JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

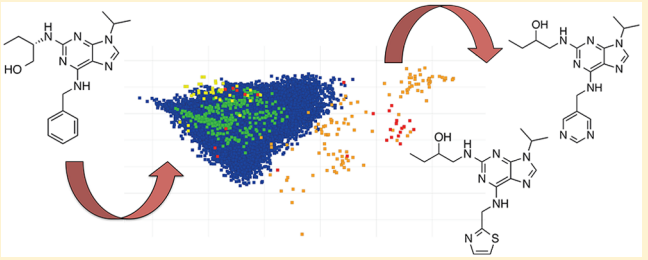
MOARF, an Integrated Workflow for Multiobjective Optimization: Implementation, Synthesis, and Biological Evaluation

Nicholas C. Firth, Butrus Atrash, Nathan Brown,* and Julian Blagg*

Cancer Research UK Cancer Therapeutics Unit, Division of Cancer Therapeutics, The Institute of Cancer Research, London, SM2 5NG, U.K.

Supporting Information

ABSTRACT: We describe the development and application of an integrated, multiobjective optimization workflow (MOARF) for directed medicinal chemistry design. This workflow couples a rule-based molecular fragmentation scheme (SynDiR) with a pharmacophore fingerprint-based fragment replacement algorithm (RATS) to broaden the scope of reconnection options considered in the generation of potential solution structures. Solutions are ranked by a multiobjective scoring algorithm comprising ligand-based (shape similarity) biochemical activity predictions as well as physicochemical property calculations. Application of this iterative workflow to optimization of the CDK2 inhibitor Seliciclib (CYC202, R-roscovitine) generated solution molecules in desired physicochemical property space. Synthesis and experimental evaluation of optimal solution molecules demonstrates CDK2 biochemical activity and improved human metabolic stability.



INTRODUCTION

A major challenge in small molecule drug discovery is the efficient exploration of chemical space toward desired program objectives with synthesis of the minimum number of molecules. An increased understanding of the many factors driving successful drug design, while also avoiding compound-related attrition, has resulted in a corresponding expansion in the number of parameters that should be considered in medicinal chemistry design: the multiobjective optimization challenge.[1−3] Improving potency and selectivity against a biological target should ideally be evaluated alongside, for example, lipophilicity and appropriate ligand efficiency metrics to increase the probability of optimal metabolic stability and membrane permeability while also minimizing binding promiscuity and transporter affinity.[4−6] The increasing number of concurrent parameters which require optimization in a modern drug discovery program may not be realizable in a single molecule without very significant exploration of chemical space. Thorough experimental exploration of all the chemical space is not realistic despite cumulative advances in rapid parallel synthetic chemistry techniques over recent decades[7] and the emergence of diversity-oriented synthesis (DOS) as a chemistry paradigm.[8] Premature focusing of synthetic efforts onto local regions of chemical space, coupled with a common desire to stay close to what is known, may neglect opportunities more likely to contain solutions that meet the most program objectives.

The need to objectively explore broad chemical space and to address the associated challenges in multiobjective optimization represents a significant task to which computational methods have been applied. *De novo* design (DND) is an example of such a computational method and multiobjective DND integrates the multiobjective optimization challenge into the design workflow. DND methods can be divided into two typical approaches: ligand-based and structure-based. Ligand-based methods use only ligand information, such as molecular similarity or activity models, whereas structure-based methods use protein structures to optimize designed ligands in the presence of the target binding site. A number of ligand-based and structure-based DND methods have been published, including CoG[9] and IADE,[10,11] and SPROUT,[12] LUDI,[13] and LigBuilder,[14] respectively. Multiobjective methods that include both structure- and ligand-based scoring of virtual compounds have also been published, including MEGA,[15] Molecule Commander,[16] and Muse.[17]

DND approaches can also be classified with regard to the method by which new potential ligands are generated: atom-based, fragment-based, or reaction-based. Atom-based methods perturb candidate solutions atom-by-atom, allowing for full exploration of the chemistry search space, but this can result in chemical structures that are synthetically unfeasible.[18] Conversely, reaction-based methods explicitly take into account synthetic feasibility by encoding precedented building blocks and reaction transforms resulting in structures that are, in theory, synthetically feasible;[19] however, this may reduce the space of potential solutions explored, and lead to issues in scoring solutions during optimization.[20]

An appropriate balance between atom-based and reaction-based methods is a fragment-based DND system.[21] Retro-

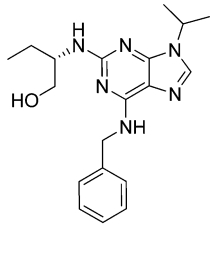
Received: February 8, 2015
Published: June 9, 2015

DOI: 10.1021/acs.jcim.5b00073
J. Chem. Inf. Model. 2015, 55, 1169−1180

synthetic fragmentation rules can be applied to databases of chemical structures to appropriately extract synthetically relevant fragments. The term fragment in this context, and henceforth in this study, refers to the constituent molecular building blocks that serve as the basis for fragment-based DND, as opposed to fragment molecules used for fragment-based screening and drug design.[22] A fragment-based DND approach implicitly takes into account synthetic feasibility, with the assumption that the retrosynthetic strategy is appropriate. Lastly, as fragment-based methods are not constrained by a defined set of reactions, they permit more extensive exploration of chemical space than reaction-based methods while also offering greater control over synthetic feasibility than atom-based methods.

We present here a fully integrated and adaptive fragment-based DND workflow, multiobjective automated replacement of fragments (MOARF). MOARF evolves to optimized, drug-like molecules using in silico fragment generation (SynDiR) and fragment replacement (RATS) algorithms. MOARF optimization may be guided by both structure-based and ligand-based prediction of target biological effects as well as simultaneous optimization of critical physicochemical parameters in medicinal chemistry (ClogP, topological polar surface area (TPSA)[23] and molecular weight (MW)) that, taken together, are likely to strongly influence important drug attributes such as metabolic stability, membrane permeability, transporter affinity and promiscuity against biological targets.[4,5,24,25] We apply a stochastic procedure, based upon circular and topological fingerprint similarity, to select replacement fragments from a comprehensive database; this approach uses no predefined rules of fragment placement, maximizing the available search space from each of the selected fragments. Molecular cut-points are computationally selected according to the feasibility of reinstating the fragment disconnection by well-explored synthetic organic chemistry methodologies.[26]

Herein, we describe fully our computational methodology for the directed design and selection of novel compounds. In developing MOARF, our aim was to enable the rapid and objective *in silico* generation of synthetically plausible solution molecules. The solution molecules were additionally required to be within the scope of medicinal chemistry program objectives to provide additional ideas to maximize the potential solution space available to medicinal chemists. We illustrate the prospective application of this methodology to the synthesis and experimental testing of the designed compounds of relevance to a historical in-house drug discovery program. These compounds are derived from the exemplar small molecule cyclin-dependent kinase 2 (CDK2) inhibitor, seliciclib (CYC202, *R*-roscovitine, **1**, Figure 1) currently in Phase II clinical trials for the treatment of various malignancies.[27−29] Seliciclib has suboptimal human microsomal stability to oxidative metabolism and research from our laboratory to improve this feature has previously been reported.[30] Consistent with the context in which we developed MOARF, we apply it to explore synthetically accessible chemical space while maintaining the core purine scaffold of **1** and limiting solutions within physicochemical property space likely to engender improved metabolic stability. Solutions that lie within desired physicochemical property space, and that are predicted to have CDK2 binding affinity, are selected in iterative cycles of multiobjective computational design. Synthesis and experimental evaluation of 14 MOARF-generated optimal solution molecules demonstrates CDK2 biochemical activity and improved human metabolic stability versus **1**.



(1)

Figure 1. Seliciclib (CYC202, R-roscovitine).

METHODS

Multiobjective Algorithm for Replacement of Fragments (MOARF). MOARF is a multiobjective evolutionary algorithm constructed from a modular workflow (Figure 2). Individual component parameters may be selected to suit the optimization objective under study. Input to the workflow is a parameter file that contains the simplified molecular-input line-entry system (SMILES) annotation of the starting molecule

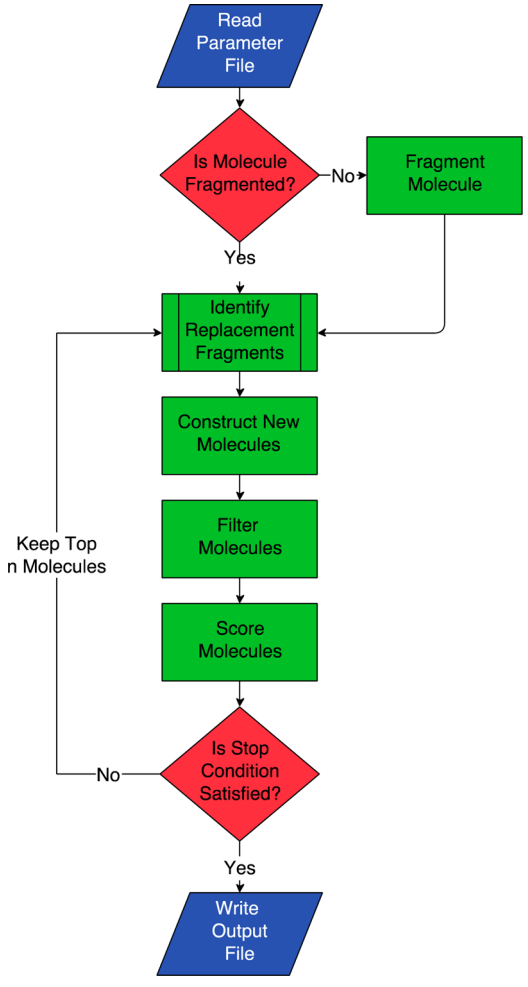
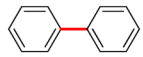
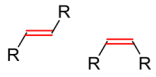
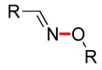
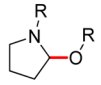
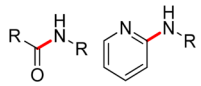
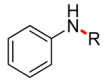
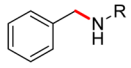


Figure 2. Outline of the MOARF workflow. Fragmentation performed by SynDiR. Rapid Alignment of Topological Scaffolds (RATS) is included in the "identify replacement fragments" component.

DOI: 10.1021/acs.jcim.5b00073
J. Chem. Inf. Model. 2015, 55, 1169−1180

Table 1. SynDiR Rules in Descending Priority Order[a]

Priority	Cut-point rule	Example disconnection
1	Biaryl, aryl-heteroaryl or heteroaryl-heteroaryl bond	
2	*E*- or *Z*-Alkene bond	
3	Carbon-iodine bond on an aromatic system	
4	Bonds between any two heteroatoms in an acyclic system	
5	Glycosidic linkages	
6	Bonds between a heteroatom and an unsaturated system containing a heteroatom in the alpha position	
7	Alkyl-heteroatom bonds	
8	Bonds between a benzylic carbon and a heteroatom	
9	Exocyclic bonds from nitrogen in a cyclic system	
10	Enolic bonds	

[a]No bond in a cyclic system is disconnected.

and the set of parameters to be optimized. If desired, one or more fragments in the starting molecule can remain unaltered throughout the workflow. The parameter file is read and the molecule is either passed into an objective, rule-based structure fragmenter (see SynDiR below) or, alternatively, may be subject to user-defined fragmentation of the chemical structure. Python modules, implemented using RDKit,[31] are used to integrate the components (Figure 2). Fingerprints are calculated in RDKit and are named according to literature conventions.[32] Machine learning is implemented using the scikit-learn API.[33]

Workflow Component: Synthetic Disconnection Rules (SynDiR). The fragmentation component SynDiR is used to deconstruct a query molecule into chemically relevant fragments for use in DND. In addition, SynDiR is applied to data sets of available small molecules to construct a library of synthetically accessible fragments for use in the fragment-replacement algorithm (see below). SynDiR applies an ordered set of rules, each of which corresponds to a plausible retrosynthetic disconnection (Table 1) and is substructure encoded as a smiles arbitrary target specification (SMARTS) query. All SMARTS matches within an input molecule are returned and, for each match, an attempt to disconnect a bond is made in order of priority (Table 1). Prioritization of the cut-point rules takes into account the synthetic tractability of the generated fragments and of the synthetic process that describes the reverse of the disconnection (Table 1).[26] Sequential, prioritized application of the cut-point rules minimizes the generation of isolated heteroatoms and, in addition, an overarching rule forbids disconnections that open a ring system or lead to an isolated heteroatom (unless the generated atom is iodine, Rule 3); this approach maintains existing ring systems and avoids the generation of very small fragments. When a cut is made, the positions of each cut-point are tagged with a dummy atom. The fragmenter function and set of rules is available in the Supporting Information (File 1 in the zip file) as a Python function using the RDKit API.[31]

Data Sets Used for Synthetic Fragment Generation. The following seven compound data sets were subject to the SynDiR disconnection rules (Table 1) to generate a library of synthetic fragments. Compound data sets were chosen to provide large coverage of synthetic chemistry space.

Sigma-Aldrich:[34] ~5.7 million Commercially available, unique structures from the Sigma-Aldrich Market Select database.

ICR Lead-Like Screening Collection:[35] ~75 000 compounds from The Institute of Cancer Research (ICR) in-house screening collection. This library includes compounds selected from commercial vendors and compounds synthesized in-house.

ICR Small Molecule Fragment Screening Library:[36] 2465 fragment-like molecules purchased from vendors.

ChEMBLdb V.15:[37] ~1.2 million compounds curated by the EBI-ChEMBL team from the medicinal chemistry literature.

DOI: 10.1021/acs.jcim.5b00073
J. Chem. Inf. Model. 2015, 55, 1169−1180

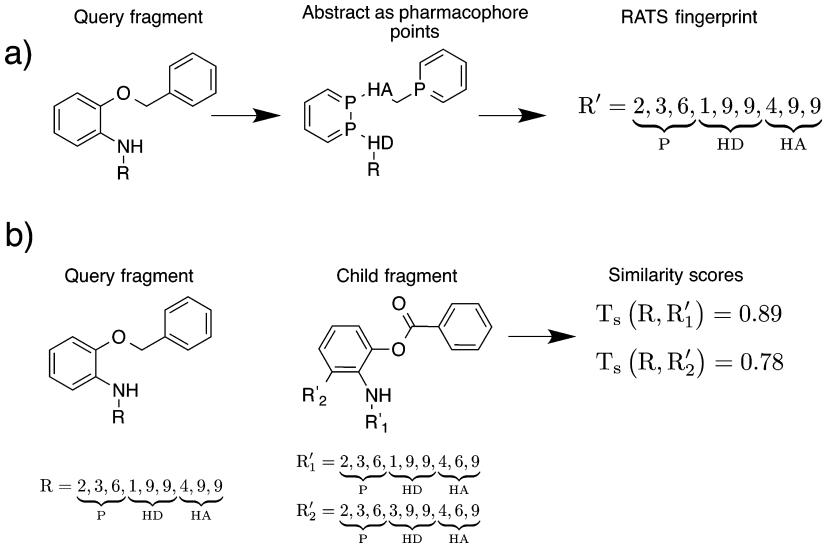


Figure 3. (a) Exemplar fingerprint generation: conversion of a two-dimensional structure to a pharmacophore graph, then conversion of this graph to a fingerprint for the described exit vector (R). The pharmacophores used are hydrogen bond donor (HD), hydrogen bond acceptor (HA), and atoms with three or more heavy atom connections (P). This exemplar molecule serves to illustrate the fingerprint generation concept; however, we recognize that it would be further disconnected by the cut-point rules of Table 1. (b) Exemplar exit vector selection: query and child fragments have fingerprints calculated for each exit vector and similarity scores are calculated for each exit vector in the child fragment. In this example R'_1 would be selected as it is the most similar to R.

BioFocus Kinase Focus Library:[38] 10 000 compounds designed to inhibit protein kinases.

eMolecules:[39] ~5.2 million commercially available unique structures from the eMolecules database.

Maybridge Screening Library:[40] 60 000 lead-like compounds and rule-of-three compliant small molecules.

Preparation of the Synthetic Fragment Database. The seven compound data sets described above were combined and all molecules which contained atoms other than H, C, N, O, S, F, Cl, Br, and I were removed. The Pipeline Pilot 8.0[41] canonical tautomer was selected and all duplicate molecules were removed. Each molecule was then fragmented, and the cut-point locations annotated. Identical fragments, including their respective cut-points, were merged to give 880 273 unique fragments. The frequency of each exit vector pattern (i.e., the relative arrangement of exit vectors on the fragment structure) was calculated, these patterns are termed "child fragments". The cut-points were then removed to give a set of 169 212 unique "parent fragments". Physicochemical properties and ECFC4[42] and CATS10[43] (implemented in RDKit) fingerprints were calculated for each parent fragment. All child and parent fragments, cut-point annotations, properties, and fingerprint information were written to a PostgreSQL[44] database.

Workflow Component: Identify Replacement Fragments. For each query fragment, identified automatically or manually from the input query molecule, a set of similar parent fragments are selected as potential replacements by first selecting a manageable subset ($n < 5000$) of the database as defined by molecular weight, TPSA, number of rings, and, if required, iterative similarity thresholds. Molecular weight, TPSA, and number of rings are applied by defining a property range around the query fragment; for instance, molecular weight of the query fragment ±25 Da as a default parameter. This filtering process has been implemented due to the computational strain of evaluating large numbers of molecules in multiple dimensions when using methods such as Pareto ranking. Also, as the results are cached in memory to speed up the fragment replacement process for subsequent generations, these similarity thresholds prevent MOARF from overloading memory. The number of parent fragments returned from the initial physicochemical descriptor-based database query varies depending on the query molecule. For instance, a purine ring returns 13 500 parent fragments; however, an isopropyl returns 197 parent fragments. The desired number of fragment replacements, as entered in the parameter file (see above), is then selected by performing a fingerprint similarity-driven virtual screen. This virtual screen is performed by identifying the most similar fragments as defined by a combination of precalculated properties, such as structural and pharmacophoric fingerprints (see Results and Discussion section).

Workflow Component: Rapid Alignment Search (RATS). For each parent fragment identified as a potential replacement, a child fragment is selected. Selection of the child fragment is achieved using a stochastic method based upon both the number of available substitution points of the parent fragment (exit vectors) and the frequency of occurrence for each of the substituted child fragments in the synthetic fragment database (i.e., how many times a child fragment results from disconnection of the complete data set). Let R_0 be the number of exit vectors in the query fragment, R_i, the number of exit vectors for each of the i children in the set, and f_i, the frequency of occurrence for each child fragment. Then we set R'_i to be the inverted distance function given by

$$R'_i = |\max(\{R_i\}) - R_0| - |R_i - R_0|$$

To scale the distribution of distance functions in accordance with the distribution of frequency distributions, a feature normalization algorithm was applied to R'_i to give R''_i

$$R''_i = (b - a)(R'_i - c)/(d - c) + a$$

where a is the minimum of f_i, b is the maximum of f_i, c is the minimum of R'_i, and d is the maximum of R'_i. A fitness score is then created for each of the children which is defined as $P_i = f_i + R''_i$, giving a probability which combines both frequency and

DOI: 10.1021/acs.jcim.5b00073
J. Chem. Inf. Model. 2015, 55, 1169−1180

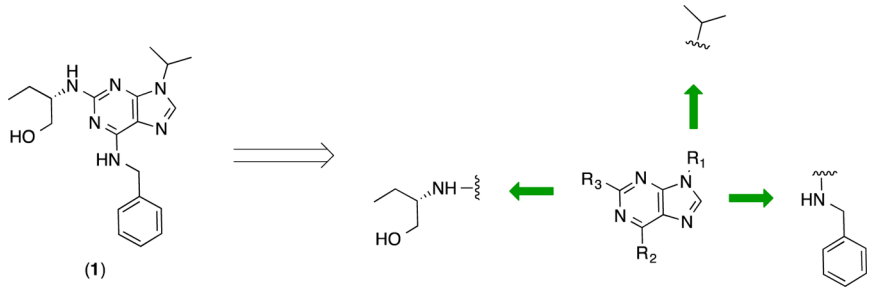


Figure 4. Fragmentation of **1**.

similarity in the number of exit vectors. Probabilistic selection is then performed using the roulette wheel algorithm.[45] If the selected fragment has fewer exit vectors than the query fragment then an exit vector is added to any atom with one or more hydrogens.

Once a child fragment is selected and potential points of substitution have been highlighted, a fragment alignment procedure is applied. Each exit vector is abstracted as a *k.n* fingerprint where *k* is the number of features and *n* is the number of occurrences of each feature included in the fingerprint. The fingerprint is computed by calculating the *n*-closest through graph distances, using the Floyd–Warshall algorithm, to each of the *k* features (Figure 3). If there are fewer than *n* features in the fragment, the remaining counts are set to the maximum path length. A similarity score is then calculated for a single alignment by summing the Tanimoto similarity of each of the individual Tanimoto scores between an exit vector in the child fragment and an exit vector in the query fragment. Once all of the possible alignments have been scored, the highest scored alignment is used. This two-stage fragment selection process has been designed to allow any of the fragments within the database to be selected as well as any possible substitution pattern. However, the virtual screen to select child fragments and the probabilistic selection of child fragments is designed to favor the most relevant replacements.

Workflow Component: Constructing and Filtering. Once each set of replacement fragments has been selected and aligned, molecules are reconstructed, by changing a single fragment within the molecule, to give an intermediate population of PS·NF·NR solutions, where PS is the population size which is carried through from generation *n* to *n* + 1, NF is the number of fragments in the original molecule, and NR is the user-defined number of fragment replacements to be made. By changing a single fragment at a time MOARF can perform a more exhaustive search of chemistry space. Before scoring is performed, this population is filtered by removing molecules with undesirable substructures or whose physicochemical properties lie outside acceptable ranges as defined by the user in the parameter file. Duplicate solutions are also removed.

Workflow Component: Scoring and Ranking. The final part of the workflow is scoring and ranking the population. Since there is a large variety of methods available, and each of these is highly customizable, we recommend that bespoke scoring functions are selected for each problem. In the example presented here, predicted CDK2 biochemical potency was modeled using the ROCS[46] Tanimoto Combo score (see exemplar optimization experiment below), Atom Pair fingerprint[47] similarity, and a statistical classifier (see bioactivity modeling below). These scores were then fused along with ClogP values using zScores.[48] This data fusion method is performed by transforming each raw score to the standard score and then averaging either some or all of the highest standard scores. Once the population of solutions has been ranked, the top PS molecules are selected to go through to the next iteration. This process is iterated until either an automated termination criterion is met or, if under inspection the user is satisfied with the results, a manual termination can be performed.

Bioactivity Modeling. To assess the ability of MOARF to design active compounds, a statistical classifier was built from in-house and literature experimental biochemical activity data for small molecule ligands versus CDK2. This data set consists of measured IC_{50} values against purified human CDK2 for 196 compounds substituted at the C2 and/or C6 positions of the purine scaffold (Figure 4). Compounds in this data set originated as follows ChEMBL v.17 (n = 24), ICR historical project compounds (n = 52) and 118 compounds purchased from Sigma-Aldrich and tested in-house for activity against CDK2 using a previously reported assay.[49] These data were defined as active or inactive using a 10 μM cutoff[50] and then split randomly into training and test sets using a 7:3 ratio. ECFC4 fingerprints, of length 1024, were calculated and a Random Forest (RF) model, using 500 trees and default scikit-learn parameters, was trained on the training set and validated using the test set (Table 2). After validation, models were retrained using all the data. An RF model was chosen as feature selection is performed implicitly. ECFC fingerprints were used due to their precedented application as descriptors to model structure–activity relationships.[42]

Table 2. Statistics for Random Forest Model Using 10 μM Cutoff

	inactive (training/test)	active (training/test)	total (training/test)
precision	1.00/0.91	1.00/0.94	1.00/0.92
recall	1.00/0.97	1.00/0.79	1.00/0.92
F1-score	1.00/0.94	1.00/0.86	1.00/0.91
support	89/40	48/19	137/59

Exemplar Optimization Experiment. Compound 1[27−29] was fragmented according to the SynDiR rules (Figure 4). The central purine scaffold and the *N*9-isopropyl moiety (R_1 in Figure 4) were retained, consistent with the scope explored in a previously reported medicinal chemistry program.[30] The two remaining R-groups (R_2 and R_3 in Figure 4) were replaced with methyl groups in the input query molecule to minimize bias from the initial structure and to maximize the potential for exploration of chemical space. This transformation will at first prevent RATS from providing optimal alignments; however as the scoring function prioritizes molecules, RATS will provide

DOI: 10.1021/acs.jcim.5b00073
J. Chem. Inf. Model. 2015, 55, 1169−1180

relevant alignments for these prioritized molecules. Ten parent fragments were selected for each fragment replacement, a number chosen to compromise between population size and scoring speed at the end of each generation. Parent fragments were evaluated using a weighted-sum of Tanimoto similarity using ECFC4 and CATS10; a final scoring term based on ClogP, where the highest score is given to the fragment with the lowest ClogP, was also added. The ClogP score was applied on every iteration to increase the probability of lower ClogP replacements and to be consistent with the aims of the previously exemplified medicinal chemistry program. A roulette wheel algorithm was also applied at each iteration to select potential replacement fragments with broad exploration of chemical space while also focusing on the best-scoring fragments. Child fragments were aligned using the RATS fingerprint-based algorithm by applying the feature list: R-groups, hydrogen bond donors, hydrogen bond acceptors, and aliphatic atoms. The number of occurrences counted in the RATS fingerprint was set to five. Once the child fragments had been aligned, the next generation of molecules was constructed. CDK2 biochemical potency for each of these solution molecules was predicted using the ligand-based methods, ROCS, atom pair fingerprints, and an RF classifier.

The three-dimensional conformation of **1** observed in PDB structure 3DDQ,[51] the highest resolution (1.80 Å) crystal structure of **1** in the kinase domain of CDK2 available at the time of our study, was used as a ROCS query. Ligands scored by ROCS were prepared using OMEGA[52] and default ROCS parameters were applied.

Atom pair fingerprints were calculated using default parameters in RDKit and similarity to **1** was assessed using the Dice similarity coefficient.[53] In addition activity prediction from the trained RF model was applied; molecules were scored either as zero or one corresponding to active and inactive, respectively. ClogP was calculated using the ChemAxon cxcalc tool,[54] with default parameters. The ClogP score was then normalized using the modified normal distribution, $f(x)$.

$$f(x) = \begin{cases} 2.89 - x, & \text{if } x > 2.89 \\ 2e^{-(x-1.44)^2/0.3}, & \text{otherwise} \end{cases}$$

Fusion of these four individual scores was performed using zScores. Although this method has been described for consensus modeling,[48] we used Z_3 in place of a normalized sum to allow the fusion of scoring functions with differing ranges. Ligands were then ranked by zScores and the top 25 molecules were progressed to the next generation of the MOARF algorithm. Scoring was distributed across 15 CPUs and made up 43% of the total run time of MOARF. Ten parent fragment replacements coupled with a population size of 25 gives an intermediate population size of 500 molecules, with redundancy. In the experiments reported here, MOARF was set to terminate after 100 generations.

RESULTS AND DISCUSSION

Development and Application of SynDiR. We developed SynDiR to break a query molecule into chemically relevant synthetic fragments for use in DND and to construct a library of synthetically accessible fragment building blocks for use in the fragment-replacement algorithm. SynDiR applies an ordered set of rules, each of which corresponds to a synthetically tractable class of disconnection. RECAP[52] is an alternative precedented and frequently used fragmentation method for the generation of fragment building blocks from small molecules.[20,56,57] In the original description of RECAP,[55] there is some ambiguity about when a fragment is too small for a disconnection to occur. This leads to inconsistent parametrizations in different implementations[31,58,59] each giving differing results.

We found no single implementation of RECAP that consistently gave synthetically relevant fragment building blocks from query molecules that we desired for the purpose of DND described here and therefore consider SynDiR to be more appropriate for application to the MOARF workflow. However, we were keen to compare SynDiR and RECAP in the context of the fragmentation of small molecules in compound libraries into their component fragments. With this in mind, we applied both SynDiR and RECAP, using the ChemAxon Fragment[59] software, to the BioFocus Kinase Focused Library and the ICR Lead-Like Screening Collection (see above) to compare the number of fragments generated and the commercial availability of the resultant fragments. RECAP settings for Fragment (Supporting Information, File 2 in the zip file)[59] were chosen as the closest comparison to the SynDiR method.

In this comparative study, more molecules remained uncut by RECAP in both data sets, and although more fragments were generated by SynDiR compared to RECAP, fewer unique fragments (after removal of duplicates) were generated by SynDiR (Tables 3 and 4). We hypothesize that these observations are likely due to the wider scope of the disconnection rules adopted in SynDiR; for example, the amide cut rule in SynDiR is part of a more generic disconnection (Rule 6, Table 1) compared to the correspond-

Table 3. Comparative Study of Fragments Generated by Applying SynDiR and RECAP to the BioFocus Kinase Focused Library[a]

	SynDiR	RECAP
no. uncut molecules	1	70
no. fragments generated	39921	35179
no. unique fragments generated	628	1414
average heavy atom count of unique fragments	9.3 ± 3	15.6 ± 6.1
no. unique fragments available in Sigma-Aldrich database	444 (71.7%)	392 (28.0%)

[a]Number of fragments generated is the total number for all molecules in the dataset including duplicate fragments. Removal of duplicates gives number of unique fragments generated.

Table 4. Comparative Study of Fragments Generated by Applying SynDiR and RECAP to the ICR Lead-Like Screening Collection[a]

	SynDiR	RECAP
no. uncut molecules	2947	5697
no. fragments generated	407475	387875
no. unique fragments generated	19116	31169
average heavy atom count of unique fragments	12.9 ± 4.2	14.7 ± 4.5
no. unique fragments available in Sigma-Aldrich database	6245 (32.7%)	8039 (26.0%)

[a]Number of fragments generated is the total number for all molecules in the dataset including duplicate fragments. Removal of duplicates gives number of unique fragments generated.

DOI: 10.1021/acs.jcim.5b00073
J. Chem. Inf. Model. 2015, 55, 1169−1180

Table 5. Alignment Results Comparing RATS Fingerprints to BROOD Default Parameters for Five Scaffolds Depicted in Figure 5[a]

drug	seliciclib (1)	atorvastatin (2)	glipizide (3)	dipyridamole (4)	sildenafil (5)
no. of exit vectors	3	3	2	3	3
no. of scaffold hops identified in BROOD	599	957	632	120	394
time taken to align using RATS (seconds)	6.3	11.8	6.5	1.5	3.8
no. of RATS-generated alignments which correspond to alignments generated in BROOD	483	698	618	120	304
RATS-generated alignments which correspond to alignments generated in BROOD (%)	80.6	69.0	98.3	100	77.2
average probability of random alignments corresponding to BROOD alignments (%)	19.5	16.7	50	33	17.4

[a]Average probability of alignments corresponding to BROOD alignments is calculated by averaging the probability of each scaffold hop random alignment corresponding to the BROOD alignment. The probability of a random scaffold hop corresponding is calculated by performing all alignments and calculating which percentage correspond to BROOD.

ing amide disconnection in RECAP. We also postulate that SynDiR generates less complex fragments which, as a result, collapse to a smaller number of unique fragment building blocks. To explore this further, we examined the difference in heavy atom count between unique fragments generated in SynDiR and RECAP.

SynDiR generated unique fragments with a statistically significant (p-value $< 2.2 \times 10^{-16}$) lower number of heavy atoms from both data sets compared to RECAP; however, the difference is small (less than two heavy atoms between the mean values). These findings indicate that the lower complexity of SynDiR-generated fragment building blocks is not driven by lower molecular weight. We also observed that SynDiR generates a higher percentage of unique parent fragments available as exact match structures in the Sigma-Aldrich catalogue (Tables 3 and 4) consistent with the notion that the SynDiR disconnection rules lead to synthetically relevant fragments useful for DND in the context of a medicinal chemistry program.

Rapid Alignment of Topological Scaffolds: RATS. Once potential replacement fragments have been identified using SynDiR, we were keen to consider all possible connection options between the selected replacement fragments, thereby broadening our exploration of potential chemical space and recognizing that the versatility of modern synthetic chemistry methodologies often allows multivector optimization of fragment scaffolds. With this goal in mind, we developed RATS search to analyze combinatorial connection options.

To the best of our knowledge, there has been no reported 2D fragment alignment algorithm for this use, however, BROOD[60] is a scaffold hopping software designed to replace selected scaffolds in a molecule with alternatives that have similar shape and pharmacophore, as described by color, combined with modified molecular properties.

We selected BROOD as a comparator methodology for RATS; we performed scaffold hops for five drug-like small molecules by manually selecting the scaffold to be replaced and using the default scaffold replacement library in BROOD (this default library is restricted to three or fewer exit vectors from a scaffold). We then applied the RATS fingerprint methodology to align potential scaffold replacements and compared the output with the corresponding BROOD default alignment.

We consider the BROOD-generated scaffold replacement to be an optimal literature-precedented alignment for a scaffold with n exit vectors; thus, to perform better than random we would expect RATS to have a greater than $(100/n!)$% consensus with BROOD. This threshold derived from considering the number of possible alignments of n exit vectors, in a nonsymmetric molecule with unique R-groups, as the permutation of n objects in a set ($n!$). Thus, for RATS to be better than random, we would expect to see agreement with BROOD more times than random, which is $(100/n!)$% of the time. The probability for a random alignment to correspond with a BROOD alignment is compound dependent as symmetry within the scaffold or having identical R-groups will increase the probability of a random alignment agreeing with BROOD, thus we give the average probability of a corresponding alignment for each data set (Table 5).

For the five exemplar molecules studied (Figure 5), Seliciclib along with four molecules selected from the top 200 brand name drugs in 2010,[61] we indeed observed higher agreement with BROOD than expected from random alignment of the scaffold replacement (Table 5).

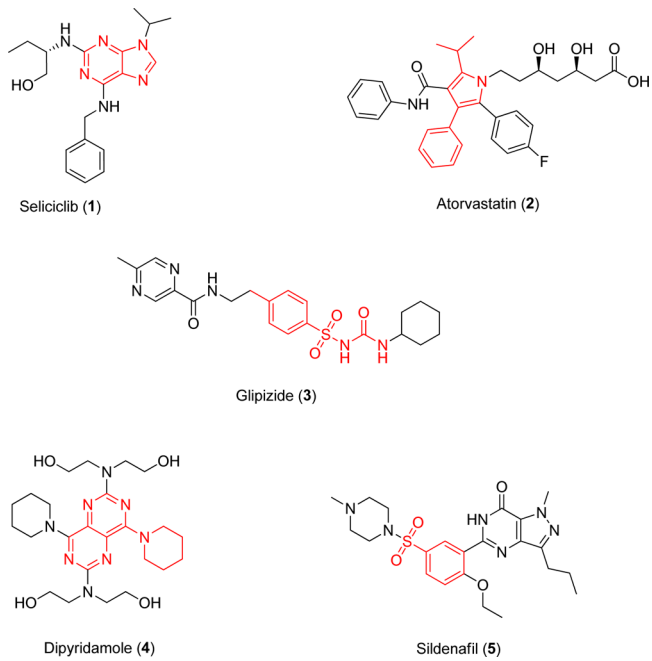


Figure 5. Molecules subject to BROOD-generated scaffold hop and comparative RATS scaffold replacement alignment. The scaffold identified for replacement is highlighted in red.

Multiobjective Optimization Case Study. We applied MOARF to a multiobjective optimization challenge encountered during a previously reported drug discovery project;

DOI: 10.1021/acs.jcim.5b00073
J. Chem. Inf. Model. 2015, 55, 1169−1180

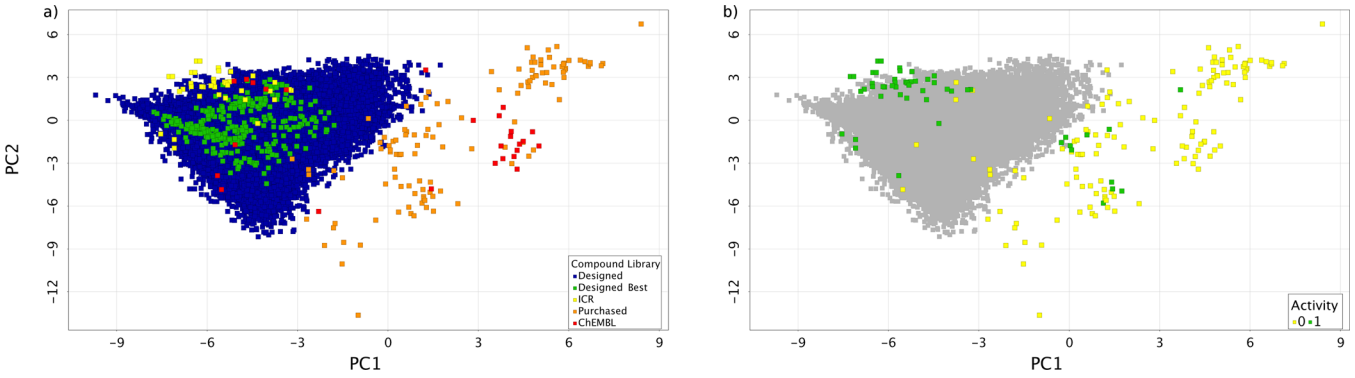


Figure 6. Principal component analysis of ECFC4 fingerprints for various compound data sets, including the 100 generations of MOARF compounds. The PCA model was built using only synthesized molecules with measured IC_{50} data versus CDK2. (b) Actives (green) are defined as measured $IC_{50} < 10\ \mu M$, and inactives (yellow) are defined as measured $IC_{50} > 10\ \mu M$. The variance explained by PC1 is 21% and that explained by PC2 is 15%.

namely to reduce the susceptibility of **1** to oxidative metabolism while retaining potency against the primary biochemical target, cyclin-dependent kinase CDK2.[30]

As described above (see Methods section), **1** was fragmented according to the SynDiR rules while retaining the central purine scaffold and *N*9-isopropyl moiety, consistent with the scope of the previously reported medicinal chemistry program (Figure 4).[30] The MOARF DND workflow, including RATS alignment, was applied to the two remaining R-groups (R_2 and R_3 in Figure 4) with the query molecule defined by R_2 and R_3 = Me. Scoring of each solution used a set of ligand-based methods (ROCS, atom pair similarity, and predicted activity using an RF model), combined with a ClogP desirability function for each generated molecule. The top 25 solutions were progressed to the next iteration of the MOARF algorithm which was set to terminate after 100 generations. After the 93rd generation no new molecules were included in the top 25 solutions suggesting that MOARF had converged to a global optimum.

Importantly, and in line with expectation, we observed a broad exploration of chemical space using MOARF (Figure 6). MOARF-designed compounds (43 844) have a large and dense coverage of an area of chemical space which spans inside the applicability domain of the trained PCA model built using compounds experimentally tested against CDK2, similar to those described in the above Bioactivity Modeling section. The number of fragment replacements made at R_2 and R_3 (Figure 4) were 17 593 and 12 110, respectively, giving a virtual space of over 210 million compounds.

MOARF-designed compounds are dissimilar to those found in the ChEMBL and Sigma-Aldrich (purchased) data sets (Figure 6a). We propose that this is due to both the physicochemical constraints applied to the MOARF-designed compounds and also that replacements were of similar size to the original fragments whereas compounds in the ChEMBL and Sigma-Aldrich data sets may bear larger substituents and substitution at additional exit vectors. Interestingly, we observed that the MOARF-designed compounds are distal to the area of the PCA space which has a high ratio of inactives to actives (Figure 6b). An enumerated library of molecules containing the purine and isopropyl motif, but with randomly selected single exit vector child fragments from the MOARF database, has also been projected onto the same PCA-space (Supplementary Figure S1). Though these compounds have a vast coverage of chemical space, they show no preference for the "active" areas of the PCA-space (Figure 6b) and can be seen to lie outside of the applicability domain.

To experimentally validate our method, we synthesized 14 exemplar compounds from the designed best top 25 compound set obtained after 100 DND generations. Compounds were prepared according to our previously published synthetic method (See Supporting Information for synthesis and characterization of all compounds).[30] This set of 14 compounds was selected for synthesis based upon the availability of synthetic building blocks; for some designed best top 25 compounds, building blocks were no longer commercially available or were prohibitively expensive.

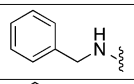
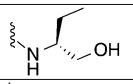
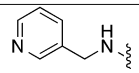
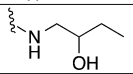
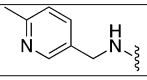
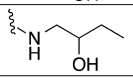
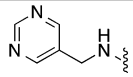
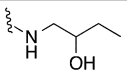
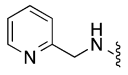
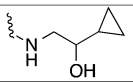
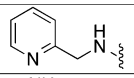
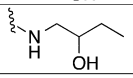
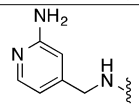
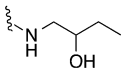
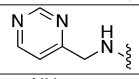
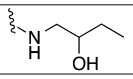
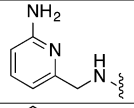
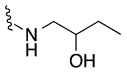
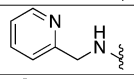
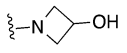
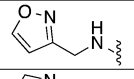
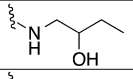
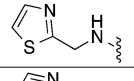
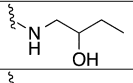
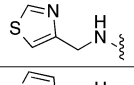
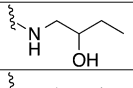
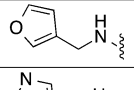
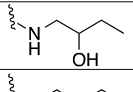
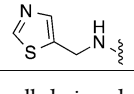
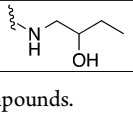
All 14 compounds were evaluated experimentally for activity in a CDK2 biochemical assay and for metabolic stability in a human microsomal metabolism preparation side-by-side with **1** (Table 6). All synthesized compounds demonstrated activity in the biochemical CDK2 assay with **6**, **7**, and **8** retaining activity within 5-fold of **1** (Table 6, Entries 1−4). All compounds, with the exception of **18**, demonstrate improved metabolic stability consistent with the lower ClogP range of this optimal set of designed molecules. Interestingly, all 14 exemplars contain 5- or 6-membered heterocyclic replacements for the phenyl substituent at C6 of the purine scaffold and it is likely that these replacements contribute to improved metabolic stability. Indeed a 3-pyridyl substituent at this position, present in solutions **6** and **7** (Table 6, Entries 2 and 3), has previously been reported to improve metabolic stability in CCT68127.[30] Compound **18**, that does not display improved metabolic stability, bears a furan heterocycle with a precedented liability for CYP450-mediated metabolism unrelated to its overall physicochemical properties.[62]

Also of note, all 14 solutions replace the primary alcohol of the purine C2 substituent with a secondary alcohol which may also contribute to improved stability through reduced propensity to oxidation.[30] Notably, the preferred 1-amino-butane-2-ol moiety at C2 of the purine scaffold (12 out of 14 compounds) is present in the most potent and stable examples **6**, **7**, and **8** (Entries 2−4, Table 6).

In summary, starting from a fragment-like query we demonstrate application of MOARF, incorporating an RF model of CDK2 activity coupled with physicochemical property space restrictions (ClogP) to optimize toward potent and metabolically stable analogues of **1**. In the course of this optimization 43 844 virtual molecules were generated in 100 generations of DND with 14 of the "designed best" last

DOI: 10.1021/acs.jcim.5b00073
J. Chem. Inf. Model. 2015, 55, 1169−1180

Table 6. CDK2 Biochemical Potency and Human Microsomal Stability for 14 Compounds Selected from the Designed Best Cohort in Comparison with 1[a]

Entry No.	Compd No	R_2	R_3	IC_{50} (μM)	HLM (% turn over)	zScore
1	**1**			0.128 ± 0.075	51.2 ± 8.6	N.C.
2	**6**			0.468 ± 0.047	24.6 ± 8.5	1.80
3	**7**			0.384 ± 0.050	15.8 ± 2.8	1.35
4	**8**			0.35 ± 0.110	1.2 ± 4.0	1.30
5	**9**			2.49 ± 0.322	4.0 ± 4.0	1.37
6	**10**			2.13 ± 0.466	4.9 ± 3.6	1.72
7	**11**			1.74 ± 0.267	18.4 ± 2.0	1.42
8	**12**			6.26 ± 0.041	14.2 ± 6.0	1.38
9	**13**			3.70 ± 0.190	14.8 ± 0.4	1.40
10	**14**			41.46 ± 22.4	6.1 ± 4.2	1.41
11	**15**			1.36 ±0.154	13.6 ± 0.0	1.56
12	**16**			2.22 ± 0.457	7.1 ± 2.7	1.53
13	**17**			2.47 ± 0.143	9.3 ± 0.0	1.52
14	**18**			1.03 ± 0.140	44.2 ± 10.7	1.36
15	**19**			0.65 ± 0.070	17.3 ± 0.8	1.49

[a]Data fusion scores (zScore) is also included for all designed compounds.

generation of synthesized compounds demonstrating CDK2 biochemical activity and improved human microsomal stability.

CONCLUSIONS

We have reported the development and application of a multiobjective optimization workflow (MOARF) with the intention of objectively broadening the exploration of potential chemical space in a medicinal chemistry program while simultaneously incorporating desirable physicochemical property design features. The MOARF system is highly extensible to new challenges in a drug design project with the ability to readily incorporate other computational methodologies in the optimization, such as pharmacophore modeling and virtual ligand docking.

To generate synthetically relevant molecular fragments, we developed a rule-based molecular fragmentation scheme (SynDiR), which we used to generate a large and diverse library of such fragments, annotated with cut-points, as potential replacements. We have demonstrated that SynDiR compares favorably with a widely used retrosynthetic fragmentation methodology (RECAP) in its ability to generate commercially available synthetic fragments from large molecule libraries. The SynDiR-generated database of fragments contains information on both the parent fragments and their children that explore the exemplified connection points and their relative frequency of occurrence.

To maximize the use of cut-point information contained within SynDiR, we also developed a pharmacophore fingerprint-based fragment replacement algorithm (RATS) based only on topology. RATS broadens the scope of reconnection options considered in molecule reconstruction and was validated and found comparable to a leading geometric bioisosteric replacement tool, BROOD.

DOI: 10.1021/acs.jcim.5b00073
J. Chem. Inf. Model. 2015, 55, 1169−1180

To demonstrate that MOARF and the new components can be combined, we developed a computational method that integrates the SynDiR-derived synthetic fragment library, RATS fragment-replacement and alignment algorithm and a multi-objective-scoring algorithm for ranking candidate solution molecules that comprises biochemical activity predictions and physicochemical property calculations. Application of this integrated and iterative workflow to the optimization of the exemplar small molecule CDK2 inhibitor **1** generated a set of candidate solutions, from a fragment-like query molecule, that occupy chemical space previously unexplored in terms of chemical structure and physicochemical properties (ClogP) in the context of a historical medicinal chemistry program.

We have shown that MOARF allows for the rapid exploration and exploitation of a vast (ca. 200 M virtual molecules) synthetically accessible chemical space using highly relevant building blocks that are likely to be commercially available. In this example three objectives for optimization were considered: ligand-based shape similarity to a known ligand of interest; RF-based biochemical activity prediction; and the restriction of physicochemical property space (ClogP). A prospective study was conducted in which 14 of the top 25 solutions were synthesized and tested. Three of these compounds retained biochemical activity within 5-fold of the query molecule and 13 of the 14 solutions demonstrated improved metabolic stability. This study demonstrates the prospective application and validation of MOARF to a relevant medicinal chemistry challenge to improve metabolic stability while maintaining biochemical potency.

There are strong economic and practical drivers for medicinal chemists and drug design teams to fully explore and exploit relevant chemical space with the synthesis of a minimum number of molecules. In addition, there is increasing understanding of the molecular properties likely to deliver molecules with good pharmacokinetic, pharmaceutical and safety profiles. A potentially powerful attribute of the fully modular and extensible MOARF workflow is the opportunity for user-defined parametrization and inclusion of additional computational objectives to direct *de novo* design as a medicinal chemistry program develops and new challenges arise.

■ ASSOCIATED CONTENT

Supporting Information

Table S1, Figure S1, and preparation of compounds **6**−**19**. Experimental protocols for CDK2 kinase activity assay and human microsomal stability assay. ^{1}H NMR spectra for compounds **6**−**19**, the fragmenter function and set of rules (File 1), and RECAP settings for Fragment (File 2) are also provided. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00073.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: julian.blagg@icr.ac.uk. Telephone: +44 (0) 20 8722 4051 (J.B.).

*E-mail: nathan.brown@icr.ac.uk. Telephone: +44 (0) 20 8722 4033 (N.B.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

N.C.F. is funded by the Institute of Cancer Research, N.B. and J.B. are funded by Cancer Research UK Grant No. C309/A8274. N.C.F. would like to thank Kathy Boxall for her help with running the CDK2 assays, Angela Hayes and Jennie Roberts for performing the HLM assays, and Yi Mok for helpful comments on the manuscript. We thank Dr. Amin Mirza, Mr. Meirion Richards and Dr. Maggie Liu for their assistance with NMR, mass spectrometry and HPLC.

■ REFERENCES

(1) Segall, M. D. Multi-Parameter Optimization: Identifying High Quality Compounds With a Balance of Properties. *Curr. Drug Metab.* **2012**, *18*, 1292−1310.

(2) Gillet, V. J.; Khatib, W.; Willett, P.; Fleming, P. J.; Green, D. V. Combinatorial Library Design Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 375−385.

(3) Brown, N.; McKay, B.; Gasteiger, J. A Novel Workflow for the Inverse QSPR Problem Using Multiobjective Optimization. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 333−341.

(4) Leeson, P. D.; Oprea, T. I. Drug-Like Physicochemical Properties. In *Drug Design Strategies: Quantitative Approaches*; Livingstone, D. J., Davis, A. M., Eds.; RSC Drug Discovery Series; RSC Publishing: Cambridge, 2012; Vol. *13*, pp 35−59.

(5) Leeson, P. D.; Springthorpe, B. The Influence of Drug-Like Concepts on Decision-Making in Medicinal Chemistry. *Nat. Rev. Drug Discovery* **2007**, *6*, 881−890.

(6) Hopkins, A. L.; Keseru, G. M.; Leeson, P. D.; Rees, D. C.; Reynolds, C. H. The Role of Ligand Efficiency Metrics in Drug Discovery. *Nat. Rev. Drug Discovery* **2014**, *13*, 105−121.

(7) Merritt, A. High Throughput Chemistry in Drug Discovery; In *New Synthetic Technologies in Medicinal Chemistry*; Farrant, E., Ed.; RSC Drug Discovery Series; RSC Publishing: Cambridge, 2012; Vol. *11*, pp 6−41.

(8) O'Connell, K. M.; Galloway, W. R. J. D.; Spring, D. R. The Basics of Diversity-Oriented Synthesis. In *Diversity-Oriented Synthesis: Basics and Applications in Organic Synthesis, Drug Discovery, and Chemical Biology*, First ed.; Trabocchi, A., Schreiber, S. L., Eds.; John Wiley & Sons: New York, 2013; pp 1−26.

(9) Brown, N.; McKay, B.; Gilardoni, F.; Gasteiger, J. A Graph-Based Genetic Algorithm and its Application to the Multiobjective Evolution of Median Molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1079−1087.

(10) Ertl, P.; Lewis, R. IADE: a System for Intelligent Automatic Design of Bioisosteric Analogs. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 1207−1215.

(11) Ertl, P.; Lewis, R. Evaluation of a Semi-Automated Workflow for Fragment Growing. *J. Chem. Inf. Model.* **2015**, *55*, 180−193.

(12) Gillet, V. J.; Newell, W.; Mata, P.; Myatt, G.; Sike, S.; Zsoldos, Z.; Johnson, A. P. SPROUT: Recent Developments in the de novo Design of Molecules. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 207−217.

(13) Böhm, H.-J. The Computer Program LUDI: a New Method for the de novo Design of Enzyme Inhibitors. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61−78.

(14) Wang, R.; Gao, Y.; Lai, L. LigBuilder: A Multi-Purpose Program for Structure-Based Drug Design. *J. Mol. Model.* **2000**, *6*, 498−516.

(15) Nicolaou, C. A.; Joannis, A.; Costas, S. P. de novo Drug Design Using Multiobjective Evolutionary Graphs. *J. Chem. Inf. Model.* **2009**, *49*, 295−307.

(16) van der Horst, E.; Marqués-Gallego, P.; Mulder-Krieger, T.; van Veldhoven, J.; Kruisselbrink, J.; Aleman, A.; Emmerich, M. T. M.; Brussee, J.; Bender, A.; IJzerman, A. P. Multi-Objective Evolutionary Design of Adenosine Receptor Ligands. *J. Chem. Inf. Model.* **2012**, *52*, 1713−1721.

(17) Nicolaou, C. A.; Brown, N. Multi-Objective Optimization Methods in Drug Design. *Drug Disc. Today Technol.* **2013**, *10*, 427−435.

DOI: 10.1021/acs.jcim.5b00073
J. Chem. Inf. Model. 2015, 55, 1169−1180

(18) Huang, Q.; Li, L.-L.; Yang, S.-Y. PhDD: A New Pharmacophore-Based *de novo* Design Method of Drug-Like Molecules Combined With Assessment of Synthetic Accessibility. *J. Mol. Graph. Model.* **2010**, *28*, 775−787.
(19) 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*, e1002380.
(20) Gillet, V. J.; Bodkin, M. J.; Hristozov, D. Multiobjective de novo Design of Synthetically Accessible Compounds. In *de novo Molecular Design*; Schneider, G., Ed.; Wiley-VCH: Weinheim, 2013; pp 267−285.
(21) Kawai, K.; Nagata, N.; Takahashi, Y. de novo Design of Drug-Like Molecules by a Fragment-Based Molecular Evolutionary Approach. *J. Chem. Inf. Model.* **2013**, *54*, 49−56.
(22) Kumar, A.; Voet, A.; Zhang, K. Fragment Based Drug Design: From Experimental to Computational Approaches. *Curr. Med. Chem.* **2012**, *19*, 5128−5147.
(23) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.
(24) Kerns, E.; Di, L. *Drug-like properties: concepts, structure design and methods: from ADME to toxicity optimization*; Academic Press: Waltham, 2008.
(25) Hughes, J. D.; Blagg, J.; Price, D. A.; Bailey, S.; DeCrescenzo, G. A.; Devraj, R. V.; Ellsworth, E.; Fobian, Y. M.; Gibbs, M. E.; Gilles, R. W.; Greene, N.; Huang, E.; Krieger-Burke, T.; Loesel, J.; Wager, T.; Whiteley, L.; Zhang, Y. Physiochemical Drug Properties Associated With in vivo Toxicological Outcomes. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 4872−4875.
(26) 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*, 3451−3479.
(27) Cyclacel http://www.cyclacel.com/research_programs_oncology_cyc202.shtml (accessed September 13, 2014).
(28) Krystof, V.; Uldrijan, S. Cyclin-Dependent Kinase Inhibitors as Anticancer Drugs. *Curr. Drug Targets* **2010**, *11*, 291−302.
(29) Aldoss, I. T.; Tashi, T.; Ganti, A. K. Seliciclib in Malignancies. *Expert Opin. Investig. Drugs* **2009**, *18*, 1957−1965.
(30) Wilson, S. C.; Atrash, B.; Barlow, C.; Eccles, S.; Fischer, P. M.; Hayes, A.; Kelland, L.; Jackson, W.; Jarman, M.; Mirza, A.; Moreno, J.; Nutley, B. P.; Raynaud, F. I.; Sheldrake, P.; Walton, M.; Westwood, R.; Whittaker, S.; Workman, P.; McDonald, E. Design, Synthesis and Biological Evaluation of 6-pyridylmethylaminopurines as CDK Inhibitors. *Bioorg. Med. Chem.* **2011**, *19*, 6949−6965.
(31) Landrum, G. RDKit: Open-source cheminformatics. http://www.rdkit.org (accessed September 13, 2014).
(32) Riniker, S.; Landrum, G. A. Open-Source Platform to Benchmark Fingerprints for Ligand-Based Virtual Screening. *J. Cheminf.* **2013**, *5*, 26.
(33) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(34) Sigma Aldrich Market Select. http://www.aldrichmarketselect.com (accessed September 13, 2014).
(35) Langdon, S. R.; Brown, N.; Blagg, J. Scaffold Diversity of Exemplified Medicinal Chemistry Space. *J. Chem. Inf. Model.* **2011**, *51*, 2174−2185.
(36) Silva-Santisteban, M. C.; Westwood, I. M.; Boxall, K.; Brown, N.; Peacock, S.; McAndrew, C.; Barrie, E.; Richards, M.; Mirza, A.; Oliver, A. W.; Burke, R.; Hoelder, S.; Jones, K.; Aherne, G. W.; Blagg, J.; Collins, I.; Garrett, M. D.; van Montfort, R. L. M. Fragment-Based Screening Maps Inhibitor Interactions in the ATP-Binding Site of Checkpoint Kinase 2. *PLoS One* **2013**, *8*, e65689.
(37) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.
(38) BioFocus. http://www.biofocus.com/offerings/compound-libraries/kinase.htm (accessed March 15, 2012).
(39) eMolecules. http://www.emolecules.com (accessed September 13, 2014).
(40) Maybridge. http://www.maybridge.com (accessed March 15, 2012).
(41) *PipelinePilot*, version 8.0; Accelrys: SanDiego, CA, USA; http://accelrys.com/products/pipeline-pilot (accessed September 13, 2014).
(42) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(43) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. Scaffold-Hopping by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894−289.
(44) The PostgreSQL Global Development Group. *PostgreSQL*, version 9.2.2; http://www.postgresql.org (accessed September 13, 2014).
(45) Goldberg, D. E.; Holland, J. H. Genetic Algorithms and Machine Learning. *Mach. Learn.* **1988**, *3*, 95−99.
(46) Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74−82.
(47) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(48) Sastry, G. M.; Inakollu, V. S.; Sherman, W. Boosting Virtual Screening Enrichments With Data Fusion: Coalescing Hits From 2D Fingerprints, Shape, and Docking. *J. Chem. Inf. Model.* **2013**, *53*, 1531−1542.
(49) Naud, S.; Westwood, I. M.; Faisal, A.; Sheldrake, P.; Bavetsias, V.; Atrash, B.; Cheung, K.-M. J.; Liu, M.; Hayes, A.; Schmitt, J.; Wood, A.; Choi, V.; Boxall, K.; Mak, G.; Gurden, M.; Valenti, M.; de Haven Brandon, A.; Henley, A.; Baker, R.; McAndrew, C.; Matijssen, B.; Burke, R.; Hoelder, S.; Eccles, S. A.; Raynaud, F. I.; Linardopoulos, S.; van Montfort, R. L. M.; Blagg, J. Structure-Based Design of Orally Bioavailable 1H-Pyrrolo[3,2-c]pyridine Inhibitors of Mitotic Kinase Monopolar Spindle 1 (MPS1). *J. Med. Chem.* **2013**, *56*, 10045−10065.
(50) Mok, N. Y.; Brenk, R. Mining the ChEMBL Database: an Efficient Chemoinformatics Workflow for Assembling an Ion Channel-Focused Screening Library. *J. Chem. Inf. Model.* **2011**, *51*, 2449−2454.
(51) Bettayeb, K.; Oumata, N.; Echalier, A.; Ferandin, Y.; Endicott, J.; Galons, H.; Meijer, L. CR8, a Potent and Selective, Roscovitine-Derived Inhibitor of Cyclin-Dependent Kinases. *Oncogene* **2008**, *27*, 5797−5807.
(52) Kirchmair, J.; Wolber, G.; Laggner, C.; Langer, T. Comparative Performance Assessment of the Conformational Model Generators OMEGA and Catalyst: a Large-Scale Survey on the Retrieval of Protein-Bound Ligand Conformations. *J. Chem. Inf. Model.* **2006**, *46*, 1848−1861.
(53) Dice, L. R. Measures of the Amount of Ecologic Association Between Species. *Ecology.* **1945**, *26*, 297−302.
(54) *cxcalc*, version 5.10.3; ChemAxon Ltd;https://www.chemaxon.com/marvin-archive/5_2_0/marvin/help/applications/calc.html (accessed September 13, 2014).
(55) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP − Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511−522.
(56) Firth, N. C.; Brown, N.; Blagg, J. Plane of Best Fit: A Novel Method to Characterize the Three-Dimensionality of Molecules. *J. Chem. Inf. Model.* **2012**, *52*, 2516−2525.
(57) de León, A. V.; Bajorath, J. Matched Molecular Pairs Derived by Retrosynthetic Fragmentation. *Med. Chem. Comm.* **2014**, *5*, 64−67.
(58) *MOE*; Chemical Computing Group, Montreal, Quebec, Canada; http://www.chemcomp.com (accessed September 13, 2014).

DOI: 10.1021/acs.jcim.5b00073
J. Chem. Inf. Model. 2015, 55, 1169−1180

(59) *Fragmenter*, 5.10.3; ChemAxon Ltd; https://www.chemaxon.com/jchem/examples/fragmenter/index.html (accessed September 13, 2014).
(60) *BROOD*, version 1.7.2; OpenEye Scientific Software, Inc.: Santa Fe, NM, USA; www.eyesopen.com (accessed September 13, 2014).
(61) McGrath, N. A.; Brichacek, M.; Njardarson, J. T. A Graphical Journey of Innovative Organic Architectures That Have Improved Our Lives. *J. Chem. Educ.* **2010**, *87*, 1348−1349.
(62) Dalvie, D. K.; Kalgutkar, A. S.; Khojasteh-Bakht, S. C.; Obach, R. S.; O'Donnell, J. P. Biotransformation Reactions of Five-Membered Aromatic Heterocyclic Rings. *Chem. Res. Toxicol.* **2002**, *15*, 269−299.

DOI: 10.1021/acs.jcim.5b00073
J. Chem. Inf. Model. 2015, 55, 1169−1180