS patterns and/or physicochemical properties such as logP, molecular weight, polar surface area, number of hydrogen bond donors/acceptors, heavy-atom count, and ring count can be automatically applied to both the reactants and the enumerated products by the user to further shape the output of the enumeration.

The enumeration step also has the added flexibility that any of the reactants in a multi-step enumeration sequence can be kept constant or varied. This is especially useful in both R-group and core hop enumerations (see Fig. 2 and Fig. 3) to produce ideas that are readily amenable to downstream computational models.

PathFinder combines curated reactant libraries, facile creation of additional reactant libraries from SMARTS patterns, filtering tools based on physicochemical properties and SMARTS patterns, with the ability to select which reactants to enumerate. This provides easy access to synthetically tractable synthetic libraries, that, at the same time, fit a range of user-specified criteria. At this stage it is also possible to filter based on metrics such as Tanimoto similarity, or filter out compounds based on known chemical matter. The filtering step is essential in reducing the size of the enumerated library so that it can be gradually advanced to more expensive and time-consuming computational assays, and at the same time it ensures that the resulting molecules are relevant to the project-specific aims and goals.

In the fourth step, a computational model that incorporates 3D information (docking, pharmacophore, etc.) is employed to quickly identify molecules that are likely to be incompatible with

the binding site. Using Glide SP^{43,44} one can quickly dock hundreds of thousands of ideas, and depending on what is known about the target, specific constraints can be employed to filter out ligands that are likely not to drive potency due to a lack of key interactions believed to be important for binding. To accelerate this step, the user could also simply align all ligands to a known reference ligand and then “score in place” to remove ligands that are creating clashes with the receptor.

Even after filtering with 3D techniques, enumerated libraries can be quite large (potentially 100K ligands or more), so a subset of the pool must be selected for profiling with more computationally expensive methods like FEP+. For the CDK2 example discussed in this work, we use an active learning^{45,46} approach using FEP+ predictions to quickly enrich for potent compounds. Initially, 1K ligands are chosen randomly and FEP simulations are performed on this set for 1 ns from a single reference molecule. The FEP+ results from the set of 1K random molecules are then used to train a machine learning model to predict the FEP+ potency of the remaining ligands. The top 1K from the machine learning model predictions are evaluated in FEP simulations. After each iteration the results were used to inform the model prior to the next round of predictions, and the process was repeated four times.

FEP Methods

FEP simulations were run using FEP+.^{41,47} All FEP simulations used 16 lambda windows, hydrogen mass repartitioning (enabling 4 fs timestep), SPC waters⁴⁸, and the OPLS3 force field.⁴⁹ Single edge FEP was run for 1 ns and full cycle closure FEP was run for 20 ns simulation length. Detailed FEP map information, including custom core SMARTS, is given in the supporting information.

Results and Discussion

A number of drugs approved by the FDA in 2017, covering a wide variety of chemical structures, targets, and modes of action were analyzed using PathFinder with a maximum depth of 2 (Table 1). In these examples the cores were kept constant, and the various R-groups were enumerated using applicable reactions. To prevent a combinatorial explosion of possible products generated in each route, which could contain up to three reactions, only one reactant was varied in each of the enumerations. This also significantly increases the possibility that the ligands can be profiled by FEP, because they are likely to be closer in structure to known reference compounds. The enumerated R-

group libraries were combined and de-duplicated (different reactants and reactions may form the same product), and the results are reported in Table 1. This table illustrates how PathFinder can be employed across a diverse set of drug-like chemical matter to quickly generate large libraries of synthetically tractable molecules that are readily amenable to downstream computational models

Table 1. PathFinder reaction-based enumeration of a set of drug molecules approved by the FDA in 2017 and Compound C73⁵⁰.^a

Input molecule	Class	# Routes	# Jobs	Median time (s) ^b	# Enumerated compounds
Letermovir	Antiviral	10	18	583.6	1,058,608
Abemaciclib	Dual CDK4/6 inhibitor	216	502	604.9	9,220,785
Enasidenib	IDH2 allosteric inhibitor	78	145	1548.2	4,077,689
Naldemedine	μ -, δ -, κ -opioid receptor antagonist	39	67	3335.8	1,390,474
Compound C73 ⁵⁰	CDK2 inhibitor	310	694	3382.6	9,828,340

^aPathFinder was run with a maximum depth of 2 and the core of each compound was kept constant (See Supporting Information; Table S5). ^bMedian time per job reported in seconds for an enumeration job run on 24 cores.

To demonstrate the utility of reaction-based enumeration followed by FEP profiling in a drug discovery campaign, we selected a recent example from the literature in which a cyclin-dependent kinase-2 (CDK2) inhibitor was developed (See Table 1; Compound C73).⁵⁰ We chose to use CDK2 as our model system to illustrate the PathFinder workflow for a number of reasons. First, kinases are relevant drug targets with many examples of compounds that have progressed from hit to clinic and ultimately FDA approval.^{51–55} Second, the literature provides chemical matter and SAR to serve as great starting points for the workflows discussed.⁵⁰ And third, CDK2 was used previously as a model system for retrospective validation of our FEP technologies, which demonstrated excellent agreement of the FEP predictions with experimental binding affinities⁵⁶, suggesting that we can have relatively high confidence in the predictions that are made in this exercise.

CDK2 Biphenyl R-group Replacement

A previous publication detailed the discovery of a potent, selective CDK2 inhibitor (Figure 2A, Compound **C73**).⁵⁰ Given the liabilities of a biphenyl containing compound for further development (solubility, protein binding, metabolic stability, etc.), and our previous success with FEP predictions in CDK2⁵⁶, we aimed to find a potent biphenyl replacement with PathFinder-driven enumerations followed by large-scale FEP simulations with FEP+. First, we simplified the input structure of **C73** by removing the terminal sulfonamide, because we did not want the relatively poor properties of the sulfonamide (PSA, MW, HBA, HBD, etc.) to result in the filtering of R-groups when shaping the library after enumeration. PathFinder analysis with a depth of 1 identified multiple aryl couplings as possible reactions to replace the biphenyl group within seconds (Figure 2a).

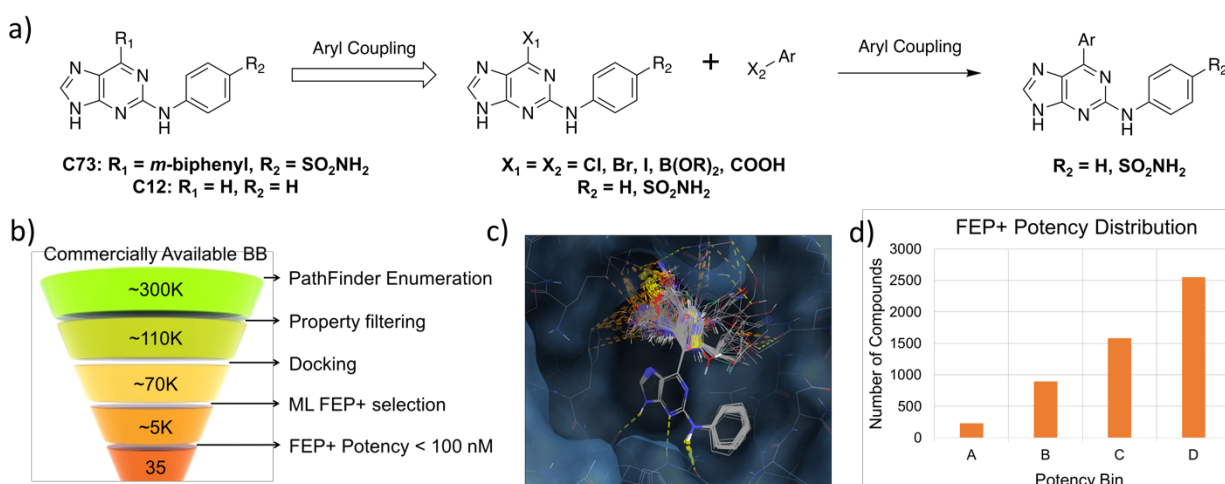


Figure 2. PathFinder R-Group Enumeration Workflow. **a)** Retrosynthetic scheme demonstrating aryl coupling opportunity and subsequent enumeration of aryl building blocks (X_2 -Ar) onto the purine core. **b)** Funnel depicting the filtering process of the enumerated library: drug-like property filtering, docking, and then ML FEP+ selection. **c)** Input poses of 5,255 ligands for FEP simulations. **d)** FEP+ predicted potency distribution of enumerated ligands in 1 ns FEP simulations. IC_{50} potency bins: A < 10 nM; 10 nM < B < 100 nM; 100 nM < C < 1000 nM; D > 1000 nM.

To exhaustively explore the commercially available chemical space compatible with these aryl couplings in the eMolecules building blocks library, we employed all four of the aryl couplings identified by PathFinder: Suzuki, Stille, Negishi, and decarboxylative couplings. After the creation of this data set *via in silico* enumeration, the virtual ligand set was filtered as depicted in Figure 2b. In the first step the ligand set was filtered based on physicochemical property criteria to remove ligands well outside reasonable drug-like space; here, we used a modified ‘rule of five’ for library shaping: MW < 500, PSA < 150, RB < 7, AlogP < 5, CC < 3, HBD < 5.¹¹ We also filtered the ligands using a

proprietary set of SMARTS patterns to remove known chemical liabilities and PAINS offenders.^{57,58} This removed approximately 60% of the compounds, trimming the set from greater than 300,000 compounds to approximately 110,000 unique, chemically stable, synthetically tractable, reasonably drug-like ligands. These compounds were then docked into the CDK2 receptor. There are a number of criteria that can be employed at this point to enrich the dataset (constraints, interactions, clustering etc.); we simply kept all ligands that fit into the binding site (~70,000; Figure 2c) and did not create any severe clashes with the receptor. At this point, too many ligands remained to profile the entire set with FEP, so we used a machine learning approach to enrich the dataset prior to FEP profiling. First, 1000 ligands were chosen at random for 1 ns FEP simulations, and then this data was used to create a machine learning model. The remaining ligands were then scored with the ML model, and the top 1K were passed on to FEP for 1 ns simulations. The results were used to re-train the model, and the model was used to select the next top 1K from the remaining ligands. This process was repeated 4 times, resulting in the profiling of 5,255 ligands in 1 ns FEP simulations. To our knowledge, this is the largest set of FEP calculations disclosed in the literature to date.⁵⁹ We should note that the 1 ns FEP calculations are used qualitatively to enrich the initial set of ligands for longer FEP simulations, analogous to how single point inhibition assays are used experimentally to select for compounds to run in a full dose-response curve. Ligands that were predicted to be < 10 nM in the 1 ns single edge simulations were passed on to 20 ns cycle closure FEP simulations with additional reference compounds to generate more accurate predictions.

Results from the 1 ns FEP potency predictions are presented in Figure 2d. We binned ligands based on predicted potencies; most (~80%) are predicted to be > 100 nM binders (Figure 2d; C & D bins), but a significant portion (~20%) are predicted to provide a boost in potency compared to the selected reference (Figure 2d; A & B bins). A representative set of A and B ligands that were run in the longer 20 ns cycle closure FEP simulations is shown in Table 2, and a more comprehensive set can be found in Supporting Tables S1-S3. Five- and six-member rings like pyrazole (**3**), thiophene (**4**), thiazole (**6**), and pyridine (Table S1: **S1**), replace the first phenyl and then an additional substituent fills out the space occupied by the second *m*-phenyl. A variety of polar and charged groups like lactams (**2**), amines (**3**, **4**, **5**), oxetanes (Table S1: **S8**), and nitriles (**8**) all provide diverse options to replace the *m*-phenyl moiety. These groups typically interact with D86 and/or K89 directly or *via* water-mediated hydrogen bonds, an example simulation interaction diagram (SID) illustrating these interactions for **2** can be found in Figure S2a. As one might expect, these polar groups are not required for potency

when replacing the *m*-phenyl, other hydrophobic pieces are predicted to be nearly equipotent to the *m*-biphenyl, as exemplified by the thiazole and cyclopentane of **6**; Figure S1b shows the SID for **6**.

Table 2. R-groups Identified with PathFinder Combined with ML/FEP Workflow.

	C12	C13	C73	C2	C3	2	3	4	5	6	7
R₁	H	H									
R₂	H	SO ₂ NH ₂	SO ₂ NH ₂	H	SO ₂ NH ₂	H	H	H	H	H	H
Exp. pIC₅₀	4.21	5.82	7.36	6.01	8.30	ND	ND	ND	ND	ND	ND
LE	0.36	0.40	0.31	0.34	0.40	ND	ND	ND	ND	ND	ND
FEP+ pIC₅₀	4.87	6.14	7.31	5.72	7.36	7.81	7.76	7.39	7.36	7.11	6.41
LE_{FEP+}	0.42	0.42	0.31	0.33	0.36	0.37	0.38	0.37	0.36	0.37	0.36

Notably, the R-groups sampled off of the first ring are all predicted to be more efficient (LE_{FEP+}) than a phenyl group, and are predicted to drive potency, not simply maintain it. Often, promising ideas like those shown in Table 2 might be considered by medicinal chemists on a project, but the perceived synthetic difficulty may lead to deprioritization. For example, while far from impossible to synthesize, **2** may be lower in the synthesis queue because it contains an sp³-sp² carbon-carbon bond and introduces a chiral center, which complicates the process relative to some other analogs. However, in this exercise we used PathFinder to focus on one-step reactions from commercially available building blocks to create the set of virtual ligands, so the products should be significantly easier to synthesize than typical virtual ligands.

In the process of mining the commercially available chemical space for diverse biphenyl replacements, some virtual SAR around the potent ligands was generated. For example, the compound with the highest predicted potency is **2** (Table 2), and many other analogs of **2** were run in FEP simulations (Table S1). Here, compounds **S2** and **S1** show that removing the fluorine from the phenyl or exchanging the fluoro-phenyl for pyridine has little effect on the potency, but installing a methyl at C4 of the pyrrolidone (**S16**), or a fluorine on the C2 position of the phenyl (**S9**) is predicted to reduce potency ~10x. This suggests that disrupting the torsion between the pyrrolidone and the phenyl interferes with the interactions that the pyrrolidone is making. It is interesting to note that Table S1

captures many of the potential modifications that a medicinal chemist may make to **2** while probing the pocket for potency: removing the fluorine (**S2**), pyridine replacing the phenyl (**S1**), lactam ring expansion (**S3**), methyl addition to the lactam (**S16** and **S17**), capping the lactam NH (**S19**), and moving the carbonyl outside of the ring (**S18**). One of the most efficient pieces identified is the thiazole of compound **6** (Table 2), and seven other thiazoles were profiled with FEP simulations (Table S2). These ligands show that the 2- and 4-position of the thiazole are predicted to be sensitive to modulation. For example, lipophilic groups were predicted to increase the potency up to ~3-fold at the 4-position (**S23-S26**), whereas substitution with a THF or THP at the 2-position led to a drastic loss in expected potency (**S27-S29**). Several thiophenes like **4** (Table 2) were also generated and profiled in FEP simulations; Table S3 summarizes the virtual SAR predicted for these ligands. There are no direct matched molecular pairs in this set of ligands, but we can infer that the two regioisomers of the thiophene appear to be equipotent (**4** & **S31**), a polar group is always present in place of the *m*-phenyl, and charged groups are predicted to be more potent than neutral groups (**4** & **S31** vs **S34** & **S35**).

Compounds **C2** and **C3** illustrate the large increase in potency that is possible with the addition of a sulfonamide at the C4 position of the aniline (Table 2).⁵⁰ Adding the sulfonamide to some of the most potent ligands identified in this exercise produced a similar effect. While the increase in potency in FEP was not as drastic for any of our ligands compared to the **C2/C3** pair ($\Delta pIC_{50} = 2.3$), a boost in predicted potency of ~10x was usually observed. Supporting Tables S1-S3 contain the FEP+ predicted pIC_{50} for some of the more potent compounds with and without the addition of the sulfonamide, where applicable. With the addition of the sulfonamide, many compounds are predicted to surpass the 10 nM IC_{50} threshold (Table S1: **2**, **S1**, **S2**, **S3**, **S4**, and **S9**; Table S2: **6**, **S23**, and **S24**). Interestingly, the largest predicted boost in potency with the addition of the sulfonamide is for **S9** (1.94 pIC_{50} units; Table S1), one of the most efficient ligands comes from the thiazole series (Table S2: **S24**), and the thiophene series experiences a much smaller boost in potency with the addition of the sulfonamide relative to the other series identified (all < 10x increase; Table S3). We must note that the addition of the sulfonamide does move these ligands into less drug-like space. However, the goal here was to demonstrate the value these technologies can add in a drug discovery environment, which is unambiguous. Further, one could imagine performing the same exercise as just described to replace the sulfonamide with more drug-like pieces.

Machine Learning and Population-Based Statistics

By randomly selecting initial molecules on which to run FEP simulations the total number of tight binders in the set of ligands from the R-group enumeration was estimated. Using the Wilson Score Interval with Continuity Correction the number of compounds with a given minimum binding affinity present in the set of 68,966 ligands from the R-group enumeration was estimated (see Table 3).⁶⁰

Table 3. Estimated number of ligands with a predicted binding affinity using the Wilson Score Interval with Continuity Correction.

Binding Affinity	Estimated number of ligands	95% Confidence Interval ^a	90% Confidence Interval ^a
< 10 nM	197	[51, 626]	[61, 539]
< 100 nM	2301	[1633, 3213]	[1722,3054]
< 1000 nM	14,332	[12682, 16131]	[12933, 15839]

^aConfidence intervals are given with their lower and upper bound respectively.

After the initial random selection of molecules a machine learning model was trained on the FEP results using AutoQSAR/DC.^{61,62} The entire set of ligands from the R-group enumeration was then scored using the machine learning model and FEP simulations were run on the compounds that were predicted to have the highest binding affinity.

As shown in Table 4, after four rounds of active learning, the ML model succeeded to recover approximately 61% of the 197 single-digit nanomolar binding affinity compounds expected to exist in the full 68,966 compound set. Furthermore, this 61% recovery of tight binding compounds required free energy calculation scoring of only 5.8% of the 68,966 compounds. Thus, the active learning approach reported here makes processing very large libraries of compounds, potentially millions of compounds, tractable achieving free energy calculation-like accuracy at acceptable computational cost.

We note in passing here that there is an ambiguity in the machine learning community regarding whether or not this workflow should be considered a reinforcement learning approach, an active learning approach, or both. Here we use the term active learning, since each iteration of the workflow can be considered to be a prompt to run additional FEP calculations, expanding the training set of the ML.⁶³

Table 4. Total number of ligands analyzed by FEP and their predicted binding affinity for each round of the ML-FEP workflow.

Round	Molecule Selection	# Molecules	< 10 nM	< 100nM	< 1000 nM
1	Random	1049	3	35	218
2	Machine Learning	1020	26	186	625
3	Machine Learning	984	52	269	670
4	Machine Learning	984	40	218	560
Total	All	4037	121	708	2073
% EV	All	5.9%	61%	31%	15%
95UB	All	5.9%	20%	22%	13%

Likewise, the approach presented here, could also be considered a reinforcement learning workflow, as the machine learning model is used to attempt to select molecules predicted to be tight binding by free energy calculations, and it is retrained after each trial round to improve its performance at this particular goal.⁶⁴

Core Hop Enumeration

In addition to the identification of a biphenyl replacement for Compound **C73**, we aimed to use PathFinder-driven reaction-based enumeration followed by free energy calculations to find alternative cores for the purine core of **C73**. This process begins by identifying a route that allows for the insertion of cores onto the already existing R-groups. We removed the sulfonamide from **C73** (Figure 3a) and used the resulting structure as input for a retrosynthetic analysis with a depth of 2 in PathFinder. This generated many routes for potential core replacement; an example is shown in Figure 3a. In this example, a Suzuki coupling is used to couple the biphenyl to the purine core, and a Buchwald coupling is used to couple the aniline to the purine core. Using this chemistry one can plug in various 6-member rings and bicyclic systems (5,6-, 6,6-, etc.) with the same geometric arrangement of chemical handles to replace the purine while maintaining similar vectors to the R-groups of the input structure (Figure 3a). Next, we expanded the set of cores by exploring all of the available aryl-

aryl couplings simultaneously; commercially available cores that were amenable to these reactions, and possessed similar vectors for the R-groups were considered (Figure 3b). To gather these cores, we searched the eMolecules database for 6-, 5,6- and 6,6-(bi)cyclic systems that have X_1 and X_2 substituents where X_1 = aryl halide (excluding F), carboxylate, or boronate (for aryl C-C coupling reactions), and X_2 = Cl, Br, or I (for Buchwald coupling; Figure 3b). To further expand the dataset, we decided to simplify the chemistry by requiring only the Buchwald coupling step. Therefore, only the leaving group for the Buchwald reaction was necessary to be present (X_2 = Cl, Br, or I). Based on the large number of cores identified and profiled (>50K), it would be very difficult and time consuming to identify all of these possibilities using traditional design approaches alone.

After enumerating all amenable cores with the biphenyl and aniline R-groups, we filtered the set using a similar protocol as previously described for the R-group enumeration (Figure 3c). The remaining ~10K ligands were docked into the CDK2 binding site (Figure 3d) and a constraint requiring two hydrogen bonds with the hinge was enforced to enrich the set of ligands for FEP simulations. At this point 725 ligands remained, which was small enough to be directly profiled through 20 ns cycle closure FEP simulations. As with the biphenyl replacements, we binned the predicted FEP+ potency values (Figure 3e). Many of the cores are predicted to be very weak binders (> 1 μ M; Figure 3e, C & D bins), but several of the new cores fall into the A category (FEP+ pred. IC_{50} < 100 nM) and many more are predicted to be < 1 μ M binders (bin B). It is important to note that in a real-world drug discovery effort, finding 1-2 novel cores would generally be considered a success.

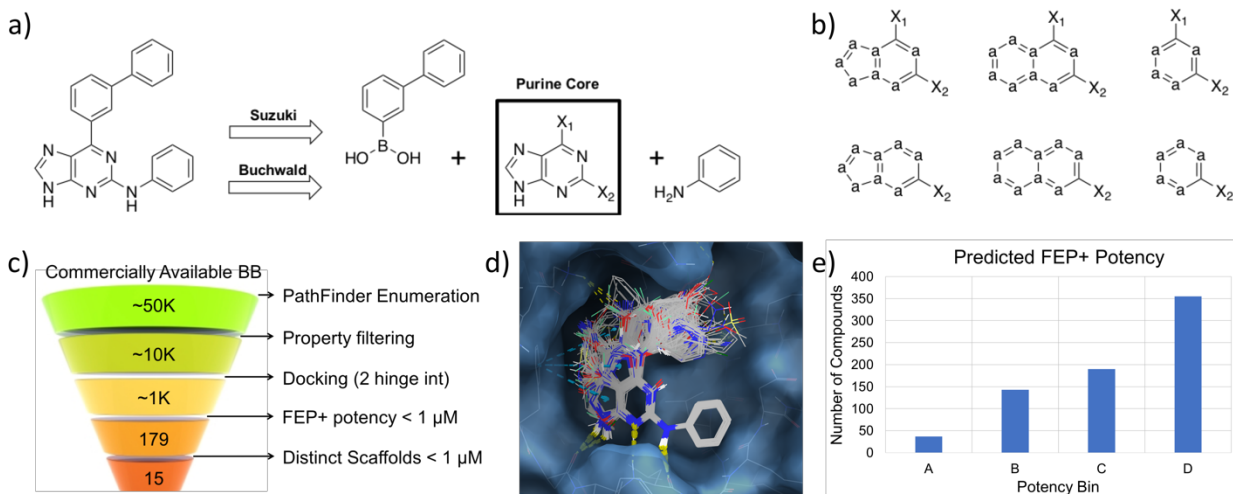


Figure 3. PathFinder Core Hopping Workflow and Results. **a)** Retrosynthetic analysis to identify purine core replacements. **b)** Commercially available cores considered for purine replacements: 6-, 5,6- and 6,6- (bi)cyclic systems requiring one- or two-step reactions. **c)** Funnel depicting the filtering process of the enumerated library: drug-like property filtering, docking with hinge constraint, and FEP+ profiling of all surviving ligands. **d)** Input poses of all 725 core hop ligands for FEP simulations. **e)** FEP+ predicted potency distribution of enumerated ligands. IC_{50} potency bins: A < 100 nM; 100 nM < B < 1 μ M; 1 μ M < C < 10 μ M; D > 10 μ M.

An example of each representative core from the A and B bins of Figure 3e is pictured in Figure 4. The A class contains four distinct cores represented by **9**, **10**, **11**, and **12** in Figure 4. Compounds containing the A bin scaffolds are observed in the B bin depending on the R-groups, and the B bin contains eleven additional distinct scaffolds (Figure 4; **12-22**). Many of the scaffolds in Figure 4 possess typical hinge-binding motifs found in kinase inhibitors, and a number of the cores are indeed present in marketed drugs: **8** (imatinib), **10** (palbociclib), **12** (vorafenib), **13** (gefitinib), and **14** (axitinib).

With this in mind, one may assume that a portion of these virtual ligands was previously evaluated against one of the 500+ kinases in the human kinome. To assess this, we searched the BindingDB^{65,66} for exact matches, and similar structures, of all 725 core hops evaluated with FEP. Unexpectedly, only 3/725 compounds have exact matches in the BindingDB, and of these three, two have reported CDK2 inhibition data (Figure 5A; **23-24**). These two data points provide blinded validation of the FEP potency values reported in the R-group and core hop exercises (Table 5), and are within the expected accuracy of FEP+.⁵⁶

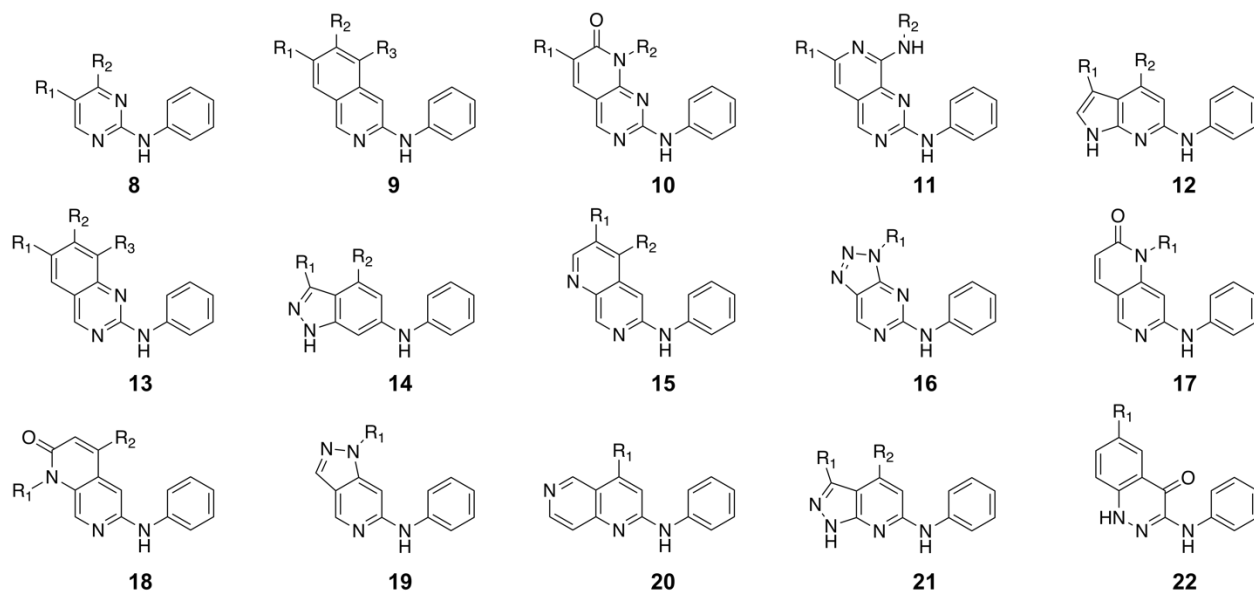


Figure 4. Purine Replacements Identified with PathFinder Core Hop Enumerations. All cores (8-22) have at least one compound with a predicted FEP+ $IC_{50} < 1 \mu M$. Cores 8-11 have analogs with predicted FEP+ $IC_{50} < 100 nM$.

Compound **25** (Figure 5a) is the most potent ligand identified in the core hop enumeration with a predicted FEP+ pIC_{50} of 7.89 (Table 6), and the most similar compound found in the BindingDB is **26** (Figure 5a), but no CDK2 inhibition data was available in the BindingDB. This, combined with only 3/725 identical matches in BindingDB suggests that the breadth of cores explored using a simple one- or two-step enumeration from commercially available building blocks is distinct relative to known chemical matter. Further, the majority of ligands that showed potent inhibition of CDK2 in FEP can be synthesized in a one-step reaction; a Buchwald amination from the commercially available cores and aniline, which makes the synthesis of these compounds relatively straightforward.

Table 5. Blind Validation of FEP+ in Core Hop Enumeration.

Compound	23	24
Exp. pIC_{50}	6.89	6.69
FEP+ pIC_{50}	7.06	6.39

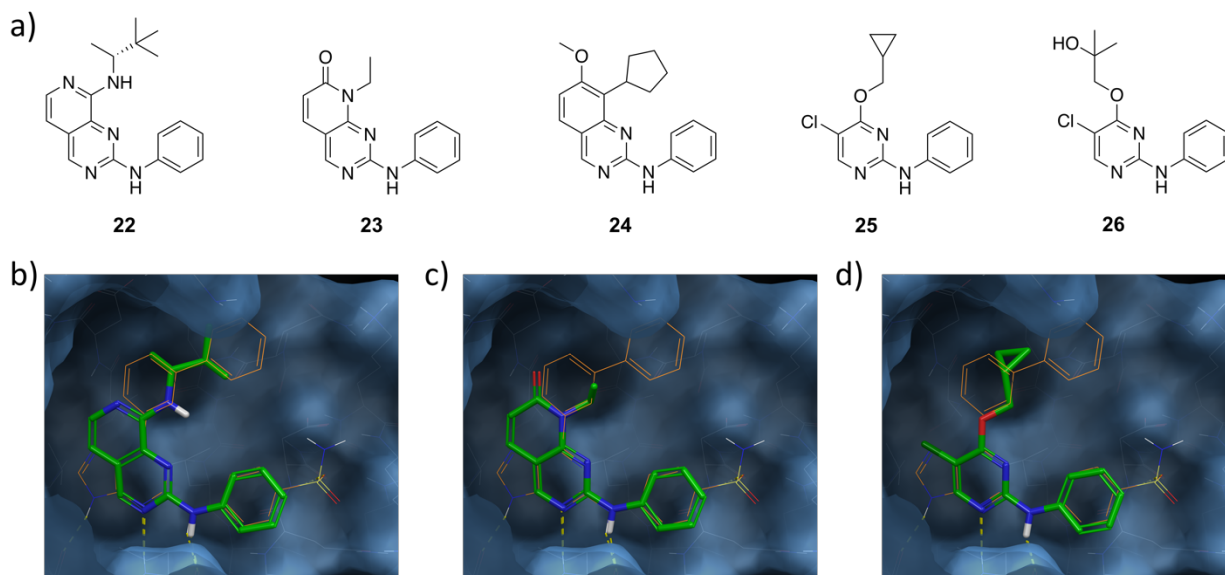


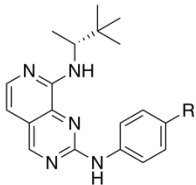
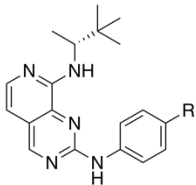
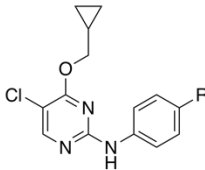
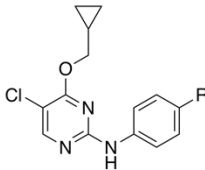
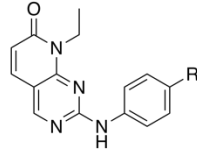
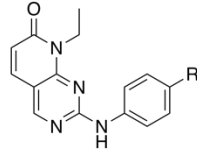
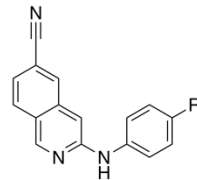
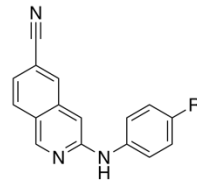
Figure 5. PathFinder Core Hopping Enumeration. a) FEP simulations predict compounds **22-25** to be potent CDK2 inhibitors. **22** is the most potent inhibitor after installation of the sulfonamide, **23** and **24** were previously assayed against CDK2, **25** is the most potent compound identified before installation of the sulfonamide, and **26** is the closest analog of **25** in the BindingDB. b) Proposed binding mode of **22**. c) Proposed binding mode of **23**. d) Proposed binding mode of **25**. Compound **C73** pictured in orange for reference.

The last step of the core hop enumeration was to install the sulfonamide in order to observe the boost in potency seen both experimentally and in the FEP simulations. To do this, compounds that were predicted to be potent from the set of cores in Figure 4 were selected and the sulfonamide was added to the 4-position of the aniline; representative compounds are depicted in Table 6. Addition of the sulfonamide is predicted to increase the potency in all cases, but not uniformly. For example, **27** is > 2 pIC₅₀ units more potent than **22**, and **28** is > 1 pIC₅₀ unit more potent than **25**.

However, **29** and **31** are only ~ 0.5 pIC₅₀ units more potent than **23** and **30**, respectively. Compounds **27** and **28** are the most potent compounds identified in the core hop enumeration workflow. In fact, **25** was the most efficient core hop found in the initial FEP core hop enumeration ($LE_{FEP+} = 0.42$), and it remained the most efficient after adding the sulfonamide **28** ($LE_{FEP+} = 0.40$). Further, both **27** and **28** have a predicted LLE (LLE_{FEP+}) > 6 which is considered of high quality in a drug discovery project. As with the ligands identified in the R-group optimization exercise, the final ligands (after sulfonamide addition) tend to move into less drug-like space due to properties, with high PSA being one of the main disadvantages. A possible future direction could be to use this workflow

to find replacements for the sulfonamide that maintain/improve potency but have better drug-like properties.

Table 6. Addition of Sulfonamide to Potent Cores Identified from PathFinder Enumeration Followed by FEP.

								
	22	27	25	28	23	29	30	31
R	H	SO ₂ NH ₂	H	SO ₂ NH ₂	H	SO ₂ NH ₂	H	SO ₂ NH ₂
FEP+ pIC ₅₀	7.12	9.86	7.89	9.26	7.06	7.62	7.77	8.24
LE _{FEP+}	0.30	0.35	0.42	0.40	0.35	0.32	0.41	0.36
LLE _{FEP+}	2.62	6.69	3.76	6.46	3.76	5.64	4.22	6.01

Chemical Space Explored During PathFinder/ML/FEP+ Workflow

In silico enumeration provides a more exhaustive exploration of available chemical space than traditional empirical SAR studies. To illustrate this, we created self-organizing maps (SOMs; Figure 6) by pooling the ligands reported by Coxon *et al.*⁵⁰ and all of the ligands evaluated by FEP+ during the R-group and core hop exercises. Each cell in the SOM indicates a section of chemical space as described by the fingerprint used to create the map; the color of the cell corresponds to the number of compounds contained in the cell, and the shading of the border indicates the degree of similarity between neighboring cells. The SOMs cover a variety of ligand populations: the chemical space considered (Figure 6a), the ligands reported by Coxon *et al.*⁵⁰ (Figure 6b), the pool of R-group ligands run in FEP simulations (Figure 6c), the pool of core hop ligands run in FEP simulations (Figure 6d), R-group ligands predicted by FEP+ to have a pIC₅₀ > 7 (Figure 6e), and core hops predicted by FEP+ to have a pIC₅₀ > 6 (Figure 6f). By comparing Figure 6b, 6c, and 6d it is clear that a larger variety of ligands was explored in the enumeration exercises (R-group and core hop) than the ligands reported by Coxon *et al.*⁵⁰ Further, Figure 6e shows the R-group ligands with FEP+ predicted pIC₅₀ > 7, and these R-groups belong to distinct regions within the chemical space profiled. This diversity provides access to chemical space that otherwise may not be considered, presents a less biased approach

towards potency optimization, and gives the researcher a myriad of avenues to pursue from a property perspective, while maintaining on-target potency.

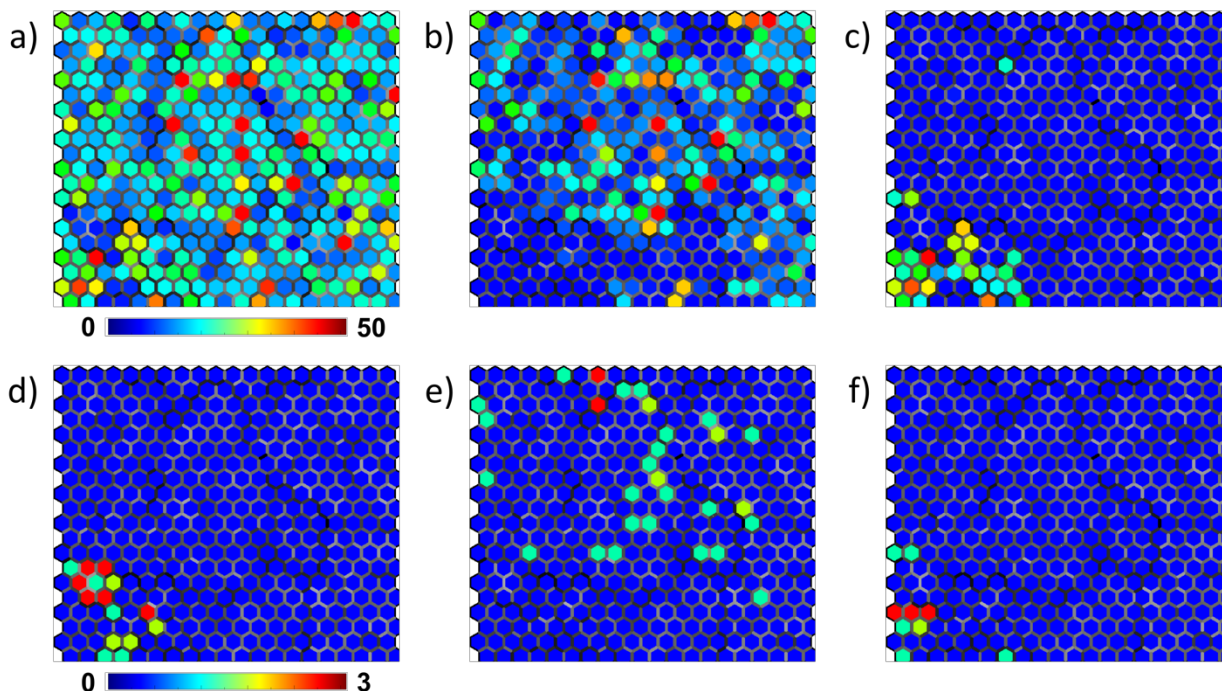


Figure 6. Self-Organizing Maps. **a)** Chemical space covered by all ligands. **b)** Chemical space covered by R-groups from FEP simulations for biphenyl replacements of Compound 73. **c)** Chemical space covered by core hops from FEP simulations to replace the purine of Compound 73. **d)** Chemical space covered by the ligands by Coxon *et al.*⁵⁰ **e)** Chemical space covered by R-group ligands (without sulfonamide) with FEP+ predicted $pIC_{50} > 7$. **f)** Chemical space covered by core hops (without sulfonamide) with FEP+ predicted $pIC_{50} > 7$. Self-Organizing Maps generated using Canvas, Schrödinger, LLC, New York, NY, 2019.

Conclusion

Recent increases in computational power allow for routine profiling of much larger ligand libraries with free energy calculations than has previously been possible. This has created a clear need for technologies that are able to rapidly generate relevant chemical matter for drug discovery on a timeline that can have impact on drug discovery projects. To address this need, we created PathFinder and demonstrated the application of enumerating a large library from commercially available building blocks, enriching the library with machine learning, and profiling the potency of those ligands with FEP simulations using CDK2 as a model system. We were able to rapidly create libraries using reaction-based enumeration in PathFinder to generate both R-group and core replacements of Compound **C73**. Four rounds of machine learning-enriched FEP profiling identified 35 unique R-

groups with predicted FEP+ $IC_{50} < 100$ nM, and one round of core hop enumeration identified 15 distinct cores with predicted FEP+ $IC_{50} < 1$ μ M potency. Further, addition of the sulfonamide, present in compound **C73**, to PathFinder-identified ligands resulted in 35 compounds (R-groups and core hops) with predicted FEP+ $IC_{50} < 10$ nM for CDK2. The potency boost of the sulfonamide moiety comes at the expense of moving the ligands into less drug-like space (Higher PSA and MW), but the principle of exploring more chemical space quickly to improve the on target potency is unequivocal.

The cores identified during the core hop exercise display traditional kinase hinge binding features, yet only 3/725 corehops profiled had exact matches in the BindingDB, and only two of those had been assayed against CDK2. This is surprising, because all of the identified cores can be made from one or two-step syntheses starting from commercially available reactants, and one would expect that most of the low hanging fruit in a target class as crowded as the kinase space would have been tested at this point. For example, our potency optimization workflow identified simple quinazolines, isoquinolines, and pyrimidines that were predicted to be potent (FEP+ $IC_{50} < 100$ nM), and to the best of our knowledge, had not been disclosed previously. This exemplifies the importance of exhaustively evaluating commercially available chemical space during the course of a drug discovery project. Intuitively, evaluating the most likely to be synthesized, and easiest to attain chemical matter against the target of interest is a good initial approach. In our workflow, PathFinder is able to generate compounds in the available chemical space, and machine learning allows for enrichment of the dataset into portions amenable to FEP+ potency profiling. The initial set of FEP+ potency predictions can be used to dive deeper into the chemical space of interest. By combining the potent R-groups and core hops identified in the first rounds of ideation one can identify more drug-like ligands with the desired potency. We expect the utility of this approach to significantly accelerate the hit-to-lead, and lead optimization processes by rapidly producing synthetically tractable, potent compounds within a desired property space.

Acknowledgments

The authors would like to thank Ronjung He for his contribution to multiple SMARTS patterns used in PathFinder, and Yutong Zhao for many useful discussions involving SMARTS and RDKit.

References

- (1) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discov.* **2004**, *3*, 935.
- (2) Lionta, E.; Spyrou, G.; Vassilatis, D.; Cournia, Z. Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. *Curr. Top. Med. Chem.* **2014**, *14* (16), 1923–1938. <https://doi.org/10.2174/1568026614666140929124445>.
- (3) Schneider, G. Virtual Screening: An Endless Staircase? *Nat. Rev. Drug Discov.* **2010**, *9*, 273.
- (4) Lyu, J.; Wang, S.; Balias, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; Meara, M. J. O.; Che, T. Ultra-Large Library Docking for Discovering New Chemotypes. *Nature*. <https://doi.org/10.1038/s41586-019-0917-9>.
- (5) Schneider, G.; Fechner, U. Computer-Based de Novo Design of Drug-like Molecules. *Nat. Rev. Drug Discov.* **2005**, *4* (8), 649–663. <https://doi.org/10.1038/nrd1799>.
- (6) Boda, K.; Seidel, T.; Gasteiger, J. Structure and Reaction Based Evaluation of Synthetic Accessibility. *J. Comput. Aided. Mol. Des.* **2007**, *21* (6), 311–325. <https://doi.org/10.1007/s10822-006-9099-2>.
- (7) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to Improve RD Productivity: The Pharmaceutical Industry's Grand Challenge. *Nat. Rev. Drug Discov.* **2010**, *9* (3), 203–214. <https://doi.org/10.1038/nrd3078>.
- (8) Ertl, P. Cheminformatics Analysis of Organic Substituents: Identification of the Most Common Substituents, Calculation of Substituent Properties, and Automatic Identification of Drug-like Bioisosteric Groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 374–380. <https://doi.org/10.1021/ci0255782>.
- (9) Bohacek, R. S.; McMartin, C.; Guida, W. C. The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective. *Med. Res. Rev.* **1996**, *16* (1), 3–50. [https://doi.org/10.1002/\(SICI\)1098-1128\(199601\)16:1<3::AID-MED1>3.0.CO;2-6](https://doi.org/10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6).
- (10) Kirkpatrick, P.; Ellis, C. Nature Insight - Chemical Space. *Nature* **2004**, *432* (December), 823 and refs. therein. <https://doi.org/10.1089/ars.2011.4276>.

- (11) 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 Deliv. Rev.* **1997**, *23*, 3–25. [https://doi.org/10.1016/S0169-409X\(00\)00129-0](https://doi.org/10.1016/S0169-409X(00)00129-0).
- (12) Hartenfeller, M.; Schneider, G. Enabling Future Drug Discovery by de Novo Design. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1* (5), 742–759. <https://doi.org/10.1002/wcms.49>.
- (13) Dean, P. M.; Lloyd, D. G.; Todorov, N. P. De Novo Drug Design: Integration of Structure-Based and Ligand-Based Methods. *Curr. Opin. Drug Discov. Devel.* **2004**, *7* 3, 347–353.
- (14) Kutchukian, P. S.; Shakhnovich, E. I. De Novo Design: Balancing Novelty and Confined Chemical Space. *Expert Opin. Drug Discov.* **2010**, *5* (8), 789–812. <https://doi.org/10.1517/17460441.2010.497534>.
- (15) Vinkers, H. M.; Jonge, M. R. De; Daeyaert, F. F. D.; Heeres, J.; Koymans, L. M. H.; Lenthe, J. H. Van; Lewi, P. J.; Timmerman, H.; Aken, K. Van; Janssen, P. A. J.; et al. SYNOPSIS: SYNthesize and OPTimize System in Silico. *J. Med. Chem.* **2003**, *46* (13), 2765–2773. <https://doi.org/10.1021/jm030809x>.
- (16) Patel, H.; Bodkin, M. J.; Chen, B.; Gillet, V. J. Knowledge-Based Approach to *de Novo* Design Using Reaction Vectors. *J. Chem. Inf. Model.* **2009**, *49* (5), 1163–1184. <https://doi.org/10.1021/ci800413m>.
- (17) Firth, N. C.; Atrash, B.; Brown, N.; Blagg, J. MOARF, an Integrated Workflow for Multiobjective Optimization: Implementation, Synthesis, and Biological Evaluation. *J. Chem. Inf. Model.* **2015**, *55* (6), 1169–1180. <https://doi.org/10.1021/acs.jcim.5b00073>.
- (18) Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G. Dogs: Reaction-Driven de Novo Design of Bioactive Compounds. *PLoS Comput. Biol.* **2012**, *8* (2). <https://doi.org/10.1371/journal.pcbi.1002380>.
- (19) Gupta, A.; Müller, A. T.; Huisman, B. J. H.; Fuchs, J. A.; Schneider, P.; Schneider, G. Generative Recurrent Networks for De Novo Drug Design. *Mol. Inform.* **2018**, *37* (1), 1–10. <https://doi.org/10.1002/minf.201700111>.

- (20) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discov. Today* **2018**, *23* (6), 1241–1250. <https://doi.org/10.1016/j.drudis.2018.01.039>.
- (21) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4* (2), 268–276. <https://doi.org/10.1021/acscentsci.7b00572>.
- (22) Kadurin, A.; Nikolenko, S.; Khrabrov, K.; Aliper, A.; Zhavoronkov, A. DruGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico. *Mol. Pharm.* **2017**, *14* (9), 3098–3104. <https://doi.org/10.1021/acs.molpharmaceut.7b00346>.
- (23) Blaschke, T.; Olivecrona, M.; Engkvist, O.; Bajorath, J.; Chen, H. Application of Generative Autoencoder in De Novo Molecular Design. *Mol. Inform.* **2018**, *37* (1), 1–11. <https://doi.org/10.1002/minf.201700123>.
- (24) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4* (1), 120–131. <https://doi.org/10.1021/acscentsci.7b00512>.
- (25) Corey, E. J. General Methods for the Construction of Complex Molecules. *Pure and Applied Chemistry*. 1967, p 19. <https://doi.org/10.1351/pac196714010019>.
- (26) Corey, E. J.; Wipke, W. T. Computer-Assisted Design of Complex Organic Syntheses. *Science (80-.)*. **1969**, *166* (3902), 178 LP-192. <https://doi.org/10.1126/science.166.3902.178>.
- (27) Corey, E. J.; Wipke, W. T.; Cramer, R. D.; Howe, W. J. Computer-Assisted Synthetic Analysis. Facile Man-Machine Communication of Chemical Structure by Interactive Computer Graphics. *J. Am. Chem. Soc.* **1972**, *94* (2), 421–430. <https://doi.org/10.1021/ja00757a020>.
- (28) 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* (3), 522–532. <https://doi.org/10.1016/j.chempr.2018.02.002>.

- (29) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. *Computer-Assisted Synthetic Planning: The End of the Beginning*, 2016; Vol. 55. <https://doi.org/10.1002/anie.201506101>.
- (30) Bøgevig, A.; Federsel, H. J.; Huerta, F.; Hutchings, M. G.; Kraut, H.; Langer, T.; Löw, P.; Oppawsky, C.; Rein, T.; Saller, H. Route Design in the 21st Century: The IC SYNTH Software Tool as an Idea Generator for Synthesis Prediction. *Org. Process Res. Dev.* **2015**, *19* (2). <https://doi.org/10.1021/op500373e>.
- (31) Law, J.; Zsoldos, Z.; Simon, A.; Reid, D.; Liu, Y.; Khew, S. Y.; Johnson, A. P.; Major, S.; Wade, R. A.; Ando, H. Y.; et al. Route Designer: A Retrosynthetic Analysis Tool Utilizing Automated Retrosynthetic Rule Generation. *J. Chem. Inf. Model.* **2009**, *49* (3), 593. <https://doi.org/10.1021/ci800228y>.
- (32) Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K. H.; Schneider, G.; Jacoby, E.; Renner, S. A Collection of Robust Organic Synthesis Reactions for in Silico Molecule Design. *J. Chem. Inf. Model.* **2011**, *51* (12), 3093–3098. <https://doi.org/10.1021/ci200379p>.
- (33) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555* (7698), 604–610. <https://doi.org/10.1038/nature25978>.
- (34) Segler, M. H. S.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. - A Eur. J.* **2017**, *23* (25), 5966–5971. <https://doi.org/10.1002/chem.201605499>.
- (35) <http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>.
- (36) Roughley, S. D.; Jordan, A. M. The Medicinal Chemist's Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates. *J. Med. Chem.* **2011**, *54* (10), 3451–3479. <https://doi.org/10.1021/jm200187y>.
- (37) Christ, C. D.; Zentgraf, M.; Kriegl, J. M. Mining Electronic Laboratory Notebooks: Analysis, Retrosynthesis, and Reaction Based Enumeration. *J. Chem. Inf. Model.* **2012**, *52* (7), 1745–1756. <https://doi.org/10.1021/ci300116p>.

- (38) Taylor, R. D.; Maccoss, M.; Lawson, A. D. G. Rings in Drugs. *J. Med. Chem.* **2014**, *57* (14), 5845–5859. <https://doi.org/10.1021/jm4017625>.
- (39) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887–2893. <https://doi.org/10.1021/jm9602928>.
- (40) <https://www.rdkit.org>.
- (41) 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* (7), 1625–1632. <https://doi.org/10.1021/acs.accounts.7b00083>.
- (42) <https://www.emolecules.com/>.
- (43) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739–1749. <https://doi.org/10.1021/jm0306430>.
- (44) Halgren, T.A.; Murphy, R.B.; Friesner, R.A.; Beard, H.S.; Frye, L.L.; Pollard, W.T.; Banks, J. L.; Glide. A New Approach For Rapid, Accurate Docking and Scoring. II. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, *2* (47), 1750–1759.
- (45) Jaques, N.; Gu, S.; Bahdanau, D.; Hernández-Lobato, J. M.; Turner, R. E.; Eck, D. Sequence Tutor: Conservative Fine-Tuning of Sequence Generation Models with KL-Control. **2016**.
- (46) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular De-Novo Design through Deep Reinforcement Learning. *J. Cheminform.* **2017**, *9* (1), 1–14. <https://doi.org/10.1186/s13321-017-0235-x>.
- (47) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; et al. 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* (7), 2695–2703. <https://doi.org/10.1021/ja512751q>.
- (48) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The Missing Term in Effective Pair

- Potentials. *J. Phys. Chem.* **1987**, *91* (24), 6269–6271. <https://doi.org/10.1021/j100308a038>.
- (49) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; et al. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, *12* (1), 281–296. <https://doi.org/10.1021/acs.jctc.5b00864>.
- (50) Coxon, C. R.; Anscombe, E.; Harnor, S. J.; Martin, M. P.; Carbain, B.; Golding, B. T.; Hardcastle, I. R.; Harlow, L. K.; Korolchuk, S.; Matheson, C. J.; et al. Cyclin-Dependent Kinase (CDK) Inhibitors: Structure-Activity Relationships and Insights into the CDK-2 Selectivity of 6-Substituted 2-Arylamino-purines. *J. Med. Chem.* **2017**, *60* (5), 1746–1767. <https://doi.org/10.1021/acs.jmedchem.6b01254>.
- (51) Cohen, P. Protein Kinases — the Major Drug Targets of the Twenty-First Century? *Nat. Rev. Drug Discov.* **2002**, *1*, 309.
- (52) Roskoski, R. A Historical Overview of Protein Kinases and Their Targeted Small Molecule Inhibitors. *Pharmacol. Res.* **2015**, *100*, 1–23. <https://doi.org/10.1016/j.phrs.2015.07.010>.
- (53) Fabbro, D. 25 Years of Small Molecular Weight Kinase Inhibitors: Potentials and Limitations. *Mol. Pharmacol.* **2015**, *87* (5), 766–775. <https://doi.org/10.1124/mol.114.095489>.
- (54) Malumbres, M.; Barbacid, M. Cell Cycle, CDKs and Cancer: A Changing Paradigm. *Nat. Rev. Cancer* **2009**, *9*, 153.
- (55) Zhang, J.; Yang, P. L.; Gray, N. S. Targeting Cancer with Small Molecule Kinase Inhibitors. *Nat. Rev. Cancer* **2009**, *9*, 28.
- (56) Wang, L.; Deng, Y.; Knight, J. L.; Wu, Y.; Kim, B.; Sherman, W.; Shelley, J. C.; Lin, T.; Abel, R. Modeling Local Structural Rearrangements Using FEP/REST: Application to Relative Binding Affinity Predictions of CDK2 Inhibitors. *J. Chem. Theory Comput.* **2013**, *9* (2), 1282–1293. <https://doi.org/10.1021/ct300911a>.
- (57) Dahlin, J. L.; Nissink, J. W. M.; Strasser, J. M.; Francis, S.; Higgins, L.; Zhou, H.; Zhang, Z.; Walters, M. A. PAINS in the Assay: Chemical Mechanisms of Assay Interference and Promiscuous Enzymatic Inhibition Observed during a Sulfhydryl-Scavenging HTS. *J. Med.*

- Chem.* **2015**, *58* (5), 2091–2113. <https://doi.org/10.1021/jm5019093>.
- (58) Filtering ~1 million compounds against 200 patterns takes approximately 30 min on 24 cores.
- (59) Vilseck, J. Z.; Armacost, K. A.; Hayes, R. L.; Goh, G. B.; Brooks, C. L. Predicting Binding Free Energies in a Large Combinatorial Chemical Space Using Multisite λ Dynamics. *J. Phys. Chem. Lett.* **2018**, *9* (12), 3328–3332. <https://doi.org/10.1021/acs.jpcclett.8b01284>.
- (60) Newcombe, R. G. Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods. *Stat. Med.* **1998**, *17* (8), 857–872. [https://doi.org/10.1002/\(SICI\)1097-0258\(19980430\)17:8<857::AID-SIM777>3.0.CO;2-E](https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E).
- (61) Dixon, S. L.; Duan, J.; Smith, E.; Von Bargen, C. D.; Sherman, W.; Repasky, M. P. AutoQSAR: An Automated Machine Learning Tool for Best-Practice Quantitative Structure–Activity Relationship Modeling. *Future Med. Chem.* **2016**, *8* (15), 1825–1839. <https://doi.org/10.4155/fmc-2016-0093>.
- (62) 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* (2). <https://doi.org/10.1039/c7sc02664a>.
- (63) Settles, B. *Active Learning Literature Survey*; 2009.
- (64) Sutton, R. S.; Barto, A. G. *Reinforcement Learning: An Introduction*, Second Edi.; The MIT Press, Cambridge MA, 2018.
- (65) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A Public Database for Medicinal Chemistry, Computational Chemistry and Systems Pharmacology. *Nucleic Acids Res.* **2016**, *44* (D1), D1045–D1053. <https://doi.org/10.1093/nar/gkv1072>.
- (66) <http://www.bindingdb.org/bind/index.jsp>.