

this article has been published as: <https://doi.org/10.1021/acs.jcim.0c00517>

***De novo* drug design of targeted chemical libraries based on artificial intelligence and pair based multi-objective optimization**

Domenico Alberga^{1†}, Nicola Gambacorta^{1‡}, Daniela Trisciuzzi¹, Fulvio Ciriaco², Nicola Amoroso¹ and Orazio Nicolotti^{1*}

¹ Dipartimento di Farmacia-Scienze del Farmaco, Università degli Studi di Bari "Aldo Moro", Via E. Orabona, 4, I-70126 Bari, Italy

² Dipartimento di Chimica, Università degli Studi di Bari "Aldo Moro", Via E. Orabona, 4, I-70126 Bari, Italy

† These authors contributed equally.

*** Author to whom correspondence should be addressed; e-mail: orazio.nicolotti@uniba.it; telephone: +39-080-5442551; fax: +39-080-5442230.

Abstract

Artificial intelligence and multi-objective optimization represent promising solutions to bridge chemical and biological landscape by addressing the automated *de novo* design of compounds as a result of a human-like creative process. In the present study, we conceived a novel pair based multi-objective approach implemented in an adapted SMILES generative algorithm based on Recurrent Neural Networks for the automated *de novo* design of new molecules whose overall features are optimized by finding the best trade-offs among relevant physicochemical properties (MW, logP, HBA, HBD) and additional similarity-based constraints biasing specific biological targets. In this respect, we carried out the *de novo* design of chemical libraries targeting Neuraminidase, Acetylcholinesterase and the main protease of Severe Acute Respiratory Syndrome Coronavirus 2. Several quality metrics were employed to assess drug-likeness, chemical feasibility, diversity content and validity. Molecular docking was finally carried out to better evaluate the scoring and posing of the *de novo* generated molecules with respect to X-ray cognate ligands of the corresponding molecular counterparts. Our results indicate that artificial intelligence and multi-objective optimization allow to capture the latent links joining chemical and biological aspects, thus providing easy-to-use options for customizable design strategies, which are especially effective for both lead generation and lead optimization. The algorithm is freely downloadable at <https://github.com/alberdom88/moo-denovo> and all the data are available as Supporting Information.

Introduction

Despite the recent progresses in high throughput screening,¹ the chemical space is still widely unexplored, in particular for prospective drug design.² Importantly, the chemical universe is estimated to encompass by far in excess more than 10^{60} molecules even limiting its exploration only to drug-like space.³ Moreover, since the drug discovery process is typically demanding, slow and expensive,⁴ academic and industrial researchers are discouraged to bet on completely novel chemotypes^{5,6} by exploring a potentially awarding off-patent chemical space, thus preferring to make me-too decorations of well-known molecular bioactive structures biasing specific biological targets.⁷ Among others, a reason behind the low rate of success of finding structurally new interesting drug-like molecules is their inherently multi-objective nature.^{8,9} In particular, drug-like candidates should match a large number of often conflicting features such as water solubility, logP, molecular weight, hydrogen bond donor/acceptor groups, toxic alerts.¹⁰ Moreover, new conceived drug-like molecules should be easy to prepare by chemical synthesis and, last but not least, biologically active.¹¹

Before the onset of computer aided molecular design, drug discovery was mostly addressed by knowledge-based human intuition.¹² A new era in this field was born with the advent of the artificial intelligence methods and the recent wide-spread of deep learning techniques, which are capable not only to uncover hidden patterns from large amounts of data, thus enabling the creation of highly predictive structure-activity models,¹³⁻¹⁵ but can serve as automated generative algorithms to design new compounds with desired properties and selective towards specific biological targets.¹⁶ As a matter of fact, there has been a mushrooming growth of novel *de novo* drug design machine learning algorithms based on diverse techniques. A few examples are the variational autoencoders,¹⁷⁻¹⁹ the generative adversarial networks²⁰⁻²² and the recurrent neural networks.²³⁻²⁶ In general, these algorithms are trained in two steps. The first step employs large high-quality molecular databases (such as ChEMBL²⁷ and ZINC²⁸) to allow models to infer learning rules concerning with chemical representation, usually SMILES^{23,26} or molecular graph,^{29,30} for the automated generation of novel molecules. In particular, the algorithm learns how creating novel chemically valid molecules based on the probability of the next character in SMILES sequences or the node and edge distribution in molecular graphs. As a second step, reinforcement learning methods³¹ are used to speed-up the exploration of the chemical space driving the generation of new samples towards unexplored regions with desired chemical, physical or structural properties.⁵ In some cases, these algorithms

proved to generate drug-like molecules with promising activity towards a protein target.¹⁶ For instance, Merk et al. exploited recurrent neural network-based methods to design novel retinoid X and peroxisome proliferator-activated receptor agonists;^{25,32} Polykovskiy et al. through entangled conditional adversarial autoencoder discovered novel Janus kinase 3 inhibitors;³³ Zhavoronkov et al. via generative tensorial reinforcement learning discovered potent inhibitors of discoidin domain receptor 1.³⁴ However, published methods are generally focused on the automated generation of novel compounds through single objective or weighted sum optimization functions.²⁹ In the present work, we employed a multi-objective optimization algorithm, which is effective in those real-life problems involving the simultaneous optimization of two or more objectives. Notably, multi-objective optimization methods do not require any a priori calibration and are able to detect a family of equivalent solutions *per* run instead than a single solution at a time. As shown in Figure 1, each equivalent solution falls on the Pareto frontier and is non-dominated because another solution does not exist that is better considering all the objectives to optimize.

INSERT FIGURE 1

Only a few applications of multi-objective methods in conjunction with reinforcement learning have been so far described for the *de novo* design.³⁵⁻³⁷ In particular, we herein propose a novel multi-objective optimization approach capable to generate *de novo* targeted chemical libraries whose compounds represent ideal non-dominated solutions for a range of simultaneously pair based optimized features. In particular, we adapted the REINVENT code proposed by Olivecrona et al.²³ in order to enable the pair based multi-objective optimization of several molecular features based on Pareto dominance.³⁸ By employing a tailored fitness function, our reinforcement learning based method is able to drive the automated generation of pair based non-dominated solutions representing chemical structures whose molecular features are customized towards specific biological targets and are constrained in drug-like ranges set by the user. The herein new proposed method was thus tested to accomplish the *de novo* generation of targeted chemical libraries.³⁹ The code is available at GitHub.⁴⁰ Performances were assessed employing quality metrics specifically suitable for evaluating *de novo* designed compounds.^{41,42} Finally, our approach was applied to three real-life case studies relative to the *de novo* drug design of new therapeutically relevant enzymatic inhibitors targeting neuraminidase (NA), acetylcholinesterase (AChE) and the main protease of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), the latter responsible of the COVID-19 global

health emergency.^{43,44} We believe that our method could be of inspiration to medicinal chemists, especially in early stages of the drug discovery process, by generating new patentable chemotypes provided with a wider spectrum of desirable physicochemical and biological properties.

Methods

The SMILES generative algorithm proposed in this work is built by adapting a method originally developed by Olivecrona et al. (REINVENT)²³, which can be briefly summarized as follows. The model consists in a Recurrent Neural Network (RNN) composed of three layers with 512 Gated Recurrent Units⁴⁵ in each layer. The SMILES is tokenized at each single character with the exceptions of Cl, Br and chars included in square brackets that are considered as one token. The RNN is trained by maximum likelihood estimation of the next token in the generated sequence given the prefix of the previous steps. The next character is sampled from the predicted probability distribution with the aim to maximize the likelihood assigned to the correct token. The RNN was trained on a subset of the ChEMBL22 database²⁷ built selecting molecules containing between 10 and 50 heavy atoms and elements H, C, N, O, F, S, Cl, Br for a total of ~1.2 million structures canonicalized with the RDKit package.⁴⁶ The generation of new molecules with specific optimized properties is tackled through a policy-based Reinforcement Learning (RL) algorithm.⁴⁷ Briefly, the probability distributions previously learnt by the trained RNN are used as the initial prior policy. The RL procedure is able to modify the prior policy in order to generate SMILES that maximize a given fitness function S . More details can be found in the original REINVENT paper.²³

The method proposed here exploits the REINVENT algorithm and implements a two-term fitness function S . The first term R plays the role of a drug-like filter to reward molecules within specific property ranges set by the user while increasing penalties are assigned when moving out of the ranges. The second term P employs a pair based multi-objective method. Unlike the classical multi-objective approach considering all the objectives at once, the pair based multi-objective optimization is applied to all the possible pairs of features in order to return a more discriminant overall ranking. This can help to better assess the quality of the *de novo* generated molecules instead of having a large number of equivalent ranked solutions whose amount raises greatly when the number of objectives increases.

In our approach, let $x=(x_1, x_2, \dots, x_N)$ be the vector of the N molecular features to optimize. Let M be the number of molecules of the final targeted chemical library generated by *de novo* design. With the aim to train the algorithm to result targeted chemical libraries with N optimized molecular features, the $S(x)$ fitness function, for each molecule in the targeted chemical library, is defined as follows:

$$S(x) = R(x) + P(x)$$

$$R(x) = \sum_{i=1}^N r(x_i)$$

$$r(x_i) = \begin{cases} e^{-\left(\frac{x_i - \min_i}{\Delta_i}\right)^2} & \text{if } x_i < \min_i \\ 1 & \text{if } \min_i \leq x_i \leq \max_i \\ e^{-\left(\frac{x_i - \max_i}{\Delta_i}\right)^2} & \text{if } x_i > \max_i \end{cases}$$

where \min_i and \max_i are the minimum and maximum acceptable values for a given feature i and $\Delta_i = (\max_i - \min_i)/4$ if $\max_i \neq \min_i$, $\Delta_i = 1$ otherwise.

$$P(x) = \begin{cases} 0 & \text{if } R(x) < N \\ \frac{1}{C(N, 2)} \sum_{i, j \in C} \frac{m - d_{ij}}{m} & \text{if } R(x) = N \end{cases}$$

where d_{ij} is the Pareto dominance of the molecule with respect to the possible pairs of sampled features i and j of the generated targeted chemical library (e. g., the number of molecules of the targeted chemical library dominating a given compound by considering pairs of features i and j); C is the set of all the possible pairs of features under multi-objective optimization that is $C(N, 2) = N(N-1)/2$; m is the number of molecules for which $r(x_i) = 1$ for all the desired features. Based on these assumptions, the $S(x)$ fitness function returns values in the range $[0, N+1]$. In the ideal case, that is $S=N+1$, a *de novo* generated molecule reflects all the considered molecular features in the desired range being all the possible pairs of its optimized features located on the Pareto frontier as optimal non-dominated solutions. Note that the direction of the Pareto frontier is defined by the choice to maximize or minimize the objectives i and j within a given desired range.

For the sake of completeness, the overall dominance with respect to all the objectives at once has been also calculated and the corresponding ranks are reported as Supporting

Information (see files NA.lib1.csv, NA.lib2.csv, AChE.lib1.csv, AChE.lib2.csv and SARS-CoV-2.lib.csv). As shown, overall non-dominated solutions accumulate in the early 5% of the top-scored *de novo* generated molecules.

The algorithm is included in a graphical user interface (GUI), written with pyqt5, freely downloadable at <https://github.com/alberdom88/moo-denovo>. The GUI is able to generate targeted chemical libraries with up to six optimized molecular features that can be selected from a collection of 203 options whose list is available as Supporting Information (File S1). Among these, 201 are molecular descriptors calculated via RDKit⁴⁶ and Moses⁴¹ packages. In addition, two further features were also included. The first accounts for the presence of a user-defined fragment inside the generated molecules. The second is the Tanimoto similarity⁴⁸ calculated using the Morgan fingerprints with radius equal to 2.⁴⁹

As recently reported,^{41,42} the goodness of the targeted chemical libraries is assessed by calculating the following quality metrics: a) validity, which represents the fraction of chemically valid SMILES; b) unicity, which stands for the fraction of unique generated SMILES; c) Internal Diversity (IntDiv)⁵⁰, which accounts for the overall molecular diversity of the targeted chemical library; d) filters, which reflect the fraction of *de novo* generated molecules devoid of medicinal chemistry filters (MCFs)^{51,52} and of pan-assay interference compounds (PAINS) alerts⁵³ and e) Synthetic Accessibility (SA) score, which provides a heuristic estimate of how hard (SA=10) or how easy (SA=1) is the chemical synthesis of a given molecule.⁴¹ SA is averaged on the entire targeted chemical library.

The herein proposed *de novo* drug design algorithm was thus challenged by generating targeted chemical libraries biased for binding NA, AChE and SARS-CoV-2 main protease. The pair based multi-objective optimization algorithm progressed through 3000 cycles, by generating 500 compounds *per* iteration. Finally, the targeted chemical libraries were built sampling 10000 potential inhibitors from the policy that maximizes the average of $S(x)$ fitness values (that is $\langle S(x) \rangle$) and were evaluated according to the quality metrics above described. Furthermore, using the Schrödinger 2019-4 suite,⁵⁴ molecular docking simulations were performed to further evaluate the goodness of new generated inhibitors towards the corresponding desired biological targets. In this respect, the X-ray solved crystal structures of NA, AChE and SARS-CoV-2 protease were retrieved from the Protein Data Bank (PDB) with the entry identifiers equal to 3B7E,⁵⁵ 4EY7⁵⁶ and 6LU7,⁵⁷ respectively. The Protein Preparation Wizard was thus used to revise the X-ray structures, eliminating the water molecules, correcting the protonation states and carrying out energy minimization. All the compounds comprised in the *de novo* targeted chemical libraries

were thus prepared for docking simulations by employing the LigPrep tool⁵⁴ and setting no more than eight stereoisomers *per* molecule to minimize structural complexity as well as avoid too expensive computational costs. Tautomers and ionization states were kept unchanged. All the dockings were performed employing Glide⁵⁴ standard precision with default settings by automatically centering the grid boxes on the co-crystallized cognate ligands of the three reference proteins. The reliability of docking simulation protocols was preliminary challenged by computing the root mean square deviation (RMSD) values (see Figure S1 of Supporting Information). The docking results were analyzed comparing posing and scoring of the co-crystallized cognate ligands with those experienced from the new generated molecules. In this respect, three terms were mostly considered: the docking scores, the chance of interacting with key binding site residues and the $S(x)$ fitness values.

Results

Our automated algorithm for *de novo* design was challenged by generating targeted chemical libraries³⁹ likely to bind NA, AChE, and novel SARS-CoV-2 main protease. More specifically two targeted chemical libraries were generated for NA (i.e., NA.lib1 and NA.lib2) and for AChE (i.e., AChE.lib1 and AChE.lib2) and one targeted chemical library (i.e., SARS-CoV-2.lib) was designed for SARS-CoV-2 main protease. Each targeted chemical library was obtained by pair based multi-objective optimization of several relevant physicochemical properties and, eventually, by considering other medicinal chemistry inspired constraints such as molecular similarity. As far as the three case studies are concerned, the learning curves indicating the progress of the average of the $S(x)$ fitness values is shown in Figure 2 while a synoptic assessment of *de novo* designed targeted chemical libraries is provided in Table 1 by reporting the calculated quality metrics.

INSERT FIGURE 2

INSERT TABLE 1

Neuraminidase case study


NAs are glycoside hydrolase enzymes with a fundamental role in the spread of the virus, especially in the late stages of infection.⁵⁵ They are classified as exosialidases and are capable of cleaving glycosidic bonds between sialic acid and sugar.⁵⁵ The virus employs this mechanism to detach itself from the host cell after the infection and, thus, to spread

out. NA inhibitors are a well-known class of drugs used against influenza A virus,⁵⁸ and the discovery of new potent biologically active agents is considerably interesting from a pharmaceutical perspective. To this purpose, two targeted chemical libraries of 10000 potential inhibitors each were designed. The first targeted chemical library (NA.lib1 is included in Supporting Information as File S2) was designed based on the wealth of physicochemical information taken from a benchmark pool comprising all the entries (that are 218) available from ChEMBL v.25 provided with $IC_{50} < 1 \mu M$ towards NA (target referenced as ChEMBL2051). The *de novo* design of NA.lib1 was addressed by considering the variation within the benchmark pool of four easy to interpret molecular descriptors that are the molecular weight (MW), the logP, the number of hydrogen bond donor (HBD) atoms and the number of hydrogen bond acceptor (HBA) atoms. In our approach the *de novo* designed compounds are optimized in a range defined within one standard deviation around the mean value of the selected features computed for the benchmark pool as shown in Table S1 of the Supporting Information. On the other hand, the *de novo* design was addressed to generate compounds including one and only one aliphatic ring like Zanamvir.

As shown in Figure 2, the algorithm reaches the convergence after approximately 500 iterations. The point corresponding to the maximum value of $\langle S(x) \rangle$ average fitness value (blue line of Figure 2) is thus used as a generative model to create NA.lib1 consisting of 10000 potential inhibitors. As reported in Table 1, the 99.5% of the generated molecules are valid and 98.6% are unique. Importantly, only the 62.0% pass typical structural alerts filters and this should alert users when prioritizing *de novo* designed compounds for further testing. Finally, the generated targeted chemical library NA.lib1 owns good internal diversity equal to 73.3% and a fair synthetic accessibility score (2.522 \pm 0.361). Figure 3 depicts the distributions of the molecular descriptors selected for pair based multi-objective optimization of NA.lib1 on the left-hand side and its average pair based Pareto dominance on the right-hand side. In particular, the 93.4% of the molecules in the library owns descriptors in the desired range. Additional details are reported in Figure S2 of Supporting Information.

INSERT FIGURE 3

The second targeted chemical library (NA.lib2 is included in Supporting Information as File S3) was instead designed to generate molecules whose structures and physicochemical properties are constrained to those of Zanamivir. To this end, the *de novo* design

progressed by maximizing the Tanimoto molecular similarity to the Zanamivir with a cutoff >0.3 and allowing deviations shown in Table S2 of Supporting Information as far as MW and logP are concerned. As above discussed, NA.lib2 included 10000 molecules created from the generative model taken from the maximum value of the average fitness value $\langle S(x) \rangle$ (red line of Figure 2). As reported in Table 1, 99.8% of the generated molecules are valid and 95.7% are unique. As expected, a drop of the internal diversity, that is now equal to 66.9%, was observed with respect to the NA.lib1 but the molecules passing the filters increases to 85.2%. The library shows an average synthetic accessibility of 4.700  0.233, a value complying that computed for Zanamivir (that is 4.287). A comprehensive view of feature distribution is shown in Figure 3. Additional details are reported in Figure S2 and Figure S3 of Supporting Information.

INSERT TABLE 2

Representative examples taken from NA.lib1 and NA.lib2 are shown in Table 2. The *de novo* drug design strategy employed to generate NA.lib1 allowed to build new chemotypes including piperidine, piperazine and phenol rings. The *de novo* drug design strategy employed to generate NA.lib2 resulted in the scaffold hopping of the dihydropyran ring of Zanamivir, replaced for instance by tetrahydropyridine and cyclohexene cores. All these engineered structures are provided with polar substituents, such as aminic, amidic, sulfamoyl, guanidinium, hydroxyl or carboxyl functional groups, potentially capable to reproduce the molecular interactions accomplished by Zanamivir. For the sake of interpretation, molecular docking analyses were thus carried out to gain insights about molecular interactions established by *de novo* designed compounds at the binding site of NA enzyme compared to those observed for Zanamivir whose observed docking score of -7.002 kcal/mol is mainly due to the number of hydrogen bond (HB) interactions engaged with R371, W178 and R152 at the NA binding site. In this respect, we reported two applicative examples taken from NA.lib1 and NA.lib2, respectively. NA.lib1_02 *de novo* designed compound returned a docking score equal to -6.920 kcal/mol being its sulfamoyl and aminic groups involved in HBs with the side chains of R118, E277, R292, and R371 as well as with the backbone of W178 (see Figure 4a). NA.lib2_01 returned a docking score equal -7.892 kcal/mol with the guanidinium group forming HBs with the side chains of E276 and E277, the amidic group interacting with side chains of R371 and R118, and the acetoamide substituents making HBs with R152 and D151 (see Figure 4b). Worthy of

mention, both NA.lib1_02 and NA.lib2_01 showed a posing and a scoring comparable to Zanamivir by experiencing very similar molecular interactions at the NA binding site.

INSERT FIGURE 4

Acetylcholinesterase case study

AChE is an enzyme located in the post-synaptic membrane of cholinergic neurons and catalyzes the hydrolysis reaction of the acetylcholine in choline and acetic acid. Excessively decreased levels of acetylcholine are hallmarks of the Alzheimer onset.⁵⁹ This is the main reason why AChE inhibition is a widely studied mechanism for the symptomatic palliation of neurodegenerative diseases.⁶⁰ Donepezil, a selective AChE dual binding site inhibitor,⁶¹ was used as reference for the automated generation of a targeted chemical library.








Again, two different strategies were adopted and, thus, two targeted chemical libraries of 10000 potential inhibitors each were designed. The first targeted chemical library (AChE.lib1 is included in Supporting Information as File S4) was designed based on five easy to interpret molecular descriptors that are MW, logP, HBD, HBA and the number of rings (nR). In particular the nR descriptor was selected considering the presence of four rings (two of which fused) in the structure of Donepezil. The minimum and maximum values of these descriptors were selected retrieving all the entries (that are 1800) available from ChEMBL v.25 provided with $IC_{50} < 1 \mu M$ towards AChE (referenced as ChEMBL220). The *de novo* designed compounds are optimized in a range defined within one standard deviation around the mean value of the selected features computed for the benchmark pool as shown in Table S3 of the Supporting Information. As shown in Figure 1, the algorithm reaches the convergence after approximately 500 iterations. The point corresponding to the maximum value of $\langle S(x) \rangle$ average fitness value (yellow line of Figure 1) is thus used as a generative model to create AChE.lib1 consisting of 10000 potential inhibitors. As reported in Table 1, 99.3% of the generated molecules are valid and 94.5% are unique. Importantly, 92.0% pass typical structural alerts filters. Interestingly, the generated targeted chemical library AChE.lib1 owns good internal diversity equal to 71.5% and a fair synthetic accessibility score (1.942 ± 0.192). Figure 5 depicts the distributions of the molecular descriptors selected for pair based multi-objective optimization of AChE.lib1 on the left-hand side and its average Pareto dominance on the right-hand side. In

particular, 96.1% of the molecules in the library own descriptors in the desired range. Additional details are reported in Figure S2 of Supporting Information.

INSERT FIGURE 5

The second targeted chemical library (AChE.lib2 is included in Supporting Information as File S5) was instead designed to generate molecules whose structures and physicochemical properties are constrained to those of Donepezil. To this end, the *de novo* design progressed by maximizing the Tanimoto molecular similarity to the Donepezil with a cutoff >0.3 and allowing deviations shown in Table S4 of Supporting Information as far as MW and logP are concerned. As above discussed, AChE.lib2 included 10000 molecules created from the generative model taken from the maximum $\langle S(x) \rangle$ average fitness value (green line of Figure 2). As reported in Table 1, 99.8% of the generated molecules are valid and 95.9% are unique. A drop of the internal diversity, that is now equal to 62.5%, as well as of the molecules passing the filters, that is now equal to 84.1%, was observed with respect to AChE.lib1. The library shows an average synthetic accessibility of 2.165 \pm 0.256 complying that computed for Donepezil (that is 2.682). A comprehensive view is shown in Figure 5. Additional details are reported in Figure S2 of Supporting Information.


Representative examples taken from AChE.lib1 and AChE.lib2 are shown in Table 2. As far as AChE.lib1 is concerned, the algorithm generated potential inhibitors with no less than three slightly decorated aromatic rings joined by proper length ether or keto bridge to ensure the sampling of the catalytic (CAS) and peripheral anionic (PAS) site of AChE.⁶² On the other hand, AChE.lib2 comprised compounds including the phenylpiperidine scaffold as Donepezil.

Hence, molecular docking simulations have been employed to inspect the molecular interactions of the *de novo* generated potential AChE inhibitors. As depicted in the left-hand side of Figure 6, AChE.lib1_02 molecule, taken from AChE.lib1, can experience    interactions with W286 at PAS and with F338 and can make HB with the backbone of F295. On the right-hand side of Figure 6, AChE.lib2_04 molecule, taken from AChE.lib2, can engage    interactions with W286 at PAS through the 3-methoxyphenyl arm and with W86 at CAS through the benzylpiperidine moiety and HB with F295. Moreover, the protonated nitrogen atom of the piperidine ring is able to engage a cation- interaction with W286. Interestingly, these interactions are also visited by

Donepezil showing a docking score value equal to -12.640 kcal/mol. AChE.lib1_02 and AChE.lib2_04 returned docking score values equal to -11.045 kcal/mol and -12.703 kcal/mol, respectively.

INSERT FIGURE 6

SARS-CoV-2 main protease case study

The dramatic spread of Covid-19 due to infective agent SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) has undoubtedly led researchers from all over the world to face this new emergency with every possible resource.^{43,44} On February 2020, the first X-ray solved structure of the protease of the new coronavirus has been released in the PDB⁵⁷ and represents an extremely important milestone for developing new and effective drug therapies. After transcription, the viral mRNA penetrates the cytoplasm and uses the cellular mechanism for the proteins production. The newly formed polypeptide chains are cleaved into smaller fragments and used by the virus for its maturation.⁶³ These cleavages are carried out by proteases. Specifically, this is classified as cysteine-histidine protease (H41-C145).⁶⁴ As far as this case study is concerned, there is no entry available in the ChEMBL repository since the target is so far still unknown. On this premise, a similarity based *de novo* design strategy was thus employed. As a reference for the similarity *de novo* design, the co-crystallized ligand (reported as N3 in the PDB entry coded as 6LU7) covalently bonded to the SARS-CoV-2 main protease was selected. As show in Table S5 of the Supporting Information, a targeted chemical library (SARS-CoV-2.lib is included in Supporting Information as File S6) was generated by pair based multi-objective optimization of the following features: the Tanimoto molecular similarity, MW and logP ranging in the intervals shown in Table S5 of Supporting Information. As far as this case study is concerned, the algorithm converged slowly after 1000 iterations (black line of Figure 2). This is likely due to the higher structural complexity of N3. Again, 10000 molecules were generated using the generative model corresponding to the maximum of $\langle S(x) \rangle$ average fitness value. As shown in Table 1, SARS-CoV-2.lib shows good values of quality metrics being validity equal to 99.9%, unicity equal to 91.0% and ability to pass structural alert filters equal to 77.3%. The internal diversity (that is 53.9%) is however lower compared to the cases of studies previously discussed. Again, this can be explained with the large structural complexity of N3 that forces the algorithm to generate molecules to some extent provided with a reduced structural variability. The library shows an average synthetic accessibility of 3.645  0.281 that is however slightly lower compared to that

computed for N3 (that is 4.701). Figure 7 depicts the distributions of the molecular descriptors selected for pair based multi-objective optimization of SARS-CoV-2.lib on the left-hand side and its average Pareto dominance on the right-hand side.

INSERT FIGURE 7

As far as SARS-CoV-2.lib is concerned, the generation of new molecules has been driven by employing a similarity-based strategy. For ease of discussion, four potential inhibitors built by automated *de novo* design were reported as examples in Table 2. The algorithm was able to generate peptide-like molecules, whose backbone included at least three peptide bonds and whose side chains explored different combinations of residues such as glycine, glutamate, aspartate, glutamine or leucine.

Molecular docking studies were thus finally performed on SARS-CoV-2.lib by using as a biological target the SARS-CoV-2 main protease (PDB entry 6LU7, the only 3D crystal structure available at the time of writing). In this particular case study, a protocol of covalent docking could have been an option except that it is very time-consuming and thus not suitable for screening large numbers of ligands. As shown in Figure 8, the *de novo* generated peptide backbone is essential for engaging key HB interactions with E166 and Q189 that actually were also observed in the N3 co-crystallized inhibitor provided with a docking score value equal to -11.213 kcal/mol. Interestingly, SARS-CoV-2.lib_03 was also able to form additional HB with C145, a key residue for the catalytic function of the SARS-CoV-2 main protease, thus returning a docking score value of -9.046 kcal/mol.

INSERT FIGURE 8

Final remarks

Based on the obtained results, we can conclude that artificial intelligence and multi-objective optimization provided a transparent framework for customizable design strategies.^{65,66} Our approach demonstrated to be particularly suited for both lead generation and lead optimization phases. In this respect, lead generation could mostly be pursued based on a merely data driven approach optimizing physicochemical properties such as MW, logP, the number of HBA and HBD. In this regard, we observed that *de novo* generated compounds (i.e. NA.lib1 and AChE.lib1) can sample new chemotypes exploring a broader and hopefully off-patent chemical space compared to a given reference domain. This approach is indeed more challenging and helpful to fuel new ideas for a target

oriented design. On the other hand, lead optimization is instead mostly addressed by adding further constraints, such as molecular similarity thresholds or the inclusion of particular privileged scaffolds, which normally reflect specific user-dependent options (i.e. NA.lib2, AChE.lib2 and SARS-CoV-2.lib). This approach is indeed more conservative but likely of more practical use and, to some extent, less prone to late stage failures. These considerations were also supported by a ligand-based drug target prediction exercise carried out by employing the Multi-fingerprint Similarity Search algorithm (MuSSel)^{67,68} which is available as a free web platform at <http://mussel.uniba.it:5000/prediction.html>. For instance, considering the top-5 targets predicted by MuSSel for each query compound, we observed that the Neuroaminidase (referenced as ChEMBL2051) and Acetylcholinesterase (referenced as ChEMBL220) were predicted as the protein targets for about 95% and about 68% compounds of NA.lib2 and AChE.lib2, respectively. On the other hand, we observed that Monoamine oxidase B (referenced as ChEMBL2039) and Monoamine oxidase A (referenced as ChEMBL1951) were predicted as the protein targets for about 31% and about 16% compounds of AChE.lib1, according to their potential action as multi-target therapeutics for neurodegenerative disorders.⁶⁰ As far as NA.lib1 is concerned, it was instead difficult to derive causative relationships with the predicted protein targets. A synopsis of all the gathered drug target predictions is shown in Table S6 of Supporting Information. The overall quality of the *de novo* designed target libraries is however ensured by the quality metrics indicating that the new generated compounds were always in optimal ranges as far as the correctness of SMILES notations and the occurrence of duplicates, the level of internal dissimilarity, the compliance with drug-likeness filters and the chemical feasibility is concerned. Furthermore, molecular docking was employed to better assess the biological potential of the new generated compounds by explicitly comparing their posing and scoring with those experimentally observed for cognate ligands co-crystallized at the binding sites of target proteins. This analysis enabled to appreciate that the *de novo* designed targeted chemical libraries contain molecules experiencing similar binding modes as they engaged specific interactions with key target protein residues. Overall, the goodness of molecular docking as a retrospective validation option is also supported by the comparison of the distribution of docking scores of the 218 pool compounds (referenced as ChEMBL2051 and provided with $IC_{50} < 1\mu M$ towards NA) and targeted libraries NA.lib1 and NA.lib2 as well as of the 1800 pool compounds (referenced as ChEMBL220 and provided with $IC_{50} < 1\mu M$ towards AChE) and targeted libraries AChE.lib1 and AChE.lib2 (see Figure S4 of Supporting Information). In a

continuing analysis aimed at providing users with a prioritization list of best compounds for synthesis and testing, we sorted the *de novo* designed libraries according to the calculated values of Ligand Efficiency (LE)⁶⁹ which is a universally accepted indicator of compound quality for prospective drug design. For the sake of comparison, the values of LE of the cognate ligands were used as a reference.

INSERT FIGURE 9

As shown in Figure 9, best LE values accumulated in the early 0.12% and 10.38% for NA.lib1 and NA.lib2, in the early 0.02% and 0.03% for AChE.lib1 and AChE.lib2, in the early 1.29% for SARS-CoV-2.lib. While it is not wise to make any beforehand speculation about these observed trends whose rationale may be case-by-case dependent, we can conclude that these are the more meaningful fractions of the *de novo* generated targeted chemical libraries. Importantly, these fractions contained those *de novo* designed compounds better awarding molecular interactions at the active sites of their biological counterparts while keeping as low as possible the structural complexity, which was comparable or even lower than that of X-ray solved cognate ligands. What descend from those observations could be of utmost importance for addressing well-informed drug design strategies.

Finally, this new proposed method makes an important step forward considering the scenario of the recent literature which already includes a number of generative *de novo* design algorithms.^{17–26,29,30,36} In this respect, we mostly focused on crafting a novel pair-based multi-objective strategy that, in conjunction with a reinforcement learning framework, allows to guide the creative generation of drugs owning multiple simultaneously optimized features based on the Pareto optimality philosophy. Indeed, this new approach allowed to enhance the quality content of the results compared to recently published methods, which enable the automated generation of novel compounds but generally through single objective or weighted sum methods,^{23,26,29,36} which require human intervention for coefficient calibration with steps that are annoying, frustrating and time consuming. Satisfactorily, our approach can be used as it is for the fast generation of *de novo* targeted chemical libraries whose features fall in a given desired ranges giving the users a pool of equivalent non-dominated solutions irrespective of calibration. Finally, note that the herein proposed fitness function can be easily implemented as a reward option in any already published reinforcement learning based molecular *de novo* design algorithm. We chose to adapt the REINVENT algorithm²³ because it is relatively easy to modify

compared to others already published; it is provided with a simple and well annotated source code; it requires limited computational resources and shows still promising performances.

Conclusions

The application of artificial intelligence and multi-objective optimization in computer assisted drug discovery have paved the way to unprecedented chances for highly relevant tasks such as structure activity relationships, target prediction, lead generation and optimization, experimental design. All these activities are expected to shorten cycle-times for the identification of new bioactive compounds for both industry and academia. In this study, several aspects of artificial intelligence and multi-objective optimization for the *de novo* drug design of targeted combinatorial libraries were investigated with the intention to support real-life project workflows. As a matter of fact, three practical case studies of therapeutic relevance have been widely discussed. In particular, this study showed as complementing artificial intelligence with pair based multi-objective optimization is effective in driving the *de novo* design of target chemical libraries whose compounds are the result of a creative process based on molecular features taken by specific portions of the chemical space, which were interfaced with particular biological targets. Last but not least, our ultimate goal is not to replace bench chemistry activities but rather to inspire the experimental work with low cost ideas.

Supporting Information

Features selected for pair based multi-objective optimization generation of the targeted chemical library NA.lib1, NA.lib2, AChE.lib1, AChE.lib2 and SARS-CoV-2.lib (Table S1-S5). Drug target predictions carried out by using the multi-fingerprint similarity search MuSSel platform (Table S6). Overlap of X-ray solved and top-scored docking poses for Zanamivir, Donepezil and N3 (Figure S1). List of all the 203 molecular features that can be optimized via the algorithm (features.txt). Targeted chemical library NA.lib1 (NA.lib1.csv). Targeted chemical library NA.lib2 (NA.lib2.csv). Targeted chemical library AChE.lib1 (AChE.lib1.csv). Targeted chemical library AChE.lib2 (AChE.lib2.csv). Targeted chemical library SARS-CoV-2.lib (SARS-CoV-2.lib.csv). The algorithm described in this paper is freely downloadable at <https://github.com/alberdom88/moo-denovo>.

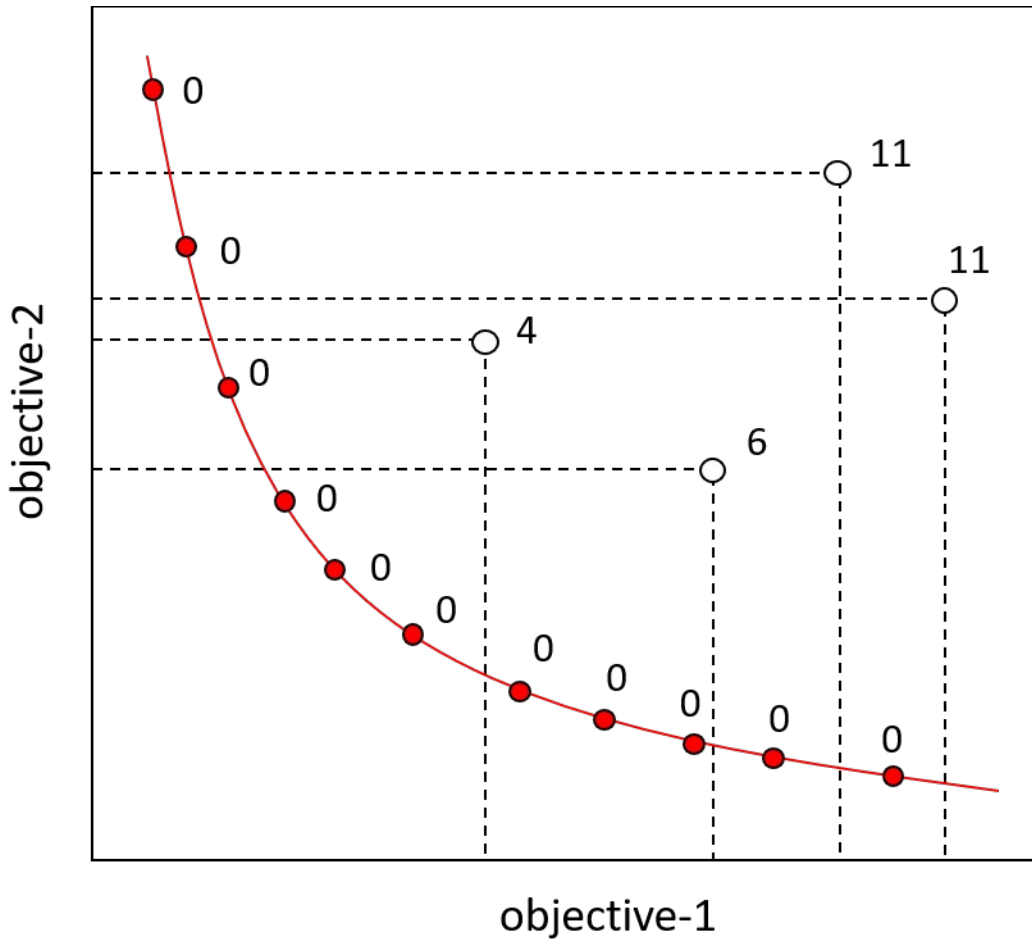


Figure 1. The solid red circles are non-dominated solutions and fall on the Pareto frontier, colored in red. Dominated solutions are shown as unfilled circles and are ranked according to the number of times they are dominated. Non-dominated solutions are given rank zero and the dominated solutions are given ranks as shown.

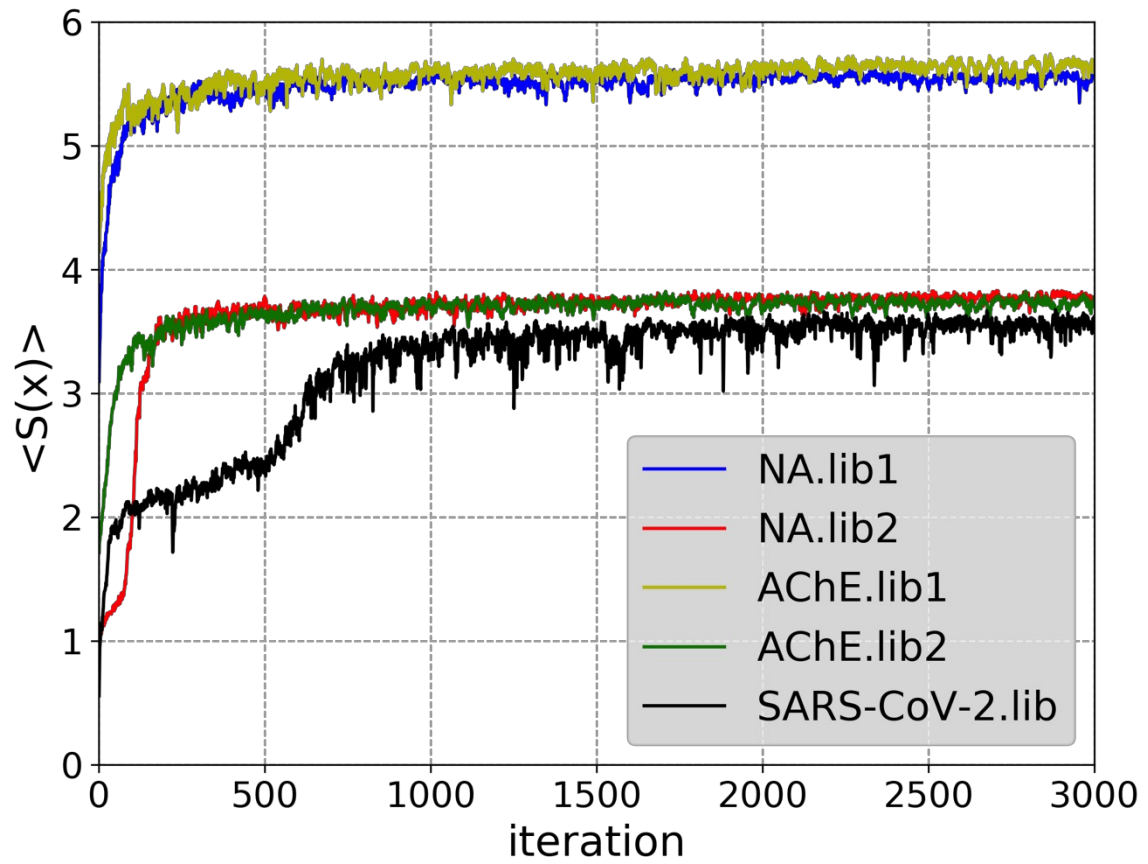


Figure 2. Average value of the $S(x)$ fitness function computed at each iteration of the pair based multi-objective optimization algorithm for each case study.

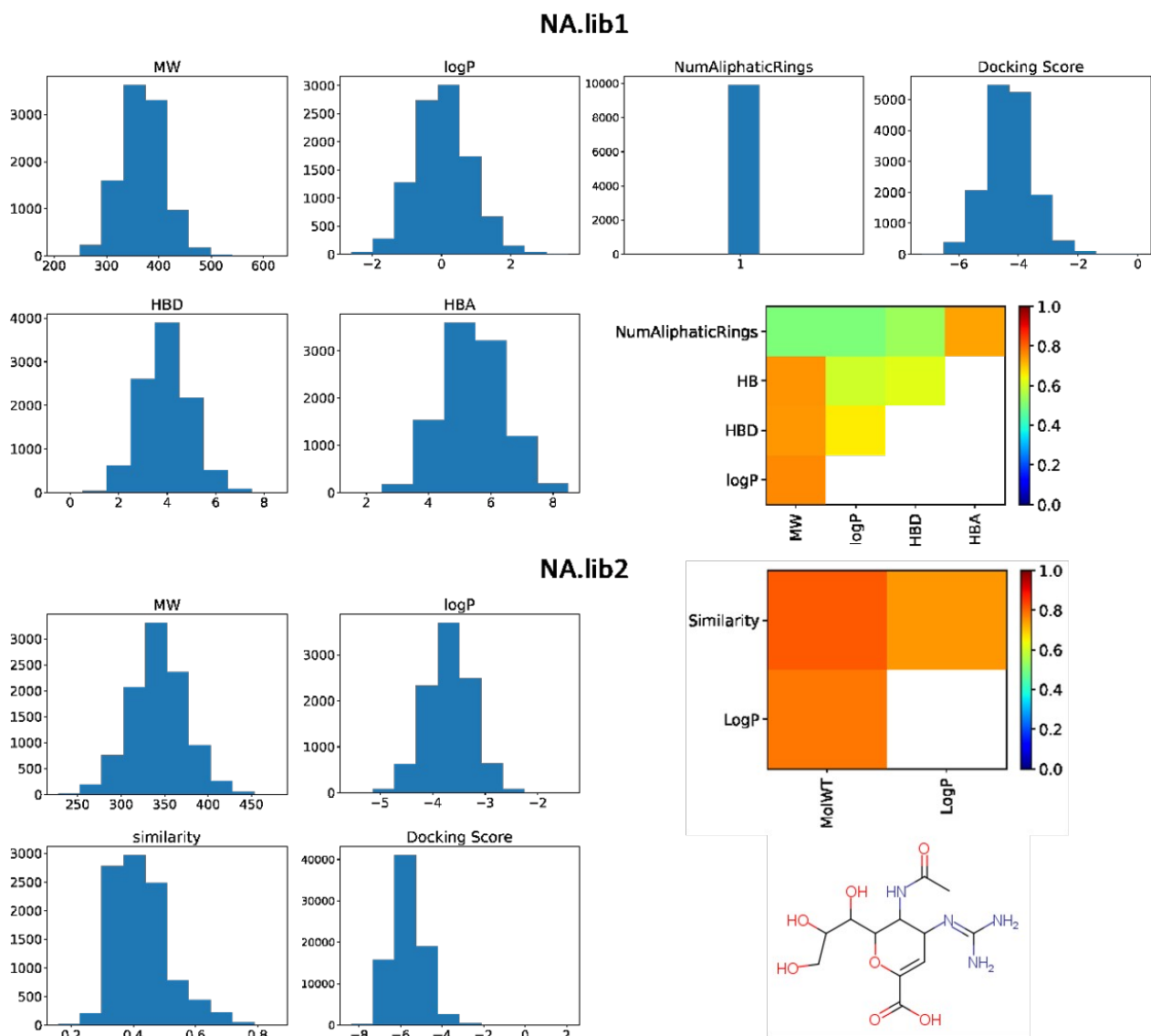


Figure 3. Distribution of the optimized features and docking scores of the 10000 compounds and heat map showing the average pair based Pareto dominance concerned with the targeted chemical libraries NA.lib1 and NA.lib2. On the bottom right corner, the structure of Zanamivir is reported.

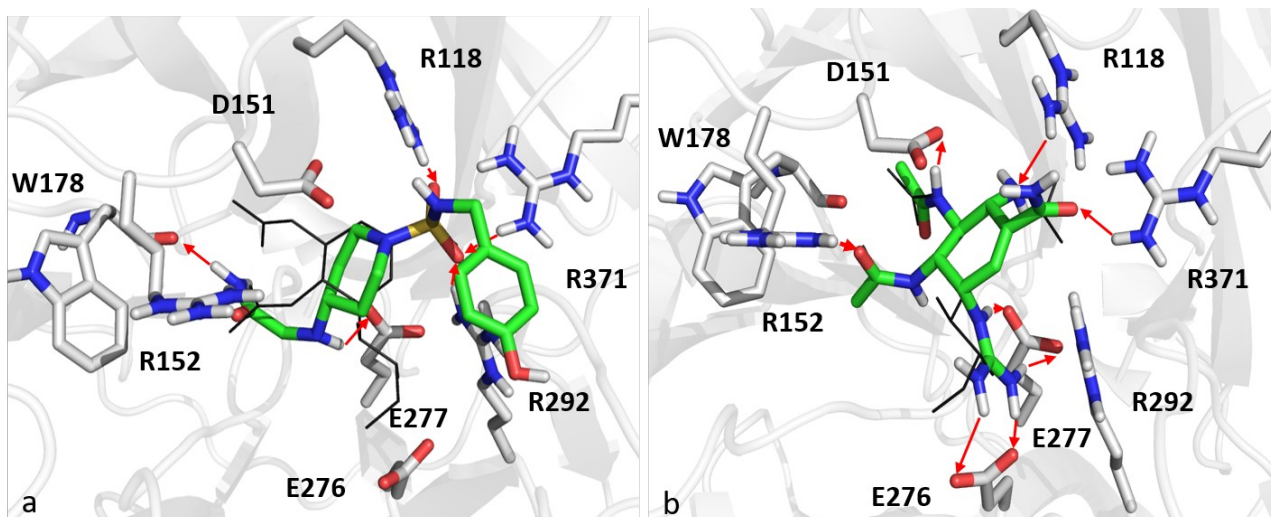


Figure 4. Panels (a) and (b) report molecular interactions between NA (PDB entry 3B7E) and *de novo* generated compounds taken from the first (NA.lib1_02) and second (NA.lib2_01) targeted chemical libraries, respectively. Ligands and the target are rendered in green sticks and gray cartoon, respectively. Zanamivir is also depicted in black wireframe. Red arrows indicate hydrogen bonds. For the sake of clarity, only polar hydrogen atoms are shown.

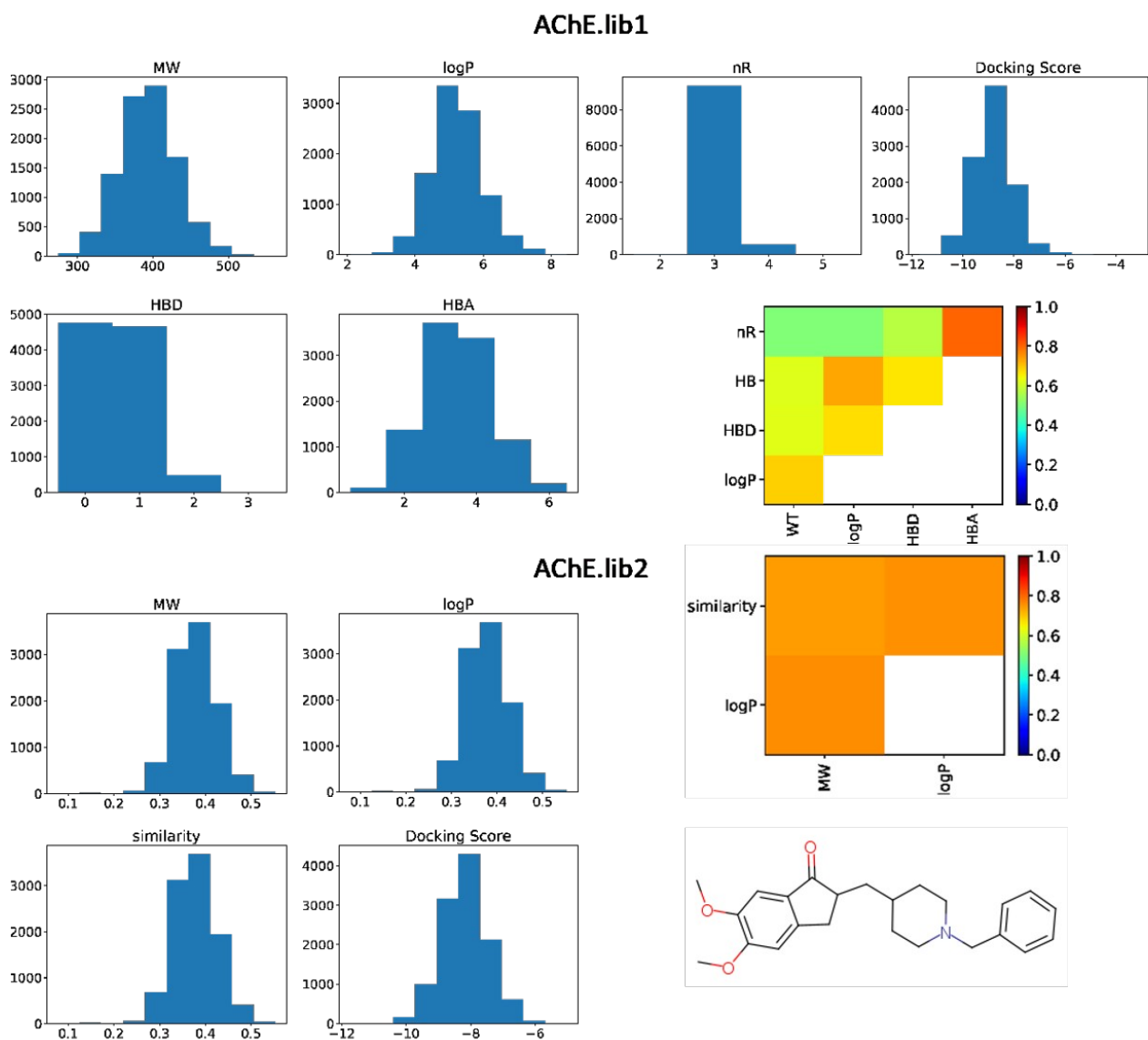


Figure 5. Distribution of the optimized features and docking scores of the 10000 compounds and heat map showing the average pair based Pareto dominance concerned with the targeted chemical libraries AChE.lib1 and AChE.lib2. On the bottom right corner, the structure of Donepezil is reported.

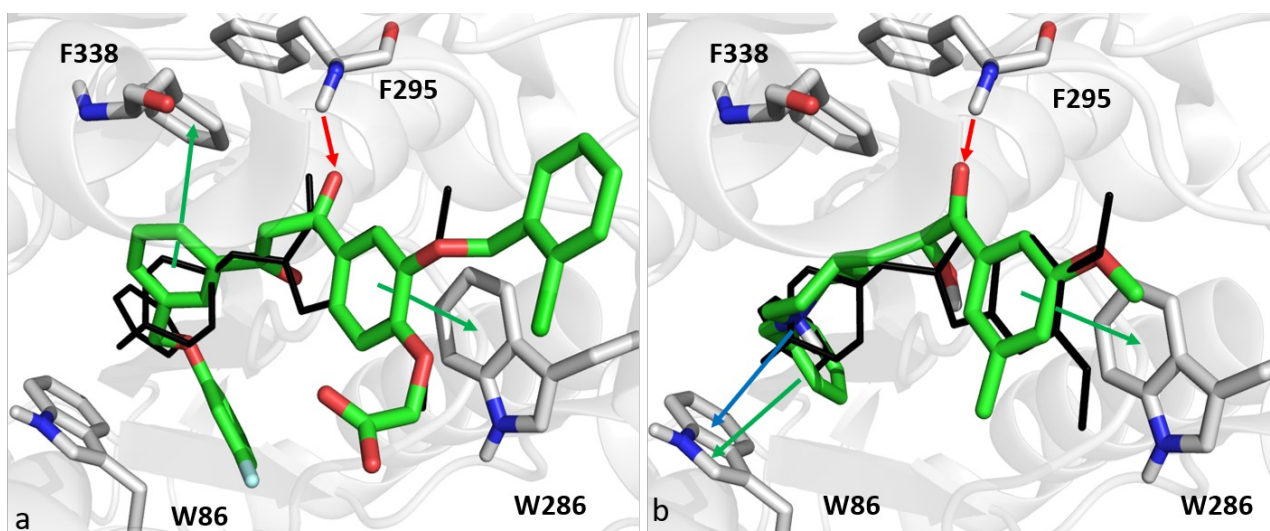


Figure 6. Panels (a) and (b) report molecular interactions between AChE (PDB entry 4EY7) and *de novo* generated compounds taken from the first (AChE.lib1_02) and second (AChE.lib2_04) targeted chemical libraries, respectively. Ligands and the target are rendered in green sticks and gray cartoon, respectively. Donepezil is also depicted in black wireframe. Red, green and blue arrows indicate hydrogen bonds, π - π and cation- π interactions, respectively. For the sake of clarity, only polar hydrogen atoms are shown.

SARS-CoV-2.lib

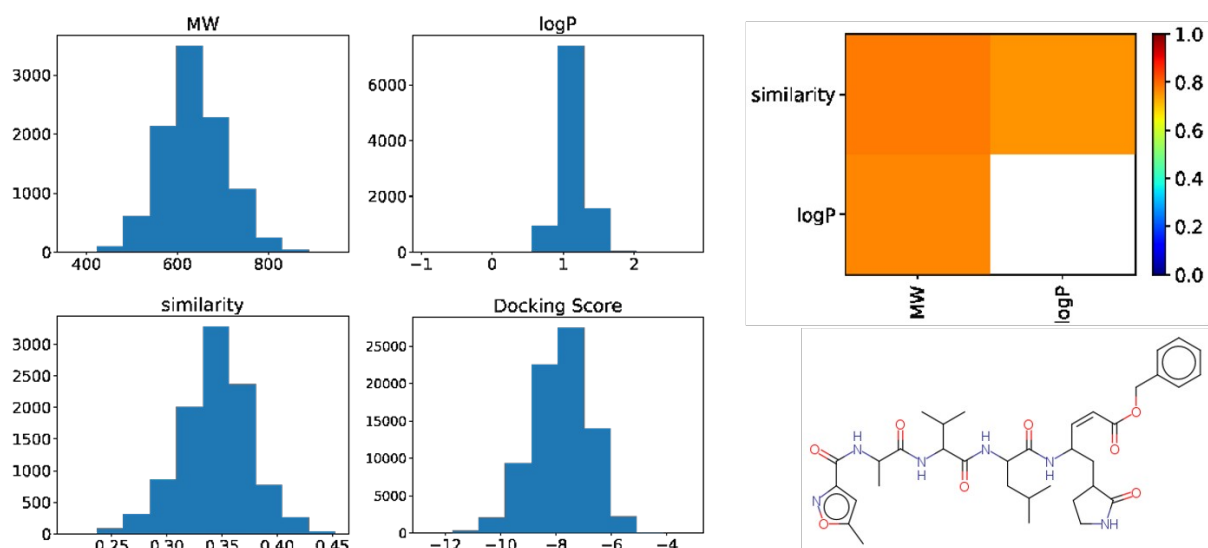


Figure 7. Distribution of the optimized features and docking scores of the 10000 compounds and heat map showing the average pair based Pareto dominance concerned with the targeted chemical library SARS-CoV-2.lib. On the bottom right corner, the structure of N3 inhibitor is reported.

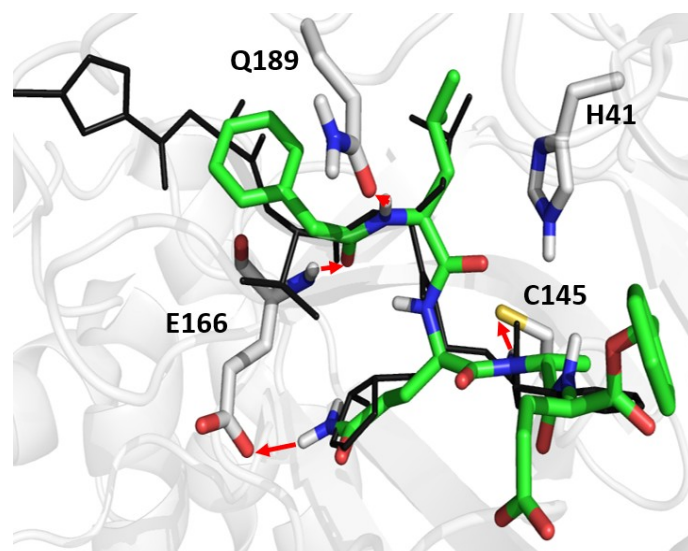


Figure 8. Molecular interactions between SARS-CoV-2 main protease (PDB entry 6LU7) and *de novo* generated compound SARS-CoV-2.lib_03. Ligand and target are rendered in green sticks and gray cartoon, respectively. N3 is also depicted in black wireframe. Red arrows indicate hydrogen bonds. For the sake of clarity, only polar hydrogen atoms are shown.

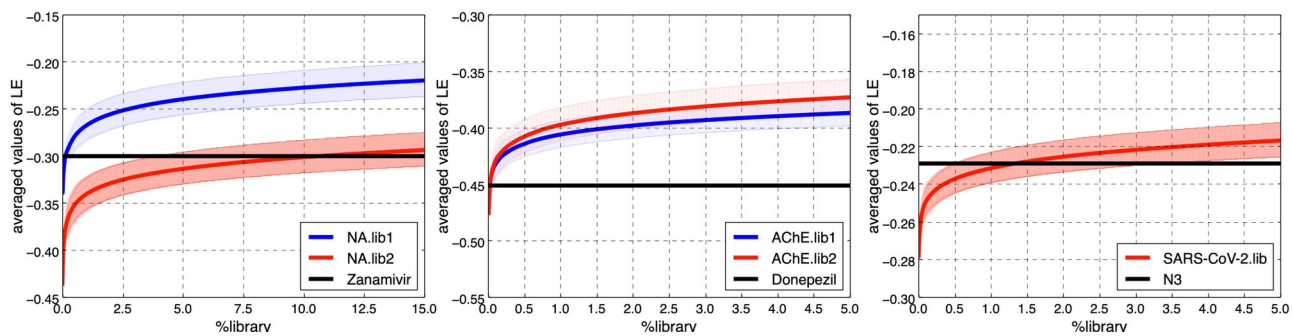
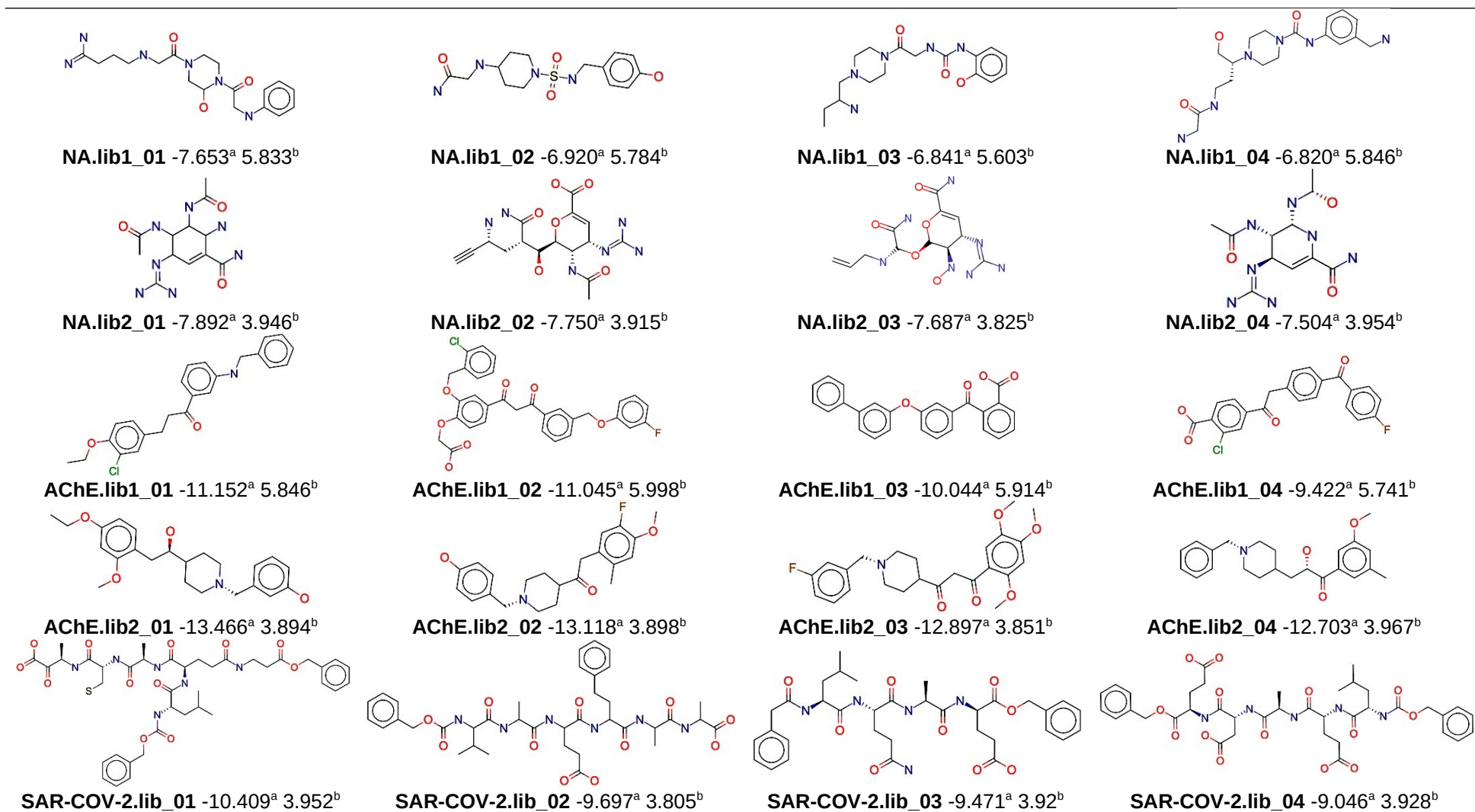


Figure 9. Red and blue solid lines as well as shaded areas indicate the average and the standard deviation of LE values, respectively, computed varying the considered percentage of each targeted library sorted at the increase of LE. The dark solid line represents the LE value for the X-ray solved cognate ligands used as reference for each case study.

Targeted chemical library	Validity	Unicity	IntDiv	Filters	SA
NA.lib1	0.995	0.986	0.733	0.620	2.522 ± 0.361
NA.lib2	0.998	0.957	0.669	0.852	4.700 ± 0.233
AChE.lib1	0.993	0.945	0.715	0.920	1.942 ± 0.192
AChE.lib2	0.998	0.959	0.625	0.841	2.165 ± 0.256
SARS-CoV-2.lib	0.999	0.910	0.539	0.773	3.645 ± 0.281

Table 1. Quality metrics of the targeted chemical libraries generated by *de novo* drug design. Note that SA values are reported as mean along with standard deviation.

Table 2. Representative examples of potential inhibitors generated through automated *de novo* drug design. The superscript letters *a* and *b* indicate docking scores (kcal/mol) and *S(x)* fitness values.



REFERENCES

- (1) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V. S.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S. Impact of High-Throughput Screening in Biomedical Research. *Nat Rev Drug Discov* **2011**, *10*, 188–195.
- (2) Reymond, J.-L.; Deursen, R. van; Blum, L. C.; Ruddigkeit, L. Chemical Space as a Source for New Drugs. *Med. Chem. Commun.* **2010**, *1*, 30–38.
- (3) Hann, M. M.; Oprea, T. I. Pursuing the Leadlikeness Concept in Pharmaceutical Research. *Current Opinion in Chemical Biology* **2004**, *8*, 255–263.
- (4) Waring, M. J.; Arrowsmith, J.; Leach, A. R.; Leeson, P. D.; Mandrell, S.; Owen, R. M.; Pairaudeau, G.; Pennie, W. D.; Pickett, S. D.; Wang, J.; Wallace, O.; Weir, A. An Analysis of the Attrition of Drug Candidates from Four Major Pharmaceutical Companies. *Nat. Rev. Drug Discovery* **2015**, *14*, 475–486.
- (5) Medina-Franco, J. L.; Martinez-Mayorga, K.; Meurice, N. Balancing Novelty with Confined Chemical Space in Modern Drug Discovery. *Expert Opin. Drug Discovery* **2014**, *9*, 151–165.
- (6) Mangiatordi, G. F.; Trisciuzzi, D.; Alberga, D.; Denora, N.; Iacobazzi, R. M.; Gadaleta, D.; Catto, M.; Nicolotti, O. Novel Chemotypes Targeting Tubulin at the Colchicine Binding Site and Unbiasing P-Glycoprotein. *Eur. J. Med. Chem.* **2017**, *139*, 792–803.
- (7) Singh, N.; Chaput, L.; Villoutreix, B. O. Virtual Screening Web Servers: Designing Chemical Probes and Drug Candidates in the Cyberspace. *Brief Bioinform.* **2020**, 1-29.
- (8) Nicolotti, O.; Gillet, V. J.; Fleming, P. J.; Green, D. V. S. Multiobjective Optimization in Quantitative Structure–Activity Relationships: Deriving Accurate and Interpretable QSARs. *J. Med. Chem.* **2002**, *45*, 5069–5080.
- (9) Nicolotti, O.; Giangreco, I.; Miscioscia, T. F.; Carotti, A. Improving Quantitative Structure–Activity Relationships through Multiobjective Optimization. *J. Chem. Inf. Model.* **2009**, *49*, 2290–2302.
- (10) Nicolotti, O.; Giangreco, I.; Introcaso, A.; Leonetti, F.; Stefanachi, A.; Carotti, A. Strategies of Multi-Objective Optimization in Drug Discovery and Development. *Expert Opin. Drug Discovery* **2011**, *6*, 871–884.
- (11) Schneider, G.; Fechner, U. Computer-Based de Novo Design of Drug-like Molecules. *Nat Rev Drug Discov* **2005**, *4*, 649–663.
- (12) Kuntz, I. D. Structure-Based Strategies for Drug Design and Discovery. *Science* **1992**, *257*, 1078–1082.
- (13) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discovery Today* **2018**, *23*, 1241–1250.
- (14) Lo, Y.-C.; Rensi, S. E.; Tornig, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discovery Today* **2018**, *23*, 1538–1546.
- (15) Schneider, P.; Walters, W. P.; Plowright, A. T.; Sieroka, N.; Listgarten, J.; Goodnow, R. A.; Fisher, J.; Jansen, J. M.; Duca, J. S.; Rush, T. S.; Zentgraf, M.; Hill, J. E.; Krutoholow, E.; Kohler, M.; Blaney, J.; Funatsu, K.; Luebke, C.; Schneider, G. Rethinking Drug Design in the Artificial Intelligence Era. *Nat Rev Drug Discov* **2019**, 1–12.
- (16) Schneider, G.; Clark, D. E. Automated De Novo Drug Design: Are We Nearly There Yet? *Angew. Chem., Int. Ed.* **2019**, *58*, 10792–10803.
- (17) Lim, J.; Ryu, S.; Kim, J. W.; Kim, W. Y. Molecular Generative Model Based on Conditional Variational Autoencoder for de Novo Molecular Design. *J. Cheminf.* **2018**, *10*, 31.
- (18) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276.

- (19) Sattarov, B.; Baskin, I. I.; Horvath, D.; Marcou, G.; Bjerrum, E. J.; Varnek, A. De Novo Molecular Design by Combining Deep Autoencoder Recurrent Neural Networks with Generative Topographic Mapping. *J. Chem. Inf. Model.* **2019**, *59*, 1182–1196.
- (20) 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*, 3098–3104.
- (21) Putin, E.; Asadulaev, A.; Ivanenkov, Y.; Aladinskiy, V.; Sanchez-Lengeling, B.; Aspuru-Guzik, A.; Zhavoronkov, A. Reinforced Adversarial Neural Computer for de Novo Molecular Design. *J. Chem. Inf. Model.* **2018**, *58*, 1194–1204.
- (22) Prykhodko, O.; Johansson, S. V.; Kotsias, P.-C.; Arús-Pous, J.; Bjerrum, E. J.; Engkvist, O.; Chen, H. A de Novo Molecular Generation Method Using Latent Vector Based Generative Adversarial Network. *J. Cheminf.* **2019**, *11*, 74.
- (23) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular De-Novo Design through Deep Reinforcement Learning. *J. Cheminf.* **2017**, *9*, 48.
- (24) Gupta, A.; Müller, A. T.; Huisman, B. J. H.; Fuchs, J. A.; Schneider, P.; Schneider, G. Generative Recurrent Networks for De Novo Drug Design. *Mol. Inf.* **2018**, *37*, 1700111.
- (25) Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De Novo Design of Bioactive Small Molecules by Artificial Intelligence. *Mol Inform* **2018**, *37*.
- (26) Popova, M.; Isayev, O.; Tropsha, A. Deep Reinforcement Learning for de Novo Drug Design. *Sci. Adv.* **2018**, *4*, eaap7885.
- (27) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res* **2019**, *47*, D930–D940.
- (28) Irwin, J. J.; Shoichet, B. K. ZINC – A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
- (29) Li, Y.; Zhang, L.; Liu, Z. Multi-Objective de Novo Drug Design with Conditional Graph Generative Model. *J Cheminform* **2018**, *10*, 33.
- (30) Pogány, P.; Arad, N.; Genway, S.; Pickett, S. D. De Novo Molecule Design by Translating from Reduced Graphs to SMILES. *J. Chem. Inf. Model.* **2019**, *59*, 1136–1146.
- (31) Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; Petersen, S.; Beattie, C.; Sadik, A.; Antonoglou, I.; King, H.; Kumaran, D.; Wierstra, D.; Legg, S.; Hassabis, D. Human-Level Control through Deep Reinforcement Learning. *Nature* **2015**, *518*, 529–533.
- (32) Merk, D.; Grisoni, F.; Friedrich, L.; Gelzinyte, E.; Schneider, G. Computer-Assisted Discovery of Retinoid X Receptor Modulating Natural Products and Isofunctional Mimetics. *J. Med. Chem.* **2018**, *61*, 5442–5447.
- (33) Polykovskiy, D.; Zhebrak, A.; Vetrov, D.; Ivanenkov, Y.; Aladinskiy, V.; Mamoshina, P.; Bozdaganyan, M.; Aliper, A.; Zhavoronkov, A.; Kadurin, A. Entangled Conditional Adversarial Autoencoder for de Novo Drug Discovery. *Mol. Pharmaceutics* **2018**, *15*, 4398–4405.
- (34) 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.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik, A. Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors. *Nat Biotechnol* **2019**, *37*, 1038–1040.
- (35) Zhou, Z.; Kearnes, S.; Li, L.; Zare, R. N.; Riley, P. Optimization of Molecules via Deep Reinforcement Learning. *Sci. Rep.* **2019**, *9*, 10752.

- (36) Ståhl, N.; Falkman, G.; Karlsson, A.; Mathiason, G.; Boström, J. Deep Reinforcement Learning for Multiparameter Optimization in de Novo Drug Design. *J. Chem. Inf. Model.* **2019**, *59*, 3166–3176.
- (37) Deng, J.; Yang, Z.; Li, Y.; Samaras, D.; Wang, F. Towards Better Opioid Antagonists Using Deep Reinforcement Learning. *arXiv:2004.04768 [cs, q-bio]* **2020**.
- (38) Nicolaou, C. A.; Brown, N. Multi-Objective Optimization Methods in Drug Design. *Drug Discovery Today: Technol.* **2013**, *10*, e427–e435.
- (39) John Harris, C.; D. Hill, R.; W. Sheppard, D.; J. Slater, M.; F. W. Stouten, P. The Design and Application of Target-Focused Compound Libraries <https://www.ingentaconnect.com/content/ben/cchts/2011/00000014/00000006/art00007> (accessed Apr 8, 2020).
- (40) <https://Github.Com/Alberdom88/Moo-Denovo>.
- (41) Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Johansson, S.; Chen, H.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *arXiv:1811.12823 [cs, stat]* **2019**.
- (42) Brown, N.; Fiscato, M.; Segler, M. H. S.; Vaucher, A. C. GuacaMol: Benchmarking Models for de Novo Molecular Design. *J. Chem. Inf. Model.* **2019**, *59*, 1096–1108.
- (43) Sohrabi, C.; Alsafi, Z.; O'Neill, N.; Khan, M.; Kerwan, A.; Al-Jabir, A.; Iosifidis, C.; Agha, R. World Health Organization Declares Global Emergency: A Review of the 2019 Novel Coronavirus (COVID-19). *Int. J. Surg.* **2020**, *76*, 71–76.
- (44) Lai, C.-C.; Shih, T.-P.; Ko, W.-C.; Tang, H.-J.; Hsueh, P.-R. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and Coronavirus Disease-2019 (COVID-19): The Epidemic and the Challenges. *Int. J. Antimicrob. Agents* **2020**, *55*, 105924.
- (45) Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv:1412.3555 [cs]* **2014**.
- (46) Landrum, G. (2006). RDKit: Open-Source Cheminformatics.
- (47) Sutton R, Barton A (1998) Reinforcement Learning: An Introduction, 1st Edn. MIT Press, Cambridge.
- (48) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- (49) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
- (50) Benhenda, M. ChemGAN Challenge for Drug Discovery: Can AI Reproduce Natural Chemical Diversity? *arXiv:1708.08227 [cs, stat]* **2017**.
- (51) Kalgutkar, A. S.; Gardner, I.; Obach, R. S.; Shaffer, C. L.; Callegari, E.; Henne, K. R.; Mutlib, A. E.; Dalvie, D. K.; Lee, J. S.; Nakai, Y.; O'Donnell, J. P.; Boer, J.; Harriman, S. P. A Comprehensive Listing of Bioactivation Pathways of Organic Functional Groups <https://www.ingentaconnect.com/content/ben/cdm/2005/00000006/00000003/art00001> (accessed Apr 8, 2020).
- (52) Kalgutkar, A. S.; Soglia, J. R. Minimising the Potential for Metabolic Activation in Drug Discovery. *Expert Opin. Drug Metab. & Toxicol.* **2005**, *1*, 91–142.
- (53) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719–2740.
- (54) Schrödinger Release 2019-4: BioLuminate, Schrödinger, LLC, New York, NY, 2019.
- (55) Xu, X.; Zhu, X.; Dwek, R. A.; Stevens, J.; Wilson, I. A. Structural Characterization of the 1918 Influenza Virus H1N1 Neuraminidase. *J. Virol.* **2008**, *82*, 10493–10501.

- (56) Cheung, J.; Rudolph, M. J.; Burshteyn, F.; Cassidy, M. S.; Gary, E. N.; Love, J.; Franklin, M. C.; Height, J. J. Structures of Human Acetylcholinesterase in Complex with Pharmacologically Important Ligands. *J. Med. Chem.* **2012**, *55*, 10282–10286.
- (57) Jin, Z.; Du, X.; Xu, Y.; Deng, Y.; Liu, M.; Zhao, Y.; Zhang, B.; Li, X.; Zhang, L.; Peng, C.; Duan, Y.; Yu, J.; Wang, L.; Yang, K.; Liu, F.; Jiang, R.; Yang, X.; You, T.; Liu, X.; Yang, X.; Bai, F.; Liu, H.; Liu, X.; Guddat, L. W.; Xu, W.; Xiao, G.; Qin, C.; Shi, Z.; Jiang, H.; Rao, Z.; Yang, H. Structure of Mpro from COVID-19 Virus and Discovery of Its Inhibitors. *bioRxiv* **2020**, 2020.02.26.964882.
- (58) Moscona, A. Neuraminidase Inhibitors for Influenza. *N. Eng. J. Med.* **2005**, *353*, 1363–1373.
- (59) Stanciu, G. D.; Luca, A.; Rusu, R. N.; Bild, V.; Beschea Chiriac, S. I.; Solcan, C.; Bild, W.; Ababei, D. C. Alzheimer's Disease Pharmacotherapy in Relation to Cholinergic System Involvement. *Biomolecules* **2020**, *10*, 40.
- (60) Pisani, L.; Farina, R.; Soto-Otero, R.; Denora, N.; Mangiatordi, G. F.; Nicolotti, O.; Mendez-Alvarez, E.; Altomare, C. D.; Catto, M.; Carotti, A. Searching for Multi-Targeting Neurotherapeutics against Alzheimer's: Discovery of Potent AChE-MAO B Inhibitors through the Decoration of the 2H-Chromen-2-One Structural Motif. *Molecules* **2016**, *21*, 362.
- (61) Birks, J.; Harvey, R. J. Donepezil for Dementia Due to Alzheimer's Disease. *Cochrane Database Syst. Rev.* **2006**, No. 1.
- (62) Conejo-García, A.; Pisani, L.; del Carmen Núñez, M.; Catto, M.; Nicolotti, O.; Leonetti, F.; Campos, J. M.; Gallo, M. A.; Espinosa, A.; Carotti, A. Homodimeric Bis-Quaternary Heterocyclic Ammonium Salts as Potent Acetyl- and Butyrylcholinesterase Inhibitors: A Systematic Investigation of the Influence of Linker and Cationic Heads over Affinity and Selectivity. *J. Med. Chem.* **2011**, *54*, 2627–2645.
- (63) Patick, A. K.; Potts, K. E. Protease Inhibitors as Antiviral Agents. *Clin Microbiol Rev* **1998**, *11*, 614–627.
- (64) Chang, Y.-C.; Tung, Y.-A.; Lee, K.-H.; Chen, T.-F.; Hsiao, Y.-C.; Chang, H.-C.; Hsieh, T.-T.; Su, C.-H.; Wang, S.-S.; Yu, J.-Y.; Shih, S.; Lin, Y.-H.; Lin, Y.-H.; Tu, Y.-C. E.; Hsu, C.-H.; Juan, H.-F.; Tung, C.-W.; Chen, C.-Y. Potential Therapeutic Agents for COVID-19 Based on the Analysis of Protease and RNA Polymerase Docking. **2020**.
- (65) Grebner, C.; Matter, H.; Plowright, A. T.; Hessler, G. Automated De Novo Design in Medicinal Chemistry: Which Types of Chemistry Does a Generative Neural Network Learn? *J. Med. Chem.* **2020**.
- (66) Struble, T. J.; Alvarez, J. C.; Brown, S. P.; Chytil, M.; Cisar, J.; DesJarlais, R. L.; Engkvist, O.; Frank, S. A.; Greve, D. R.; Griffin, D. J.; Hou, X.; Johannes, J. W.; Kreatsoulas, C.; Lahue, B.; Mathea, M.; Mogk, G.; Nicolaou, C. A.; Palmer, A. D.; Price, D. J.; Robinson, R. I.; Salentin, S.; Xing, L.; Jaakkola, T.; Green, William. H.; Barzilay, R.; Coley, C. W.; Jensen, K. F. Current and Future Roles of Artificial Intelligence in Medicinal Chemistry Synthesis. *J. Med. Chem.* **2020**.
- (67) Alberga, D.; Trisciuzzi, D.; Montaruli, M.; Leonetti, F.; Mangiatordi, G. F.; Nicolotti, O. A New Approach for Drug Target and Bioactivity Prediction: The Multifingerprint Similarity Search Algorithm (MuSSeL). *J. Chem. Inf. Model.* **2019**, *59*, 586–596.
- (68) Montaruli, M.; Alberga, D.; Ciriaco, F.; Trisciuzzi, D.; Tondo, A. R.; Mangiatordi, G. F.; Nicolotti, O. Accelerating Drug Discovery by Early Protein Drug Target Prediction Based on a Multi-Fingerprint Similarity Search †. *Molecules* **2019**, *24*, 2233.
- (69) Cavalluzzi, M. M.; Mangiatordi, G. F.; Nicolotti, O.; Lentini, G. Ligand Efficiency Metrics in Drug Discovery: The Pros and Cons from a Practical Perspective. *Expert Opin. Drug Discovery* **2017**, *12*, 1087–1104.