JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

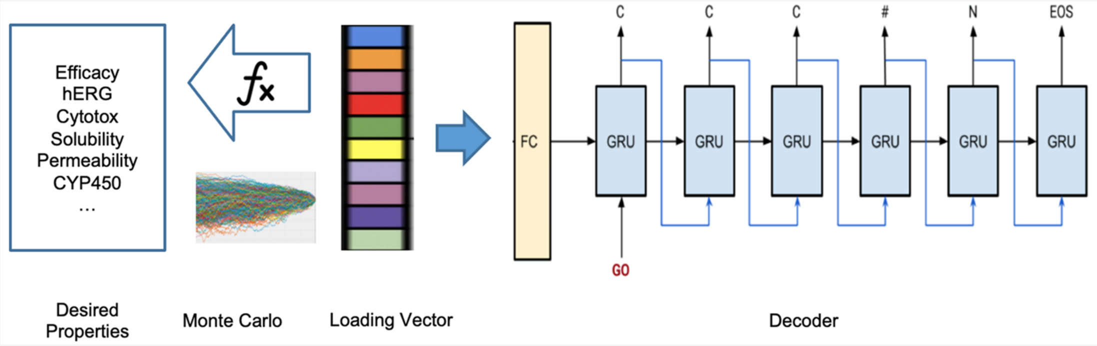
pubs.acs.org/jcim Article

Retro Drug Design: From Target Properties to Molecular Structures

Yuhong Wang,* Sam Michael, Shyh-Ming Yang, Ruili Huang, Kennie Cruz-Gutierrez, Yaqing Zhang, Jinghua Zhao, Menghang Xia, Paul Shinn, and Hongmao Sun*

Cite This: *J. Chem. Inf. Model.* 2022, 62, 2659−2669 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

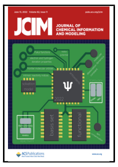


ABSTRACT: To deliver more therapeutics to more patients more quickly and economically is the ultimate goal of pharmaceutical researchers. The advent and rapid development of artificial intelligence (AI), in combination with other powerful computational methods in drug discovery, makes this goal more practical than ever before. Here, we describe a new strategy, retro drug design, or RDD, to create novel small-molecule drugs from scratch to meet multiple predefined requirements, including biological activity against a drug target and optimal range of physicochemical and ADMET properties. The molecular structure was represented by an atom typing based molecular descriptor system, optATP, which was further transformed to the space of loading vectors from principal component analysis. Traditional predictive models were trained over experimental data for the target properties using optATP and shallow machine learning methods. The Monte Carlo sampling algorithm was then utilized to find the solutions in the space of loading vectors that have the target properties. Finally, a deep learning model was employed to decode molecular structures from the solutions. To test the feasibility of the algorithm, we challenged RDD to generate novel kinase inhibitors from random numbers with five different ADMET properties optimized at the same time. The best Tanimoto similarity score between the generated valid structures and the available 4,314 kinase inhibitors was < 0.50, indicating a high extent of novelty of the generated compounds. From the 3,040 structures that met all six target properties, 20 were selected for synthesis and experimental measurement of inhibition activity over 97 representative kinases and the ADMET properties. Fifteen and eight compounds were determined to be hits or strong hits, respectively. Five of the six strong kinase inhibitors have excellent experimental ADMET properties. The results presented in this paper illustrate that RDD has the potential to significantly improve the current drug discovery process.

■ INTRODUCTION

The primary goal of modern drug discovery is to identify molecules of therapeutic benefits. A successful drug molecule usually shares two features: 1. It modulates the biological function of its therapeutic target(s) selectively with optimal binding affinity; 2. It has a balanced ADMET (absorption, distribution, metabolism, excretion, and toxicity) profile, such that it reaches its target(s) unchanged and in sufficient quantity. Traditionally, a drug discovery project starts with screening a compound library against a proposed drug target or literature search,[1] followed by an optimization process to fix existing issues associated with original hit compounds, such as potency, selectivity, pharmacokinetics (PK), etc. This traditional drug discovery process requires tremendous input of resources and time. Computational generation of high-quality drug candidates with desired properties, a long-sought goal of pharmaceutical research, not only will reduce the unprecedented cost of bringing a drug to the market dramatically[2,3] but also will greatly speed up the whole process. Accelerated drug discovery and development is of paramount importance for public health threats of pandemics such as COVID-19.[4]

Received: February 1, 2022
Published: June 2, 2022



One approach is to explore virtual chemical space more broadly and efficiently. Chemical space is vast, like a galaxy. The past decade observed tremendous efforts to expand the coverage of both physical and virtual chemical space.[5] Merck MASSIVE 2018 contains 10^{20} virtual compounds, seconded by AZ space of 10^{17} capacity.[5] Even though supercomputers can handle these huge compound libraries, they are ignorable compared with the size of the estimated druglike chemical space of 10^{60}. To rebuild the galaxy is not an efficient way for drug hunting, although occasional successes have been achieved.[6]

The deep learning (DL) technology[7] brings hope to a new era of drug discovery and development and has the potential of substantially improving or even revolutionizing the current drug discovery paradigm. DL utilizes multiple layers of neurons to model high-level abstractions, complex and nonlinear relation in data, and has outperformed humans in many fields including image processing, text and voice recognition, protein structure prediction, and GO game; yet, this potential in drug discovery remains to be fulfilled.

Various machine learning and deep learning algorithms have been proposed over the past decade for generation of novel molecules with therapeutic benefits. Kadurin et al.,[8] Blascheke et al.,[9] and Lim et al.[10] used autoencoder, variational autoencoder, and adversarial autoencoder to identify and generate new molecular fingerprints with predefined properties. Bjerrum and Threlfall,[11] Cherti et al.,[12] and Segler et al.[13] utilized recurrent neural networks, in particular, the long short-term memory (LSTM) model,[14] to generate novel molecular structures with certain target properties.[15,16]

The autoencoder and RNN models are quite limited for generating novel molecules of predefined properties. A recurrent neural network (RNN) is a class of artificial neural networks in which connections between nodes form a directed graph along a temporal sequence. These models are not designed to assess, or optimize, the properties of the generated molecules. Furthermore, quality of deep learning models is largely determined by data quality and quantity on which they are based, and unfortunately the sample size and data quality of available experimental drug discovery data are usually insufficient for deep learning methods.[15]

To address the limitations of autoencoder and RNN models, various reinforcement learning (RL)[17] and generative adversarial networks (GANs)[18] have been proposed and implemented for sequence generation. These models typically consist of a sequence generation model, RL, and GAN. The RL and GAN models are used to optimize and revise the generated molecules toward the target properties.

Olivecrona et al.[9] introduced a method to tune a sequence-based generative model for de novo molecular design. Sanchez-Lengeling et al.[19] presented ORGANIC, a framework based on both GAN and RL, which can produce a distribution over molecular space that matches with a certain set of desirable metrics. Popova et al.[20] devised and implemented a novel computational strategy for de novo design of molecules with desired properties. As a typical strategy, it includes a generative model that produces a chemically valid SMILES string, predictive models that forecast the desired properties of the de novo-generated compounds, and a reinforcement learning module that tips the generated structures toward the desired properties. Putin et al.[21] reported a deep neural network, ATNC or Adversarial Threshold Neural Computer, for the de novo design of novel small-molecule organic structures with druglikeness properties. Zhou et al.[22] presented a framework, called Molecule Deep Q-Networks (MolDQN), for molecule optimization by combining domain knowledge of chemistry and state-of-the-art reinforcement learning techniques (double Q-learning and randomized value functions). One feature of MolDQN is that it can produce structures of 100% chemical validity. Zhavoronkov[23] developed a deep generative model, generative tensorial reinforcement learning (GENTRL), for de novo small-molecule design, and GENTRL produced several compounds, which were active in biochemical assays and cell-based assays. Ikebata et al.[24] used the Bayesian model to identify promising hypothetical molecules with a predefined set of desired properties.

Most of the efforts in de novo molecular design are based upon deep neural networks RNN, GAN, and RL. One latest example is REINVENT 2.0,[25] a powerful tool for de novo drug design. These DL methods demonstrated the good potential in drug discovery; however, there exist several limitations. First, deep learning models require a very large number of good quality samples, while the biological data tend to be noisy, limited in quantity, and severely imbalanced, representing a long-standing bottleneck for DNN methods.[15] Second, current methods are not efficient in sampling molecular structural space while optimizing ADMET properties. Local, small, and slow perturbations applied on existing molecules, represented either as SMILES strings (Daylight) or graphs, can hardly be efficient in exploring the vast possible chemical and structural space. This may partially explain why GENTRL took 21 days to produce several active compounds. In addition, sampling regardless of chemical structure validity inevitably leads to a very low percentage of valid output structures. For example, only 7% of the generated structures by ORGANIC[19] is valid. Third, RL algorithms tend to have difficulties in achieving a good balance between exploration and exploitation, in making long-term credit assignments, and in achieving good stability, likely due to aiming at moving targets. Fourthly, RL, although theoretically possible to optimize multiple target properties, in practice, is mostly found to optimize either one target property or the weighted sum of multiple target properties.

In order to overcome the above-mentioned limitations, we propose a new strategy called Retro Drug Design (RDD), starting with multiple preselected target properties and their optimal ranges, working backward to generate "qualified" compound structures.

RDD is based upon following rationales and considerations. First, given that the quantity of available biological data for small-molecule drug discovery is insufficient to train deep learning models, where from thousands to millions of variables are involved, it is sufficient for traditional, or shallow, machine learning models. Second, over the past decades, we developed a generic molecular descriptor system, called optATP, of 269 descriptors,[26] which have achieved outstanding performance in traditional machine learning prediction models of all the physicochemical and ADMET properties, accessible to us. Furthermore, optATP is originally designed to have good one-to-one correspondence with SMILES; in other words, one SMILES produces one optATP, and one optATP corresponds to as few molecules as possible. Third, optATP disassembles a molecule to atom types and functional groups, which greatly increases its coverage of chemical space.[27]

RDD is conceptually similar to the inverse QSPR/QSAR analysis for chemical structure generation.[28] However, there exist substantial differences in implementations. First, inverse

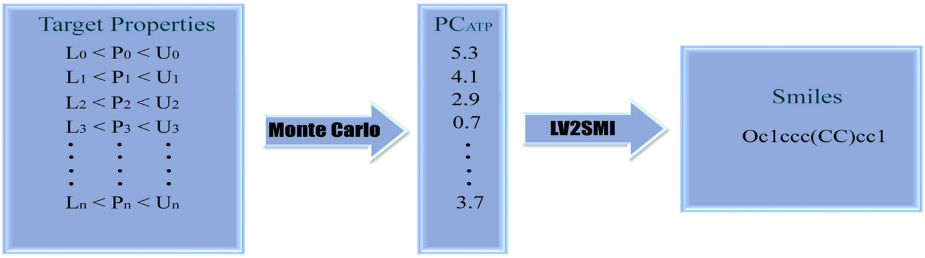


Figure 1. Flowchart for retro drug design.

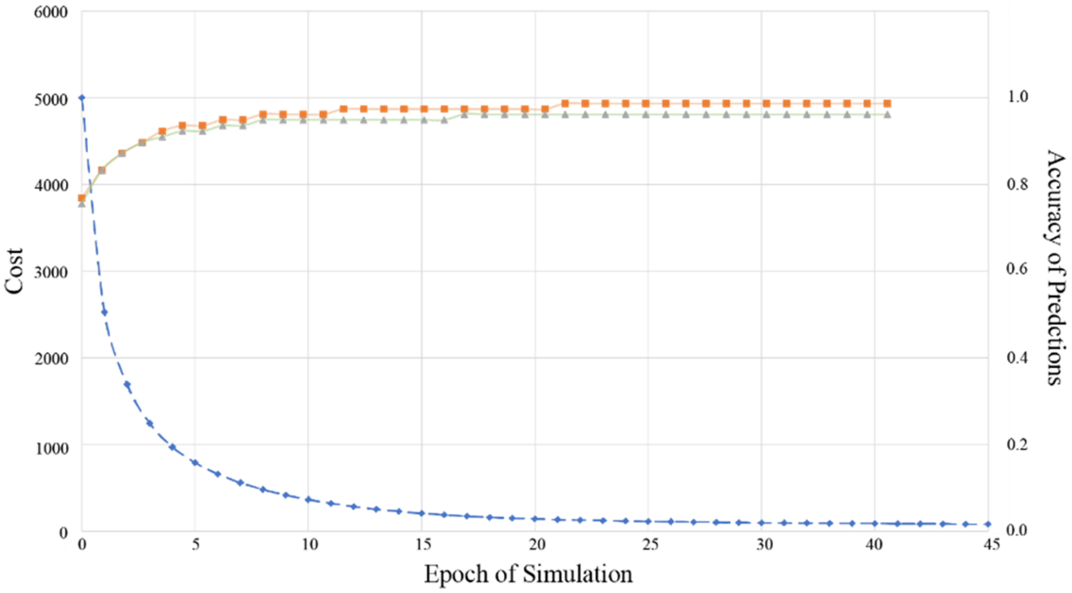


Figure 2. Cost and accuracy vs epoch on the training and validation data sets. The number of the units in both the fully connected layer and the GRU cell is 2048.

QSPR/QSAR used Gaussian mixture models and cluster-wise multiple linear regression, was a linear model, and seemed to support one target property only. Second, de novo generation of chemical structures with this inverse QSPR/QSAR algorithm, by combining ring systems and atom fragments in every possible way, was complex and less reliable and efficient than the deep learning algorithm utilized in RDD.

Kotsias et al. described an algorithm for de novo molecular generation with descriptor conditional recurrent neural networks.[29] This is an ingenious way to tackle the inverse design problem and generate molecules near the specified conditions. In RDD, we independently came up with and utilized the same idea for de novo generation of molecular structures.

To evaluate the performance of RDD, we challenged it to generate 120,000 novel small molecules that could inhibit protein kinase activity and satisfy five other important ADMET properties, including solubility and cytotoxicity. Twenty top-ranking hits were selected, synthesize, and tested for active-directed competition kinase binding and ADMET activities. Protein kinases are involved in regulation of nearly all aspects of cell life, representing an important class of drug targets for antitumor therapeutics and other diseases.[30] Most kinase inhibitors compete with ATP, the common substrate of kinases; thus, they share some common structural features—a flat aromatic moiety with an adjacent hydrogen bond donor and acceptor to facilitate favorable interactions with the highly conserved hinge region of kinases. This moiety tends to stack and aggregate in aqueous solutions, leading to poor solubility.[31,32] Being primarily antitumor targets, inhibition of many protein kinases causes cytotoxicity if their functions are essential for cell growth, apoptosis, or survival.[33,34] Therefore, to design three key properties, kinase inhibiting, soluble, and noncytotoxic, into one molecule itself represents a great challenge. If RDD could generate such kinds of molecules, it would be strong evidence that the algorithm is capable of navigating chemical space efficiently to locate a small niche where multiple seemingly exclusive properties are made concordant.

■ RESULTS

General RDD Workflow. RDD is a computational drug discovery platform that generates novel small-molecule drugs from scratch to meet predefined requirements, including but not limited to biological activity against a drug target and optimal range of physicochemical and ADMET properties. Molecular structures are represented by an optimized atom-type-based molecular descriptor system of 269 descriptors or optATP. Furthermore, principal component analysis (PCA) was performed on the optATP of the 906,727 unique molecular structures in the compound collection of the National Center for Advancing Translation Sciences (NCATS), and the feature or loading vectors of the principal components were employed to transform the representation of molecules from the space of optATP to the space of loading vectors, called ATP_LV, without sacrifice of much structural information. The target properties are computed by evaluators, which could be either predictive models trained over experimental data or mathematical functions.

The RDD workflow is illustrated in Figure 1. RDD starts with a list of target properties and their preferred ranges, such as logP between 2 and 5. The number of properties to be designed into a molecule is only limited by available computing resources. The properties can be physicochemical properties, such as molecular size, solubility, or biological properties. In this study, all properties are computed from the same molecular descriptor system, optATP, through an evaluator. The property range is defined by lower and upper boundaries of L_0−L_n and U_0−U_n (Figure 1). The Monte Carlo sampling algorithm was then utilized to find the solutions in the space of ATP_LV that have the target properties. Finally, the deep learning model of LV2SMI was employed to decode molecular structures from the solutions.

LV2SMI. We used all of the 906,727 unique molecules in the NCATS compound collection, computed optATP, performed principal component analysis (PCA), and found that 7, 14, and 38 principal components account for 95, 97, and 99% of the total variance, respectively. This collection is of pharmaceutical interest, consisting of the marketed drugs, drugs that have reached clinical trials, and other bioactive molecules. Using feature or loading vectors of the principal components, we transformed the representation of molecules from the space of optATP to the space of loading vectors, called ATP_LV.

We designed and trained a deep learning model LV2SMI over the entire collection of 906,727 molecular structures to decode molecular structures as represented by SMILES strings[35] from loading vectors. For each molecule, three ATP_LVs were calculated using loading or feature vectors from 7, 14, and 38 principal components. Each ATP_LV and the corresponding ground truth SMILES form a data sample of input and output. The 906,727 samples were randomly split into a training data set of 816,424 samples (90%), a validation data set of 45,306 samples (5%), and a test data set of 44,998 samples (5%).

The training process took from about 1 day for an ATP_LV dimension of 7 and a GRU cell of 1,024 units to 2 days for an ATP_LV dimension of 38 and a GRU cell of 2,048. Gated recurrent units (GRUs) are a gating mechanism in RNN.[36] The cost and the accuracies versus epoch on the training and validation data sets were plotted in Figure 2. The cost dropped dramatically in the first 10 epochs and then continued to decrease smoothly. The accuracies on both training and validation data sets also increased rapidly in the first 10 epochs and then continued to improve slowly.

We trained 12 LV2SMI models using ATP_LV dimensions of 7, 14, and 38 elements and GRU cells of 1024, 1280, 1536, and 2048 units, and the accuracies of the optimized models on the test data set are given in Table 1. The accuracy improves as both the ATP_LV dimension and the unit number of the GRU cell in the decoder increase. The striking factor is that even the simplest model with a dimension of 7 and a GRU cell of 1024 units achieved excellent accuracy.

A dimension of 38 and a unit size of 2048 were adopted to test the performance of the LV2SMI model. Among the 44,998 SMILES generated for the test set, 42,053 (93.5%) are chemically valid. Among the 42,053 valid SMILES, 19,716 (46.9%) molecules are identical to the ground truth, and 40,762 (96.9%) have a cosine similarity score of optATP greater than 0.95.

Table 1. Accuracies of the Optimized Models on the Test Data Set

epoch	ATP_LV dimension	no.of units in GRU cell	accuracy
120	7	1024	0.885
40	7	1280	0.885
30	7	1536	0.885
70	7	2048	0.897
120	14	1024	0.923
40	14	1280	0.923
110	14	1536	0.936
30	14	2048	0.936
250	38	1024	0.962
80	38	1280	0.962
50	38	1536	0.962
120	38	2048	0.974

Generation of Ligands for Kinase Competition Binding with Desired ADMET Properties. Protein kinases regulate nearly all aspects of cell life,[30] so they are the major drug targets of small-molecule drug discovery. As of March 2021, there are 65 FDA-approved small-molecule protein kinase inhibitor drugs.[37] Protein kinase inhibitors represent a large, structurally diverse compound collection, the majority of which target the same adenosine triphosphate binding pocket. It has been found that a large fraction of the collection is poorly soluble and cytotoxic.[31,33] Therefore, we determined to challenge RDD to create soluble and noncytotoxic kinase inhibitors with a favorable ADMET profile, in order to evaluate its efficiency in sampling vast chemical space.

The Monte Carlo sampling was performed in the space of 38 loading vectors from PCA. Starting from vectors of random numbers, RDD generated 120,000 chemical structures, of which 75% (90,215) were chemically valid. On average, RDD found 100 solutions per CPU Core-Hour with six desired target properties. Out of the 90,215 valid structures, about 26% (31,484) were predicted to be active kinase inhibitors. For the other five ADMET properties, 72,326, 40,955, 68,016, 40,196, and 54,559 were predicted to have the desired target property (Table 3) for logP, solubility, hERG, PAMPA, and Cytotox, respectively. 3,040 structures are predicted to have all six target properties.

Experimental Confirmation of the Target Properties. From the top-ranking structures for kinase activity, 20 compounds with a balanced ADMET profile were selected for synthesis and experimental measurement of competition binding activity over 97 representative kinases and ADMET properties. The main criteria for the selection are predicted kinase activity, solubility and cytotoxicity, synthesizability, and structure diversity that aim to provide some structure−property relationship (SPR). The SMILES of these 20 selected structures are given in Table S1. All compounds are >95% pure by HPLC analysis.

These 20 selected compounds were assayed by Eurofins Discovery (San Diego, USA) using the KINOMEscan panel of scanEDGE. scanEDGE includes 97 kinases distributed throughout the AGC, CAMK, CMGC, CK1, STE, TK, TKL, lipid, and atypical kinase families, plus important mutant forms. scanEDGE is an economical alternative to scanMAX with maximized coverage of the kinome space. The assay results are given in Table S2.

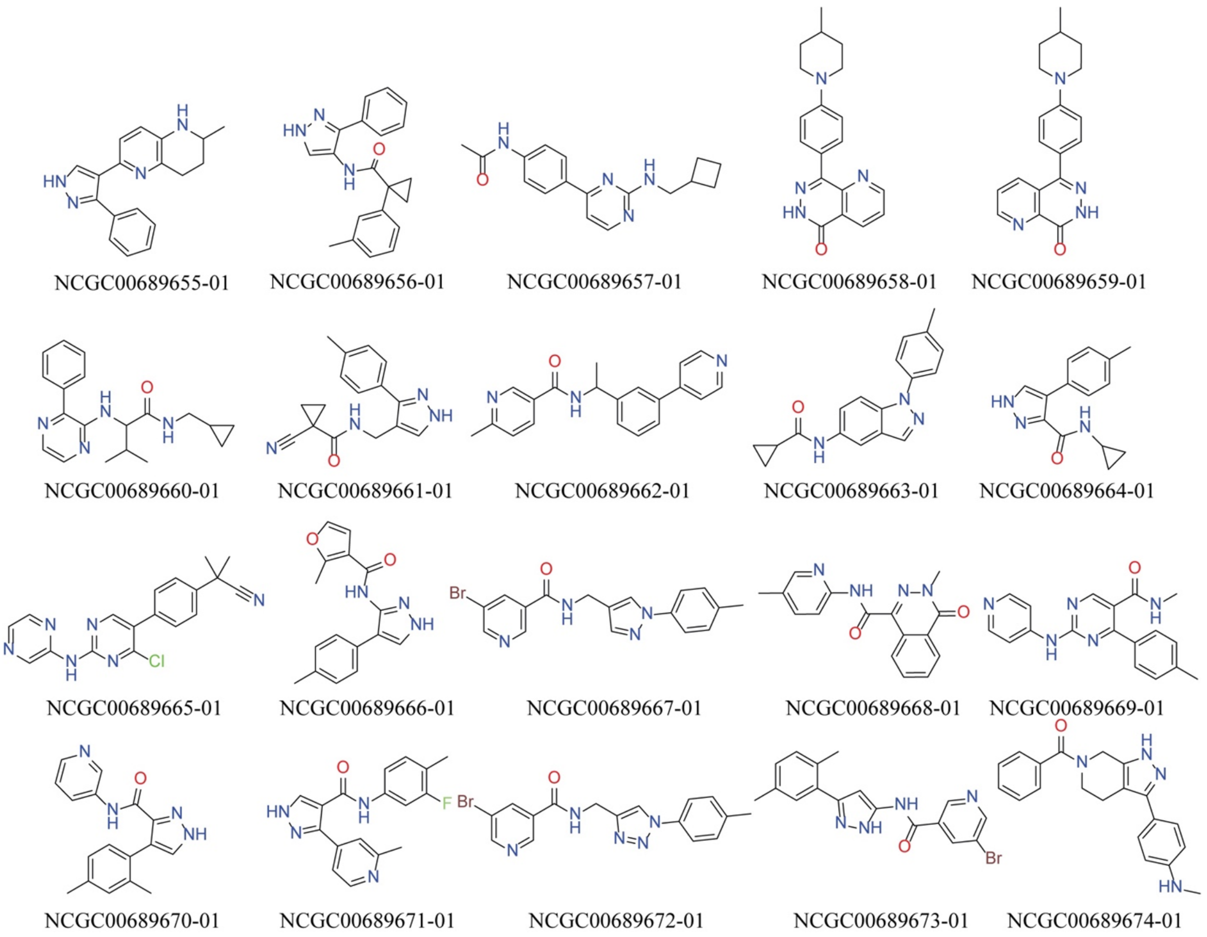


Figure 3. Chemical structures of the 20 compounds designed by RDD. Six compounds exhibit strong kinase inhibitory activity, and their sample IDs and best ECFP4 similarity scores[38] (RDKit 2022.03.1) with the 4,426 kinase inhibitors in the PKIDB[39] are NCGC00689657, NCGC00689660, NCGC00689661, NCGC00689669, NCGC00689670, and NCGC00689674 and 0.59, 0.42, 0.44, 0.49, 0.45, and 0.45, respectively.

Table 2. Predicted and Experimentally Measured Kinase Competition Binding and Five Other ADMET Properties of the 20 Compounds Generated from the Retro Drug Design Pipeline and Synthesized by Wuxi AppTec[b]

	Predicted						Measured				
ID	Kinase	logP	Solubility	Cytotoxicity	HERG	PAMPA	Kinase	Solubility	Cytotoxicity	HERG	PAMPA
NCGC00689655	0.99	3.43	0.56	0.38	0.05	0.98	w	s	n	n	w
NCGC00689656	0.99	3.78	0.38	0.45	0.02	0.69	w	s	n	n	s
NCGC00689657	0.98	3.10	0.72	0.69	0.06	0.95	s	n	n	n	w
NCGC00689658	0.97	3.46	0.38	0.46	0.17	0.97	w	s	n	n	s
NCGC00689659	0.97	3.46	0.38	0.46	0.17	0.97	w	s	n	n	s
NCGC00689660	0.95	3.51	0.98	0.50	0.10	0.81	s	s	n	n	s
NCGC00689661	0.98	1.56	0.87	0.25	0.01	0.53	s	s	n	n	s
NCGC00689662	0.98	3.27	0.83	0.29	0.16	0.70	w	s	n	n	n
NCGC00689663	0.97	3.62	0.08	0.27	0.03	0.85	n	n	n	n	s
NCGC00689664	0.99	2.49	0.96	0.17	0.01	0.53	w	s	n	n	s
NCGC00689665	0.99	4.67	0.20	0.78	0.10	0.50	n	w	n	n	s
NCGC00689666	1.00	3.30	0.77	0.41	0.01	0.76	n	s	n	n	s
NCGC00689667	0.97	3.05	0.84	0.28	0.04	0.32	n	w	n	n	s
NCGC00689668	0.93	2.17	0.72	0.38	0.02	0.68	n	w	n	n	s
NCGC00689669	0.99	2.68	0.57	0.71	0.03	0.64	s	s	n	w	s
NCGC00689670	1.00	2.91	0.83	0.48	0.01	0.78	s	s	w	n	s
NCGC00689671	0.99	3.11	0.64	0.47	0.02	0.87	w	s	n	n	s
NCGC00689672	0.95	2.71	0.81	0.34	0.02	0.30	w	s	n	n	s
NCGC00689673	1.00	3.69	0.37	0.42	0.03	0.88	w	n	n	n	s
NCGC00689674	0.98	3.17	0.61	0.78	0.09	0.86	s	s	n	n	s

[a]s: strong; w: weak; n: not detectable. [b]The predicted numbers are binary classification model probability scores from SVM with the exception of logP, which is predicted using a regression model.

Fifteen of the 20 compounds exhibited kinase competition binding activity with normalized kinase competition binding percent activity < 35%, among which six compounds were identified as strong kinase inhibitors with kinase competition binding percent activity < 10% (Figure 3). Five compounds did not have detectable activities at 10 μM concentration against the 97 kinases in the panel. One possible explanation of the observation is that this kinase panel only covers less than one-fifth of the human kinome, while the kinase inhibitors in the training set exhibit a broader coverage.

As shown in Figure 3, three kinase inhibitors are pyrimidine-2-amines, and three compounds have pyrazoles. Both fragments are recognized as privileged structures of kinase inhibitors.[40] Indeed, pyrimidine-2-amines and pyrazoles appear in 48 and 53 of 244 kinase drug candidates entered clinical trials and 756 and 1,056 of 4,426 kinase inhibitors in the PKIDB.[39] This is strong evidence that RDD is capable of locating subspace for kinase inhibitors, although the searching starts from random numbers without any structural information provided.

Predicted and experimentally measured kinase competition binding and five other ADMET properties of the 20 compounds generated from RDD pipeline and synthesized by Wuxi AppTec are given in Table 2. The primary goal of the experiment is to design three seemingly conflicting properties into one molecule, i.e., to discover novel kinase inhibitors with good aqueous solubility but without cytotoxicity. The chemical space that satisfies these three requirements is limited. It is challenging for RDD to navigate the vast chemical space to locate this tiny subspace. The results we presented here clearly indicated that RDD was capable of spotting the specific area in the vast chemical space, where multiple subspaces overlapped. On the other hand, hERG, an important toxicity end point, and permeability, a determinant factor of a drug's PK profile, are not seemingly correlated with the kinase activity of a compound. In other words, the subspaces for kinase inhibitors, for hERG inactives, and for good permeability might share a larger common area in the chemical space; therefore, generating novel molecules to meet these three requirements is less challenging than designing novel kinase inhibitors with good solubility and without cytotoxicity. In this study, we demonstrated the power of RDD to design all the properties into one molecule. The six strong kinase inhibitors are structurally diverse (Figure 3), having excellent profiles of lipophilicity, solubility, cytotoxicity, permeability, and cardiotoxicity (hERG), except NCGC00689657. The compound NCGC00689657 has poor solubility and permeability, although its predicted solubility and permeability are good (Table 2).

■ DISCUSSION AND CONCLUSIONS

We demonstrated that RDD was able to generate novel chemical structures with multiple targeted properties. In this computational drug discovery platform, optATP is used to dismantle a molecule to smaller pieces, such as atoms and functional groups, which greatly expand the coverage of chemical space. SVM and other ML models are used to extract the important or discriminant atom types associated with a protein target or a property, in other words, to identify subspaces shaped by predefined properties, whereas the Monte Carlo algorithm serves as a search engine to locate the intersection of such subspaces, Finally, LV2SMI reassembles the atomic information represented by the subspaces back to molecules.

When optATP was designed, one of the major motivations was to produce a universal molecular descriptor system; in other words, the same system can be applied to generate QSAR models for all properties.[41] Each atom type was also designed to carry sufficient information on its chemical environment, so optATP descriptors and the corresponding SMILES could be as interconvertible as possible. Our following efforts have confirmed that optATP meets our design goals. optATP provided excellent QSAR models for all the data sets available to us, and most of the constructed QSAR models achieved accuracy comparable to experimental determinations.[26,31,33,42] Even the simplest LV2SMI model with a dimension of 7 and a GRU cell of 1024 units achieved excellent accuracy in reproducing a molecular structure from optATP. Furthermore, optATP has an excellent dimension reducibility. Through PCA, the space of 269 optATP descriptors could be reduced to the space of 38 loading vectors (ATP_LV) with 99% coverage of the variance.

Unlike typical deep generative models that usually start from existing molecules and introduce different levels of perturbation to generate new structures in the space of the molecular graph, RDD did not directly search the chemical structural space; instead, the Monte Carlo search was performed in the ATP_LV space, i.e., the space of 38 loading vectors. Consequently, RDD can explore a much larger structural space in a much more efficient manner without invalidating molecular structures. Starting from random numbers, furthermore, RDD aims at searching chemical space in an unbiased manner. In this study, a high percentage of the structures generated by RDD is both chemically valid and novel. Out of 120,000 generated structures, 90,215 were chemically valid. Out of the 90,215 valid structures, the best Tanimoto similarity score with the available 4,314 kinase inhibitor drugs/ligands is < 0.35, and only 491 are available in Aldrich's catalog of over 20 million compounds. These results also indicate that the collection of commercially available chemicals is tiny and insufficient to provide a good coverage of the vast chemical space.

To generate a valid SMILES from scratch is not trivial. There are numerous underlying rules to follow in order to avoid invalid moieties, such as 5-carbon aromatic rings or 5-bond carbons. RDD can learn these chemical and structural rules and incorporate them into the process of generating new structures, as indicated by the results—3/4 of the RDD-generated compounds are chemically valid. In other words, RDD has not only learned these rules but also knows how to apply these rules and compose valid chemical structures.

Designed as a platform, RDD allows multiple properties to be optimized individually and yet simultaneously using various evaluators of individual properties. Since typical machine learning models such as SVM could be trained over specific experimental data sets of moderate size, as few as several hundred data points, RDD does not need a very large quantity of quality samples to generate new structures with multiple desired properties. The Monte Carlo sampling algorithm is very stable and fast.

RDD allows scientists dialing in or dialing out biological and ADMET properties to meet their specific designing requirements, paving the road toward true rational drug design. While RDD has the potential of starting a new paradigm for drug discovery, one major limitation is apparent. In this study, the prediction model for ligand-kinase binding affinity is derived from available, 4,314 known and experimentally determined kinase inhibitors from the PKIDB[39] and 5,894 nonkinase inhibitors from the NCATS pharmaceutical collection (NPC),[43] and it is not for a specific kinase. The generated ligands are expected to modulate various kinases and are promiscuous; surprisingly, the hit compounds are quite selective (Table S3) for reasons we do not fully understand. As more 3D structures of proteins become available,[44] a generic evaluator or a model that could predict a ligand's affinity with a target is strongly desired; such a model would make RDD applicable to most if not all druggable targets. Current available docking algorithms such as AutoDock Vina[45] are too slow or inaccurate to be feasible for this purpose. We are currently testing a model that is based upon a decoupled 3D fingerprint and trained over a very large number of experimental data sets.[46]

In conclusion, we proposed and developed a new computational drug discovery platform, Retro Drug Design. The RDD platform is capable of generating highly novel structures from scratch to meet predefined requirements, such as kinase competition binding activity and five other physical-chemical and ADMET properties. The availability of such a highly efficient and productive drug discovery platform is essential in handling emergent public health threats, such as the pandemic of COVID-19.

■ EXPERIMENTAL SECTION

Molecular Representation. The atom-type-based molecular descriptor system, or optATP, consisting of 221 atom types and 48 correction factors, was employed to represent small molecules. The details of the molecular descriptors have been elaborated elsewhere.[41] Atom types are assigned according to the properties of an atom and its chemical environment (Figure S1). An atom type casting tree was designed to assign atom types, based on whether the atom is aromatic, whether the atom is in a ring, whether the atom is next to different functional groups, etc. This original tree, largely based on a medicinal chemist's intuition, was subject to recursive optimization cycles in terms of where to further split the tree, where to stop splitting, and where to combine the branches, in order to make the best prediction of logP values in the Starlist data set containing about 11,000 structurally diverse compounds. Here, an atom in a molecule is like a piece of puzzle chip in a puzzle, which has its unique edge. When a set of puzzle chipsis provided, the puzzle can be solved unambiguously, because of the unique shape of each piece.

The optimized tree output 221 atom types, featuring 88 different carbon types, 7 hydrogen types, 58 nitrogen types, 31 oxygen types, 8 halide types, 23 sulfur types, and 6 phosphorus types. Forty-eight correction factors are appended to catch several whole molecule features, such as the molecular globularity, molecular rigidity, lipophilicity, etc. In total, a series of 269 numerical values comprise the final set of the atom type molecular descriptors.

Using optATP representation, we could design molecules of predefined properties by sampling in the optATP space of 269 dimensions. As is well-known, sampling in such high dimensional space is extremely challenging. Fortunately, there exists a substantial amount of dependency among the 269 descriptors. To reduce the dependency and the dimension of the sampling space, we performed principal component analysis (PCA) on the optATP of 906,727 unique molecular structures in the entire compound collection of NCATS and found that 7, 14, and 38 principal components account for 95, 97, and 99% of the total variance, respectively. Using feature or loading vectors of the principal components, we successfully transformed the representation of molecules from the space of optATP to the space of loading vectors, called ATP_LV, without sacrifice of much structural information.

The ATP_LV space not only has reduced dimension (from 269 to 7, 14, or 38) but also provided orthogonal principal vectors with highly loaded information. Furthermore, different dimensions of the ATP_LV space correspond to molecular structural properties at different scales. For example, the first dimension, defined by the first principal component, captures the molecular characteristics at the largest scale. By sampling the molecular structure in the ATP_LV space, RDD can explore a much larger structural space in a much more efficient manner without invalidating sampled molecular structures.

Monte Carlo Search. With the input list of target properties and their desired ranges, RDD uses a number of evaluators of P_0-P_n (Figure 1) and a Monte Carlo (MC) sampling algorithm to find solutions in the ATP_LV space that have the desired ranges according to the corresponding evaluators. The Monte Carlo sampling algorithm is very powerful for numerical optimization[47] of multivariable function. In order to sample in the more druglike region of the ATP_LV space, we calculated the mean (ATP_LV_Mean) and the standard deviation (ATP_LV_Std) from the ATP_LV for 906,727 unique molecular structures in NCATS's compound collection for MC sampling.

The MC algorithm was carried out as follows:

1. Start from a randomly initialized ATP_LV according to eq 1

$$\mathrm{lv[i]} = \mathrm{ATP\_LV\_Mean}[i] + \mathrm{gr} \times \mathrm{ATP\_LV\_Std}[i] \tag{1}$$

where gr is a Gaussian random number generator with a mean of 0.0 and standard deviation of 1.0. Reversely transform the current lv back to optATP using the feature vectors from PCA, apply the evaluators, and compute the output score S_i. Then, calculate the initial cost according to eq 2

$$C = \sum_{i=0}^{n} w_i(L_i - S_i)_{S_i < L_i} + w_i(S_i - U_i)_{S_i > U_i} \tag{2}$$

where n is the number of properties or evaluators, w_i is the weighting factor for the ith evaluator, and L_i and U_i are the lower and upper boundaries of the ith property.

2. Randomly pick eight elements (optional) from a previous lv and perturb each by adding 0.5*(r − 0.5) * ATP_LV_Std[i] to propose a new lv. r is a uniform random number generator between 0.0 and 1.0.
3. Recalculate the cost of the current lv. If the current cost is < 0.01 (optional), stop the sampling process and output the solution. If the current cost is smaller than the previous one, accept the perturbation; otherwise reject the perturbation. Go back to step 2.
4. If no solution is found after 50 steps, go back to step 1.
5. If no solution is found after 40,000 steps, stop and terminate.

LV2SMI. Taking the solution of lv in the ATP_LV space from the MC algorithm, RDD uses a deep learning model, called LV2SMI, to map lv to molecular structures as represented by SMILES. LV2SMI is adapted from the widely used SEQ2SEQ model.[48] A typical SEQ2SEQ model consists of an encoder and a decoder; the encoder transforms the sequence input to a vector of latent variables, which is then transformed to the output sequence by the decoder. In LV2SMI, we removed the encoder and directly used ATP_LV as the vector of latent variables and then used a decoder to transform ATP_LV to output sequence-SMILES (Figure 4).

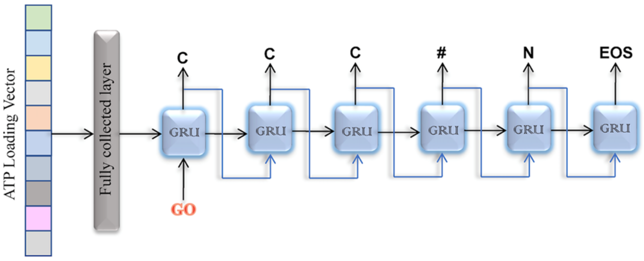


Figure 4. LV2SMI network. The loading vector of ATP_LV is fed to a fully connected layer, FC, and the output is fed to a decoder, a recurrent neural network of GRU cells. Each GRU cell outputs a letter of a SMILES.

SMILES uses a line notation to represent a molecular structure as ASCII strings. In this study, we ignored chirality in a molecular structure and used a vocabulary of only 39 letters or words—A, T, E, U, = # % ( ) [ ]/\ 0 1 2 3 4 5 6 7 8 9 N O P S F I c n o p s - + Br Cl—to encode a SMILES. A, T, E, and U are tokens for padding, starting, stopping, and extraordinary letters outside this vocabulary list. Each element in this vocabulary list is encoded by a one hot vector. We ignored chirality in a molecular structure, mainly for simplification purposes, and plan to include chirality in future studies.

In the LV2SMI, as shown in Figure 4, a loading vector is fed to a fully connected layer, FC, and its output is fed to a decoder, a recurrent neural network of GRU cells. The number of units in the FC is the same as that in the GRU cell. Four different numbers of units, 1,024, 1,280, 1,536 and 2,048, were used in search for a balance between accuracy and efficiency.

Sparse categorical cross entropy is used as the loss or cost function. Backpropagation is employed for training the network.[49] Optimization of the loss function is carried out by minibatch of a size 128 and the ADAM optimizer,[50] which is implemented as tf.train.AdamOptimizer in the Tensorflow library.[51] For the ADAM optimizer, a learning rate of 0.001 was selected according to our previous experience to produce a satisfactory result.

The model training process was monitored by two metrics: cost function and accuracy on both the training and validation data sets. The model with the best accuracy on the validation set was saved and applied to the test data set to collect chemistry specific benchmarks. In this study, we computed three chemistry specific metrics. The first one was the percentage of the generated structures that was chemically valid. A structure is considered valid if it is successfully parsed by ChemAxon's molecular parser.[52] The second metrics was the percentage of the valid structures that was identical with the ground truth. The third was the percentage of the structures having a cosine similarity score of optATP > 0.95.

SVM Models as Evaluators of Kinase Competition Binding Activity and ADMET Properties. Designed as a platform, RDD supports various evaluators/plugins, and it allows multiple properties to be optimized individually and simultaneously. In this study, we chose six target properties and built corresponding support vector machine (SVM) models from experimental data as evaluators.

Model for Predicting Kinase Competition Binding. The 4,314 known kinase inhibitors from the PKIDB[39] and 5,894 nonkinase inhibitors from the NCATS pharmaceutical collection (NPC)[43] comprised the data set for kinase/nonkinase classification. optATP of 269 descriptors for each molecular structure was computed as molecular descriptors.[26] The 10,208 samples were randomly split into training (80%) and testing (20%) data sets. The SVM model was trained on the training data set using the software package of LIB-SVM (C-C Change 2001). The parametrization of the penalty for misclassification, *C*, and the nonlinearity parameter in the kernel function of a Gaussian Radial Basis Function (RBF), γ, was accomplished on a grid-based search to minimize the mean standard error (MSE) of 5-fold cross-validation (CV) on the training data. The prediction results on the test data set were used for benchmarking the accuracy, root-mean-square error, or AUC-ROC^3. The benchmarking results are given in Table 3.

Table 3. Data, Performance, and Predefined Boundaries of SVM Models for Kinase Competition Binding Activity and Five ADMET Properties

property	logP[a]	solubility	hERG	PAMPA	Cytotox[b]	kinase inhibition
total[c]	10850	21993	3022	5435	5275	10208
Pos[d]		10827	482	2406	620	4314
accuracy		0.86	0.83	0.83	0.89	0.92
AUC[e,f]	0.14	0.93	0.92	0.90	0.89	0.97
lower[g]	1.00	0.70	0.00	0.52	0.00	0.50
upper[h]	5.00	1.00	0.19	1.00	0.50	1.00
cutoff		0.46	0.3	0.46	0.57	0.45

[a] Regression model; all of the other five models are classification models. [b] Seven submodels were used due to the imbalance data.[33] [c] Total number of samples. [d] Number of positive samples. [e] Area under the Receiver Operating Characteristics (ROC) curve. The number for logP is mean squared error. [f] Root-mean-square error. [g] Lower boundary of model score for target property. [h] Upper boundary of the model score for the target property. The classification criteria for solubility, hERG, PAMPA, Cytotox, and kinase inhibition are 0.458, 0.296, 0.459, 0.570, and 0.453, respectively.

The SVM models for logP, solubility, hERG, PAMPA, and Cytotox were described elsewhere.[26,31,41,42,53] The same procedure as the one used for kinase competition binding prediction was followed for constructing these SVM models. The details for training data and model performances are given in Table 3. The lower and upper boundaries for classification models were selected according to the optimal cutoffs of the corresponding classifiers.

Kinase Competition Binding Assay. The KINOMEscan screening platform by Eurofins Discovery Services[54] is a novel and proprietary active site-directed competition binding assay to quantitatively measure interactions between test compounds and more than 489 kinase assays and disease relevant mutant variants. KINOMEscan assays do not require ATP and thereby report true thermodynamic interaction affinities, as opposed to IC_{50} values, which can depend on the optATP concentration.

In this study, we tested the 20 compounds generated from the RDD pipeline using the KINOMEscan panel of scanEDGE. scanEDGE includes 97 kinases distributed throughout the AGC, CAMK, CMGC, CK1, STE, TK, TKL, lipid, and atypical kinase families, plus important mutant forms. scanEDGE is an economical alternative to scanMAX with reasonable coverage of the diversity of the whole KINOME.

The 20 compounds were screened at 10 μM, and results for primary screen binding interactions are reported as "% Ctrl" according to eq 3, where lower numbers indicate stronger inhibition activity.

$$\%\mathrm{Ctrl} = \frac{\text{test compound signal} - \text{positive control signal}}{\text{negative control signal} - \text{positive control sigmal}} \tag{3}$$

Computer Hardware and Software. The computations were performed on a Dell PowerEdge R940xa server with four Intel Xeon Platinum 8160 processors (each with 24 cores), 3TB of RAM, and four 16GB NVIDIA Tesla V100 graphic processing units, installed with Ubuntu 16.04.6 distribution, Python 3.5, CUDA driver version 10.0, cuDNN version 7.4, TensorRT 5.1, and TensorFlow 1.13.1. A Java program was developed in-house to use JOELib (JOELib) for molecule structure parsing and optATP calculation. The Monte Carlo algorithm was coded in Java, and the ATP2SMI module was coded in Python.

DATA AND SOFTWARE AVAILABILITY

The source codes and the required data and library files are all available from Google Drive: https://drive.google.com/drive/folders/1nirsvwKeMKrhC2nvU9gvPyhrWvqN1hmD?usp=sharing. A readme file is also included about how to execute the programs.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.2c00123.

> Experimental details of kinase inhibition assay, SMILES of 20 compounds synthesized and experimentally tested, hit kinases and % Ctrl values (<35%) for 20 compounds, S-scores of 20 compounds from KINOMEscan/scanEDGE screening, and basic design considerations of optATP (PDF)

AUTHOR INFORMATION

Corresponding Authors

Yuhong Wang − *National Center for Advancing Translational Sciences (NCATS), Rockville, Maryland 20850, United States;* orcid.org/0000-0001-6480-9528; Email: yuhong.wang@nih.gov

Hongmao Sun − *National Center for Advancing Translational Sciences (NCATS), Rockville, Maryland 20850, United States;* orcid.org/0000-0003-4042-6498; Email: hongmao.sun@nih.gov

Authors

Sam Michael − *National Center for Advancing Translational Sciences (NCATS), Rockville, Maryland 20850, United States*

Shyh-Ming Yang − *National Center for Advancing Translational Sciences (NCATS), Rockville, Maryland 20850, United States*

Ruili Huang − *National Center for Advancing Translational Sciences (NCATS), Rockville, Maryland 20850, United States*

Kennie Cruz-Gutierrez − *National Center for Advancing Translational Sciences (NCATS), Rockville, Maryland 20850, United States*

Yaqing Zhang − *National Center for Advancing Translational Sciences (NCATS), Rockville, Maryland 20850, United States*

Jinghua Zhao − *National Center for Advancing Translational Sciences (NCATS), Rockville, Maryland 20850, United States*

Menghang Xia − *National Center for Advancing Translational Sciences (NCATS), Rockville, Maryland 20850, United States;* orcid.org/0000-0001-7285-8469

Paul Shinn − *National Center for Advancing Translational Sciences (NCATS), Rockville, Maryland 20850, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.2c00123

Author Contributions

Y.W. contributed the key ideas, developed/implemented algorithms/codes, and prepared the manuscript. S.M. contributed several ideas. S.M.Y. provided the medicinal chemistry support. K.C.G. installed/supported the deep learning tools. Y.Q.Z. performed the cytotoxicity assay. R.H. analyzed the assay data. M.X. and J.Z. performed the hERG assay, P.S. provided compound management, and H.S. contributed key ideas and coprepared the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was supported in part by the Intramural/Extramural research program of the NCATS, NIH.

REFERENCES

(1) Dragovich, P. S.; Haap, W.; Mulvihill, M. M.; Plancher, J. M.; Stepan, A. F. Small-Molecule Lead-Finding Trends across the Roche and Genentech Research Organizations. *J. Med. Chem.* **2022**, *65* (4), 3606−3615.

(2) DiMasi, J. A.; Grabowski, H. G.; Hansen, R. W. Innovation in the pharmaceutical industry: New estimates of R&D costs. *J. Health Econ* **2016**, *47*, 20−33. Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov* **2010**, *9* (3), 203−214.

(3) Hughes, J. P.; Rees, S.; Kalindjian, S. B.; Philpott, K. L. Principles of early drug discovery. *Br. J. Pharmacol.* **2011**, *162* (6), 1239−1249.

(4) Administration, U. S. F. D. *Coronavirus Treatment Acceleration Program*; 2020. https://www.fda.gov/drugs/coronavirus-covid-19-drugs/coronavirus-treatment-acceleration-program-ctap (accessed 2022-05-26). Grobler, J. A.; Anderson, A. S.; Fernandes, P.; Diamond, M. S.; Colvis, C. M.; Menetski, J. P.; Alvarez, R. M.; Young, J. A. T.; Carter, K. L. Accelerated Preclinical Paths to Support Rapid Development of COVID-19 Therapeutics. *Cell Host Microbe* **2020**, *28* (5), 638−645.

(5) Hoffmann, T.; Gastreich, M. The next level in chemical space navigation: going far beyond enumerable compound libraries. *Drug Discov Today* **2019**, *24* (5), 1148−1156.

(6) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; et al. Ultra-large

library docking for discovering new chemotypes. *Nature* **2019**, *566* (7743), 224−229.

(7) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521* (7553), 436−444. Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw* **2015**, *61*, 85−117.

(8) 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. Pharmaceutics* **2017**, *14* (9), 3098−3104.

(9) Blaschke, T.; Olivecrona, M.; Engkvist, O.; Bajorath, J.; Chen, H. Application of Generative Autoencoder in De Novo Molecular Design. *Mol. Inform.* **2018**, *37* (1−2), 170012.

(10) Lim, J.; Ryu, S.; Kim, J. W.; Kim, W. Y. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *J. Cheminform* **2018**, *10* (1), 31.

(11) Bjerrum, E. J.; Threlfall, R. Molecular Generation with Recurrent Neural Networks (RNNs). 2017, *arXiv:1705.04612*. https://arxiv.org/abs/1705.04612 (accessed 2022-05-26).

(12) Medhd, C.; Kegl, B.; Kazakci, A. DE NOVO DRUG DESIGN WITH DEEP GENERATIVE MODELS: AN EMPIRICAL STUDY. In *International Conference on learning Representations*, Toulon, France; 2017.

(13) 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.

(14) Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput* **1997**, *9* (8), 1735−1780.

(15) Schneider, P.; Walters, W. P.; Plowright, A. T.; Sieroka, N.; Listgarten, J.; Goodnow, R. A., Jr; Fisher, J.; Jansen, J. M.; Duca, J. S.; Rush, T. S.; et al. Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov* **2020**, *19* (5), 353−364.

(16) Reutlinger, M.; Rodrigues, T.; Schneider, P.; Schneider, G. Multi-objective molecular de novo design by adaptive fragment prioritization. *Angew. Chem., Int. Ed. Engl.* **2014**, *53* (16), 4244−4248.

(17) Sutton, R. S.; Barto, A. *Reinforcement Learning: An Introduction*; Bradford Book: 1998.

(18) Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*; 2014.

(19) Sanchez-Lengeling, B.; Outeiral, C.; Guimaraes, G. L.; Aspuru-Guzik, A. Optimizing distributions over molecular space. An Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry (ORGANIC). 2017, https://chemrxiv.org/articles/preprint/ORGANIC_1_pdf/5309668 (accessed 2022-05-26), DOI: 10.26434/chemrxiv.5309668.v3.

(20) Popova, M.; Isayev, O.; Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **2018**, *4* (7), eaap7885.

(21) Putin, E.; Asadulaev, A.; Vanhaelen, Q.; Ivanenkov, Y.; Aladinskaya, A. V.; Aliper, A.; Zhavoronkov, A. Adversarial Threshold Neural Computer for Molecular de Novo Design. *Mol. Pharmaceutics* **2018**, *15* (10), 4386−4397.

(22) Zhou, Z.; Kearnes, S.; Li, L.; Zare, R. N.; Riley, P. Optimization of Molecules via Deep Reinforcement Learning. *Sci. Rep* **2019**, *9* (1), 10752.

(23) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **2019**, *37* (9), 1038−1040.

(24) Ikebata, H.; Hongo, K.; Isomura, T.; Maezono, R.; Yoshida, R. Bayesian molecular design with a chemical language model. *J. Comput. Aided Mol. Des* **2017**, *31* (4), 379−391.

(25) Blaschke, T.; Arus-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. REINVENT 2.0: An AI Tool for De Novo Drug Design. *J. Chem. Inf Model* **2020**, *60* (12), 5918−5922.

(26) Hongmao, S. *A Practical Guide to Rational Drug Design*; Elsevier: 2015; DOI: 10.1016/C2014-0-02348-9.

(27) Sun, H.; Veith, H.; Xia, M.; Austin, C. P.; Tice, R. R.; Huang, R. Prediction of Cytochrome P450 Profiles of Environmental Chemicals with QSAR Models Built from Drug-like Molecules. *Mol. Inform* **2012**, *31* (11−12), 783−792.

(28) Miyao, T.; Kaneko, H.; Funatsu, K. Inverse QSPR/QSAR Analysis for Chemical Structure Generation (from y to x). *J. Chem. Inf Model* **2016**, *56* (2), 286−299.

(29) Panagiotis-Christos Kotsias, J. A.-P.; Chen, H.; Engkvist, O; Christian, T.; Esben, J. B. Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nature. Machine Intelligence* **2020**, *2*, 254−265.

(30) Cohen, P.; Cross, D.; Janne, P. A. Kinase drug discovery 20 years after imatinib: progress and future directions. *Nat. Rev. Drug Discov* **2021**, *20* (7), 551−569.

(31) Sun, H.; Shah, P.; Nguyen, K.; Yu, K. R.; Kerns, E.; Kabir, M.; Wang, Y.; Xu, X. Predictive models of aqueous solubility of organic compounds built on A large dataset of high integrity. *Bioorg. Med. Chem.* **2019**, *27* (14), 3110−3114.

(32) Lovering, F.; Bikker, J.; Humblet, C. Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.* **2009**, *52* (21), 6752−6756.

(33) Sun, H.; Wang, Y.; Cheff, D. M.; Hall, M. D.; Shen, M. Predictive models for estimating cytotoxicity on the basis of chemical structures. *Bioorg. Med. Chem.* **2020**, *28* (10), 115422.

(34) Kamb, A.; Wee, S.; Lengauer, C. Why is cancer drug discovery so difficult? *Nat. Rev. Drug Discov* **2007**, *6* (2), 115−120.

(35) Daylight. *SMILES*. https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html (accessed 2022-05-26).

(36) Cho, K.; van Merrienboer, B.; Bahdanao, D.; Bengio, Y. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. 2014, arXiv:1409.1259. *arXiv*. https://arxiv.org/abs/1409.1259 (accessed 2022-05-26).

(37) Roskoski, R. *List of FDA approved small molecule protein kinase inhibitors*. 2021. http://www.brimr.org/PKI/PKIs.htm (accessed 2022-05-26).

(38) Muchmore, S. W.; Debe, D. A.; Metz, J. T.; Brown, S. P.; Martin, Y. C.; Hajduk, P. J. Application of belief theory to similarity data fusion for use in analog searching and lead hopping. *J. Chem. Inf Model* **2008**, *48* (5), 941−948.

(39) Carles, F.; Bourg, S.; Meyer, C.; Bonnet, P. PKIDB: A Curated, Annotated and Updated Database of Protein Kinase Inhibitors in Clinical Trials. *Molecules* **2018**, *23* (4), 908.

(40) Aronov, A. M.; McClain, B.; Moody, C. S.; Murcko, M. A. Kinase-likeness and kinase-privileged fragments: toward virtual polypharmacology. *J. Med. Chem.* **2008**, *51* (5), 1214−1222.

(41) Sun, H. A universal molecular descriptor system for prediction of logP, logS, logBB, and absorption. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (2), 748−757.

(42) Sun, H. A naive bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. *J. Med. Chem.* **2005**, *48* (12), 4031−4039. Sun, H.; Huang, R.; Xia, M.; Shahane, S.; Southall, N.; Wang, Y. Prediction of hERG Liability - Using SVM Classification, Bootstrapping and Jackknifing. *Mol. Inform* **2017**, *36* (4), 1600126.

(43) Huang, R.; Zhu, H.; Shinn, P.; Ngan, D.; Ye, L.; Thakur, A.; Grewal, G.; Zhao, T.; Southall, N.; Hall, M. D.; et al. The NCATS Pharmaceutical Collection: a 10-year update. *Drug Discov Today* **2019**, *24* (12), 2341−2349.

(44) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583.

(45) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31* (2), 455−461.

(46) Sun, H.; Wang, Y.; Chen, C. Z.; Xu, M.; Guo, H.; Itkin, M.; Zheng, W.; Shen, M. Identification of SARS-CoV-2 viral entry

inhibitors using machine learning and cell-based pseudotyped particle assay. *Bioorg. Med. Chem.* **2021**, *38*, 116119.

(47) Spall, J. C. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*; Wiley, 2003; DOI: 10.1002/0471722138.

(48) Google. *Seq2Seq*. https://google.github.io/seq2seq/ (accessed 2022-05-26).

(49) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533−536.

(50) Diederik, P.; Kingma, J. B. Adam: A Method for Stochastic Optimization. 2017, *arXiv:1412.698*. https://arxiv.org/abs/1412.6980 (accessed 2022-05-26).

(51) Tensorflow. *An end-to-end open source machine learning platform*. http://tensorflow.org (accessed 2022-05-26).

(52) Chemaxon. *Software solutions and services for chemistry & biology*. http://chemaxon.com (accessed 2022-05-26).

(53) Sun, H.; Veith, H.; Xia, M.; Austin, C. P.; Huang, R. Predictive models for cytochrome p450 isozymes based on quantitative high throughput screening data. *J. Chem. Inf Model* **2011**, *51* (10), 2474−2481.

(54) Fabian, M. A.; Biggs, W. H., 3rd; Treiber, D. K.; Atteridge, C. E.; Azimioara, M. D.; Benedetti, M. G.; Carter, T. A.; Ciceri, P.; Edeen, P. T.; Floyd, M.; et al. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.* **2005**, *23* (3), 329−336.