

SOFTWARE

Open Access



# Systemic evolutionary chemical space exploration for drug discovery

Chong Lu<sup>1†</sup>, Shien Liu<sup>1†</sup>, Weihua Shi<sup>1</sup>, Jun Yu<sup>1</sup>, Zhou Zhou<sup>1</sup>, Xiaoxiao Zhang<sup>1</sup>, Xiaoli Lu<sup>1</sup>, Faji Cai<sup>1</sup>, Ning Xia<sup>2</sup> and Yikai Wang<sup>1\*</sup>

## Abstract

Chemical space exploration is a major task of the hit-finding process during the pursuit of novel chemical entities. Compared with other screening technologies, computational de novo design has become a popular approach to overcome the limitation of current chemical libraries. Here, we reported a *de novo* design platform named systemic evolutionary chemical space explorer (SECSE). The platform was conceptually inspired by fragment-based drug design, that miniaturized a “lego-building” process within the pocket of a certain target. The key to virtual hits generation was then turned into a computational search problem. To enhance search and optimization, human intelligence and deep learning were integrated. Application of SECSE against phosphoglycerate dehydrogenase (PHGDH), proved its potential in finding novel and diverse small molecules that are attractive starting points for further validation. This platform is open-sourced and the code is available at <http://github.com/KeenThera/SECSE>.

**Keywords:** Chemical space exploration, Fragment-based drug discovery, Deep learning, De novo drug design, PHGDH

## Introduction

Developing a new drug is an enduring process that is estimated to take 10–15 years with a cost of 1.5 billion US dollars or more. At the early drug discovery stage, the hit-finding program is crucial for a successful R&D campaign, especially for the challenging targets, which usually yield meager hit rates. There are many options for hit-finding, such as high-throughput screening (HTS), affinity selection-mass spectrometry (AS-MS), fragment-based drug design (FBDD), DNA-encoded library technology (DEL), and virtual screening (VS). However, all the above approaches suffer from the requirement of a predefined (real or virtual) compound library. To address the limitation, make-on-demand libraries [1–3] have gained some recent popularity in expanding the chemical

space. Nevertheless, even the most extensive collection of compounds claimed so far with the size of  $10^{26}$  [4, 5] is still a very tiny fraction of the estimated chemical space in the order of  $10^{63}$  [6]. Therefore, a systemic chemical space searching strategy is needed to provide optimal starting points against the target of interest.

De novo design is one such strategy that is conceptually able to overcome the limitation of existing compound libraries, which produces novel compounds based on the 3D crystal structure of a given target from scratch. A comprehensive summary [7–11] of the recent development in de novo design is out of the scope of this paper though, several seminal works that inspire us will be briefly reviewed in the following section.

LUDI [12] was an example of early attempts, where fragments from a predefined library were positioned into sub-pockets of the target. Then the fitted fragments were bridged together to form a new compound that better occupied the pocket. A similar approach called Lig-Builder [13] used module POCKET [14] to analyze and

\*Correspondence: [wang\\_yikai@keenthera.com](mailto:wang_yikai@keenthera.com)

<sup>†</sup>Chong Lu and Shien Liu are contributed equally to this work

<sup>1</sup> Keen Therapeutics Co., Ltd., Shanghai, China

Full list of author information is available at the end of the article



parameterize protein pockets and then applied module GROW or LINK to build up new molecules. A genetic algorithm was implemented in the growing and linking steps to avoid the combinatorial explosion of the molecular generating process. Subsequently, module SCORE predicted the binding affinity of the molecules. Synthesis accessibility (SA) analysis and more druglike filters were incorporated in the upgraded program LigBuilder v2 [15]. While in the latest version LigBuilder v3 [15], the authors began to consider the flexibility of pockets by including several samples from a particular target or different targets with similar binding pockets in the generation workflow. Wang et al. developed two versions of AutoT&T [16, 17] to automatically generate analogs for hit compounds under the spatial constraints of the targeted binding pocket. Their valuable attempts opened a way for computational de novo design methods in lead optimization. OpenGrowth [18] was an open-sourced de novo design program which also based on the fragment-based growing strategy. The 3-mers screening method required that generated molecules be made by defined fragments derived from the drug library, which warranted druglike properties. Like LigBuilder v3, different conformations of the target were considered to address the protein flexibility issue. Durrant *et al.* developed AutoGrow [19] to integrate fragment-based growing and docking with an evolutionary algorithm. The latest version is AutoGrow4 [20], which employed reaction-based rules for growing as mutation operators in the genetic algorithm and merging two molecules with maximum common substructure as crossover operators. Substructure or property filters (like the rule of 5 [21], PAINS [22]) were used to control the quality of generated molecules. At the same time, open-source docking programs were invoked to evaluate the binding affinity. Although AutoGrow4 performed well in some cases, reaction-based molecular generation is intrinsically limited for constructing novel chemical entities. Polishchuk published an open-sourced tool called CReM [23] to produce highly diverse structures by fragment manipulation (mutate, grow and link). Nigam *et al.* proposed STONED for efficient search of chemical space using a SELFIES modification method [24]. Recently, Steinmann and Jensen reported a non-fragment-based approach [25], which used a set of reaction-like rules to build up chemical structures, yielding molecules with acceptable glide docking scores and synthetic feasibility by genetic algorithm.

In addition to rule-based generators, deep generative models have also been extensively explored. MolAICal [26] used generative deep learning models for 2D structure construction and classical methods for 3D evaluation and simulation. Recently, Ma *et al.* [27] developed SBMolGen, which contained an RNN based SMILES

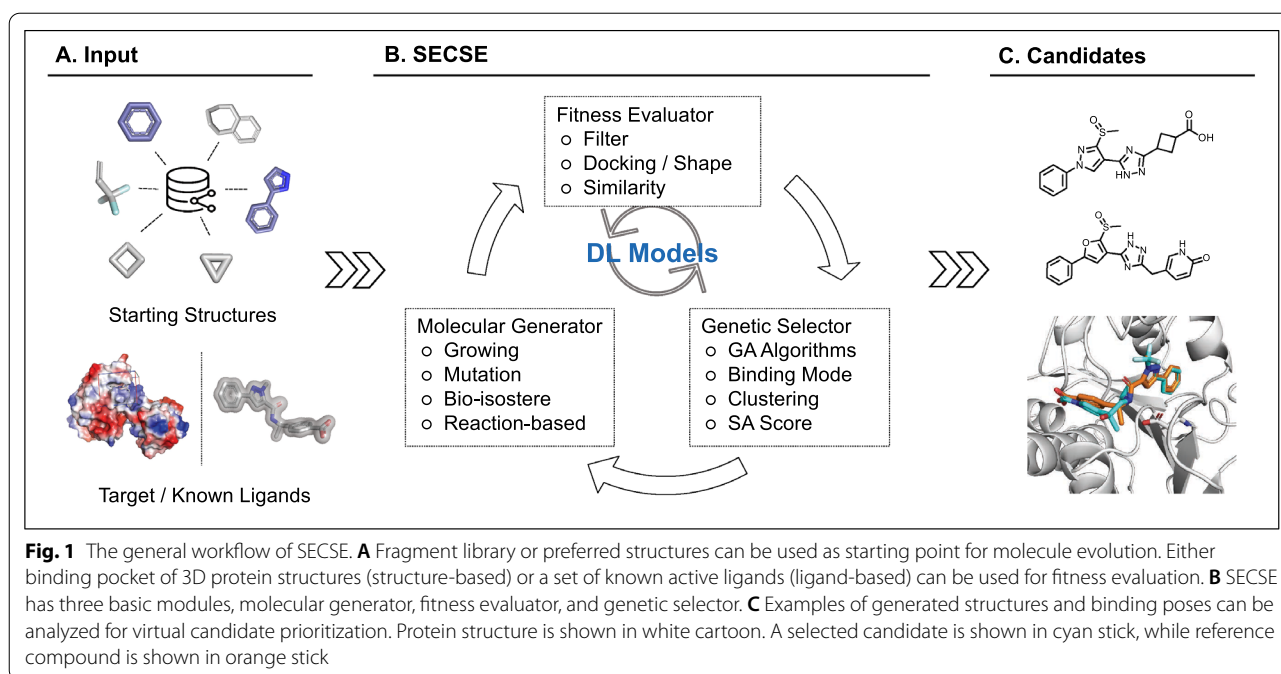
generator called ChemTS [28], a Monte Carlo tree search, and docking simulations. Lai *et al.*, the authors of LigBuilder, developed DeepLigBuilder [29] to generate 3D molecules directly from deep generative models. Several other new approaches utilizing deep learning methods to generate 3D molecules have been reported [30–33]. Compared with 1D/2D generative models or rule-based methods [34, 35], the competitive advantage of using these 3D models is direct 3D conformation generation with high speed. However, it is not easy to directly converge when training deep learning models end to end. Researchers have to introduce some special treatments for the data type and model architecture to terminate the training process, which is usually difficult to interpret.

Inspired by previous attempts in the field, we present our work setting up a platform to explore the chemical space against a given target systemically. Analogous to other programs, the SECSE platform consists of three modules, namely, a molecular generator, a fitness evaluator, and a genetic selector. The output is chemical structures that specifically fit the evaluation model for a defined pocket or other criteria. Moreover, PHGDH is chosen to demonstrate the potential of the SECSE platform. Virtually generated molecules are shown, and the corresponding structure-activity relationship is analyzed for this target. Their high docking scores and reasonable binding poses, in addition to structural novelty and patentability, warrant further exploration.

## Implementation

SECSE is a de novo design software that mainly combines rule-based molecular generation and structure-based drug design methods using genetic algorithms, as well as a deep learning module. The SECSE platform is implemented by a molecular generator, a fitness evaluator, and a genetic selector. In the molecular generator module, we have created more than 3000 rules for molecular transformations based on knowledge and expertise from the literature domain and our internal medicinal chemists. These rules are comprehensively curated and strategically categorized for optimal output. In the fitness evaluator module, molecular docking is utilized for compound assessment, which can also be replaced by shape-/pharmacophore-based evaluation methods. In the genetic selector module, a genetic algorithm is used given the similarity between the triage strategy of fragment growing and the genetic rule “fitness to survive”.

The workflow of SECSE is described in Fig. 1. In the first place, fragments/groups are docked/positioned into the pocket, from which the ones with high docking scores or ligand efficiency are picked as elites. It is noteworthy to point out that fragments with less than 13 heavy atoms are exhaustively enumerated as initial input, yet



any given structures or functional groups can be used as starting points. Then all the elites are evolved to generate novel chemical structures by applying the rules. The child molecules are clustered and sampled to represent the pool. The sampled molecules are docked into the pocket again. Highly scored molecules adopting hereditary or reasonable 3D orientation are chosen as new elites. This process concludes one cycle. After multiple cycles of iteration, a considerable number of compounds are generated and accumulated. To comprehensively evaluate all compounds, we introduced a graph-based machine learning module to speed up elite selection in each generation. Finally, hit compounds are visually inspected and selected before wet lab synthesis.

### Fragments collection

As starting points of the entire workflow, the quality of the fragment collection would determine the final output to some extent. Fragments derived from compounds in co-crystal structures or based on hypotheses can be used as proprietary input. However, it would inevitably be limited by the size of existing fragment collections or human bias. To ensure the diversity of the starting fragment library, we proposed an algorithm analogous to GDB13 [36] that can potentially enumerate molecules containing up to 12 heavy atoms with MW ranging from 50 Da to 210 Da. We provided the source code in order to explicitly present the details of the fragment space exploration.

As described in Additional file 1, sequential carbon strings, such as “CCCCCC”, are the starting point of

fragment generation with fixed heavy atom numbers. The SMILES string is then modified to construct aliphatic rings, which are subsequently submitted for structure transformations (aromatic ring formation, sidechain rearrangement, and atom/bond replacement, etc.). A series of filters (the same filter rules in SECSE) are applied to remove fragments with undesired architecture/topology or functional groups. Final structures of 121,860,917 fragments are stored in an SQLite database (See Data availability section).

### Input preparation

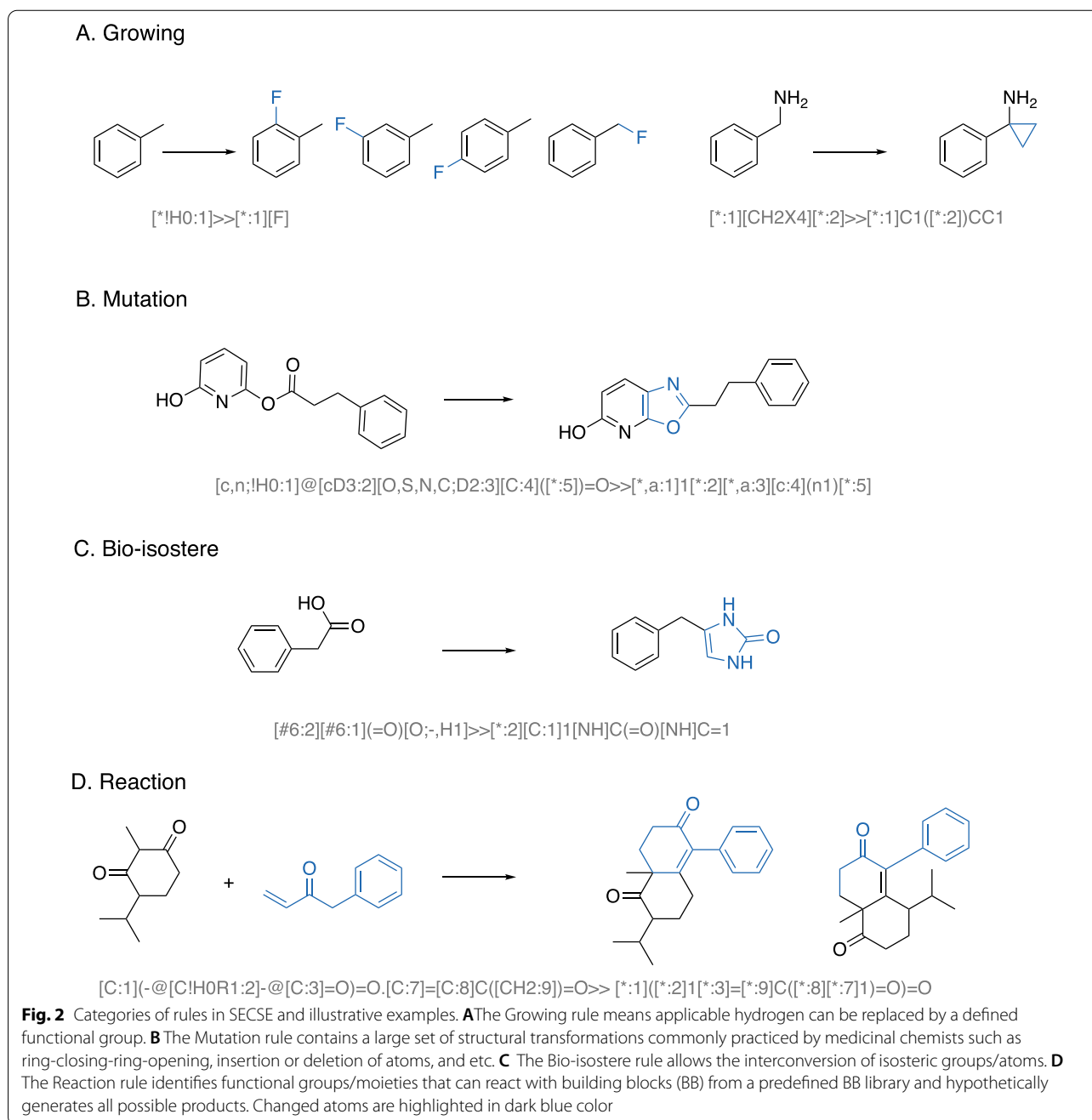
In the workflow, chemical structures and protein structures are the primary inputs. Depending on different purposes, the chemical input can be an atom, a fragment structure, or a fragment library in the format of a tab-separated file containing structure SMILES and ID. If needed, the provided SMILES can be converted into a 3D structure using ETKDG v2 built in RDKit [37]. Tautomer and spiro centers are also enumerated on demand. For AutoDock Vina docking, the ligands are converted from SDF format to PDBQT format using Open Babel v3.1.1 [38]. Fragment libraries are recommended for hypothesis-driven hit discovery, especially when limited binders against the target of interest are reported. Protein 3D structures are prepared from crystal structures from the Protein Data Bank (PDB) [39]. Homology models or predicted structures from AlphaFold2 [40] /RoseTTAFold [41] are also acceptable although with compromised accuracy and predicting power. In our demo case, protein

structures are prepared for docking with ADFR v1.2 [42, 43].

### Molecular generator

The molecular generator we have developed provides a rule-based generation approach. There are four types of transformation rules (growing, mutation, bioisostere, and reaction) in our database. Some representative cases of each class are shown in Fig. 2.

- (1) In the grow rules, any of the replaceable hydrogen atoms on the seed compound can be replaced with a new substructure, such as an atom, a functional group, a ring, or a ring with a linking spacer.
- (2) The mutation rules include the following three categories: atom replacing, insertion, and deletion; ring-closing, ring-open, ring modification (expansion, reduction, contraction); as well as aromatic-aliphatic exchange.



- (3) The bioisostere rules refer to classical or non-classical bioisosteric replacements, which are commonly used by medicinal chemists.
- (4) The reaction-based rules contain common organic reactions confined to one or two steps. A library of commercial building blocks is used as starting materials. Applying the chemical reaction rule is beneficial to efficiently increase the scaffold diversity of the resulting molecules, although they can be generated from multiple rounds of rules from the previous three categories.

All the rules are represented in the reaction-like format using the SMARTS. Hybridization extension defined in RDKit was also used in the rules. A few examples from each category are provided in the SQLite database.

#### Property and structure filter

To ensure that the platform generates molecules with decent chemical beauty [44], we construct several filters that define molecular properties, ring system count, and substructures.

- (1) The default parameters of the molecular property filters are shown as follows: molecular weight (MW)  $\in [81, 450]$ ; LogP  $\in [0, 5]$ ; the number of hydrogen bond donors (HBD)  $\leq 5$ ; the number of hydrogen bond acceptor (HBA)  $\leq 10$ ; the number of rotatable bonds (RB)  $\leq 4$ ; and topological polar surface area (TPSA)  $\leq 200$ . All properties here are calculated by RDKit v2021.03.5. The bond between two connecting parts that both have  $\pi/p$  orbitals is not rotatable because it tends to form a conjugated system with up to two stable rotamers. To count the real rotatable bonds, the definition of RB was replaced as '[C^3!D1;!\$(C(F)(F)F)]-!@[Br!F!Cl!!H3 &!\$(\*\*\*)!D1;!\$([Br!F!Cl!I](F)(F)F)]' here.
- (2) The default constraints for ring systems are: total ring system count  $\leq 4$ ; the number of rings in a polycyclic ring system  $\leq 3$ ; ring size  $\leq 7$ ; fused ring count  $\leq 3$ ; bridged ring count  $\leq 1$ ; and spiro ring count  $\leq 1$ .
- (3) Undesirable structures are also discarded by identity filters (sulfur, phosphorus, or structure alert), and count filters (e.g., max number of carboxylic acid or alkyne in one compound). PAINS filters are also included.

One thing worthy of note is that the filters are arbitrarily set depending on project requirements, which can be adjusted if the output is not ideal.

#### Fitness evaluator

Structure-based virtual screening engines such as molecular docking or pharmacophore-based screening methods are optional for fitness evaluation. Docking is the first choice for fitness evaluation. The default docking software in our platform is AutoDock Vina v1.2.0 [45, 46]. We also provide a Glide interface for users with commercial licenses. Additionally, we offer shape-based screening and similarity scoring functions to evaluate fitness for ligand-based drug design (i.e., the initial input is not a protein structure but one or more ligands with known activity).

Several scoring functions are optimized to achieve the evaluator function for different scenarios. If the docking mode is selected, both docking score and ligand efficiency (LE) [47, 48] are considered as ranking criteria, where

$$LE = \frac{\text{Docking score}}{1 + \ln(\text{Number of heavy atoms})}$$

The docking score tends to favor larger molecules in our previous tests. In contrast, LE can correct the issue by preventing premature enrichment of large molecules before reaching the upper molecular weight cutoff. Root Mean Square Deviation (RMSD) of aligned atoms between docking poses of the previous and current generation is calculated to determine whether the binding mode has changed in the two consecutive generations. If the similarity search mode is selected, the optional scoring functions will be a Tanimoto index of different molecular fingerprints from the generated molecules and reference compounds. In addition, the retrosynthesis module from Chemical.AI [49] is invoked to assess the synthetic availability.

#### Seed selector

After scoring, molecules with RMSD less than 2 Å or with significantly decreased docking scores are selected as seeds for the next generation. The purpose of the selector is to make sure compounds with consistent binding modes are maintained while compounds with much better binding modes won't be carved out. Then we apply a genetic algorithm [50–52] to select seeds from all eligible molecules. In our platform, the default GA operator is the tournament selection which is the most widely used selection strategy. Consequently, it can quickly converge to the optimal solution within noisy environments and introduce some randomness to avoid the limitations caused by local optimization.

Because of the limited computing resources, we sample data from the molecules generated by all the rules. Likewise, we use a partition clustering algorithm (see

Additional file 1) before sampling to ensure the diversity of the selected molecules. We calculate the molecular fingerprint and Tanimoto index to evaluate the distance/dissimilarity between generated molecules, based on which the sampling is executed.

### Deep learning-based fitness prediction

Although SECSE can generate a significant number of molecules, most of them are not evaluated due to limitations in computing and storage resources. Therefore, we apply deep learning (DL) modeling to reduce computational costs and make it possible to evaluate the fitness of all molecules. We use the data generated after each generation to train the model and then predicts the fitness of unsampled molecules. Docking score or ligand efficiency can be considered as target for prediction if the docking mode was selected. Fitness prediction models are constructed using package Chemprop (version 1.3.1). Chemprop builds a directed message passing neural network and learns to predict molecular properties directly from the graph representation of molecules. [53] Two strategies are provided here for the integration of DL technology. One is the combined mode, where top-ranked molecules prioritized by predicted scores were evaluated by the fitness module. These molecules were applied for seed selection together with docked molecules from sampling procedure. Moreover, in the combined mode deep learning models will be updated with each round of molecular generation. The other one is called clean mode. The DL model is trained based on the docking results after a SECSE campaign is finished. Data from each generation can be trained independently or together. The model can then be applied on undocked molecules for fitness prediction. Molecules with good performance from DL models can be subjected for further inspection. Additionally, these two modes can be used alone or in combination.

The platform uses some open-source packages: RDKit v2021.03.5, Open Babel v3.1.1, AutoDock Vina v1.2.0, Chemprop v1.3.1, and GNU parallel v20190922 [54].

## Results

### Properties of generated molecules

We constructed a random library using SECSE without any other evaluation constraint to estimate the molecular properties of generated compounds. Benzene was assigned as the only input fragment. During each iteration, one hundred molecules that passed the filter were randomly selected as seeds for the next iteration. The final random library after ten rounds of iteration contains 2,042,863 molecules, which are included in the Figshare. More information can be found in the Data availability section.

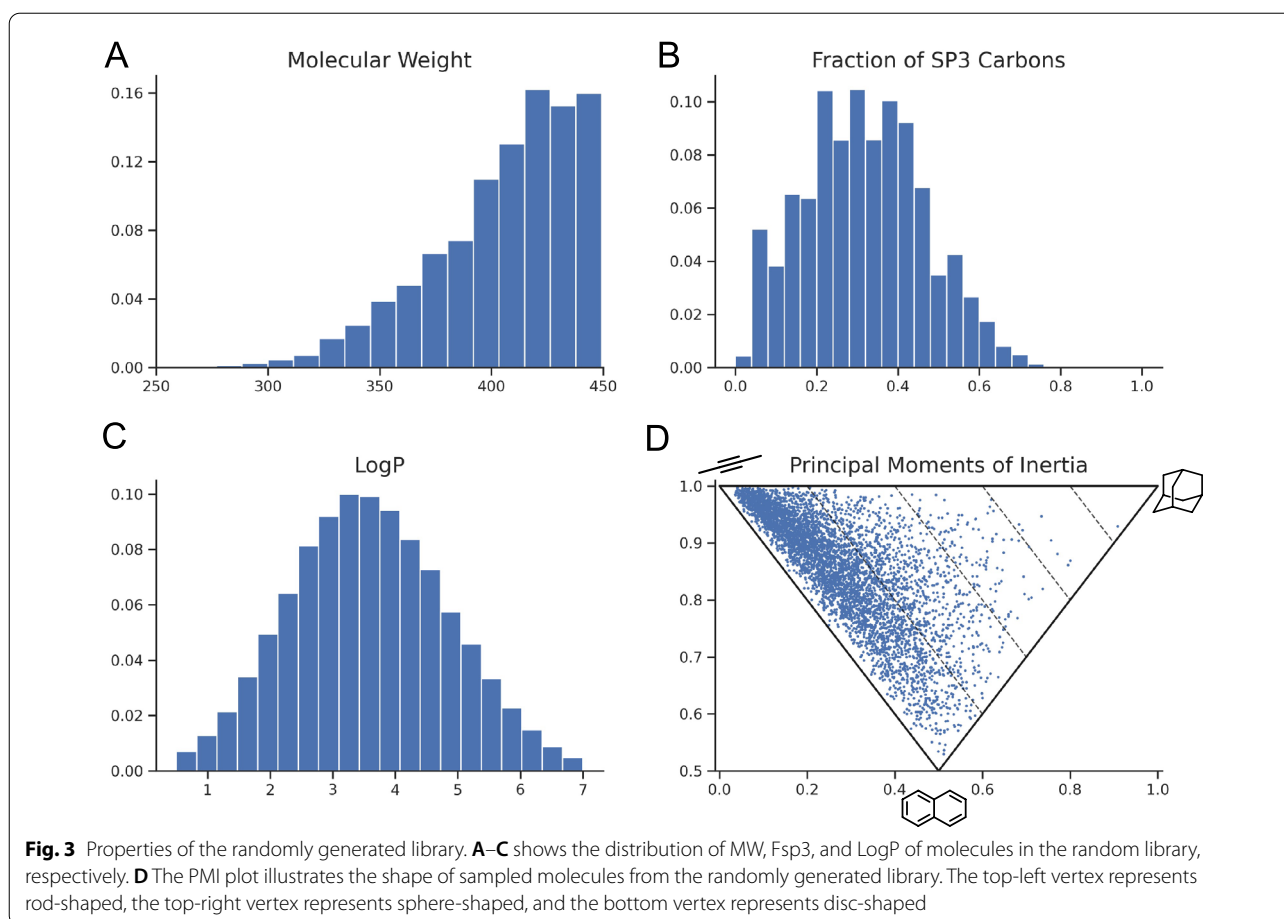
We calculated the physicochemical properties of the random library, such as molecular weight (MW), LogP, and the fraction of *sp*<sup>3</sup> hybridized carbons (Fsp3) [55], as well-illustrated in Fig. 3. Despite the upper limitation of MW in the filters, we could find that the peak falls around 450 Da. The distribution of LogP showed that the majority of molecules have a value between 0 and 5. Molecules with a high Fsp3 tended to be more three-dimensional in shape. The Fsp3 of the random library was well-distributed from 0 to 0.8. In addition, five thousand molecules were randomly sampled to plot the principal moments of inertia (PMI) [56], a more direct descriptor for assessing the distribution of molecular geometry (rod-shaped, disc-shaped, and sphere-shaped). We used the MMFF94 force field in RDKit to optimize the conformers of sampled molecules. As presented in Fig. 3D, the molecules scattered more towards the top-right vertex (sphere-shaped) in comparison with traditional HTS compounds that predominately dropped between the top-left vertex (rod-shaped) and bottom vertex (disc-shaped) (data not shown). HTS Compounds are usually built by two or three building blocks linearly. However, the transformation rules here can be used to extend fragments in any direction without human bias. That is why structures in our data set are more spherical than structures from other libraries designed by medicinal chemists. The results presented here indicate that SECSE can generate structures with suitable druglike properties and diverse geometry.

### Case: phosphoglycerate dehydrogenase (PHGDH)

PHGDH is a crucial enzyme that catalyzes the first committed step of the de novo serine synthesis pathway. It converts 3-phosphoglycerate to 3-phosphopyruvate in a reduced nicotinamide adenine dinucleotide (NADH)/nicotinamide adenine dinucleotide (NAD<sup>+</sup>)-dependent oxidation reaction. Many reports [57, 58] have indicated that overexpression of PHGDH is associated with various diseases, especially cancer. Inhibition of PHGDH may be a promising strategy for cancer therapy [59–63].

### Regeneration of PDB ligands

To validate the retrospective accuracy of SECSE, we performed a test to reproduce experimental binding modes for co-crystal ligands against PHGDH. By the time of our analysis, there are fourteen co-crystal structures of PHGDH containing drug-like ligands in PDB. Seven of the ligands are fragments with molecular weights of less than 150 Da. To narrow down the chemical space under limited computational resources, we selected a small set of rules (see Additional file 4) to generate molecules for this test. The set of rules can be found in the Additional file 4 in SMARTS format. The protein structure of



6RJ3 was selected as input. All water were removed from the original crystal structure. Benzene was employed as the initial fragment. AutoDock Vina was selected for molecular docking. After seven rounds of iterations, ten ligands were reproduced. The growing paths are shown in Fig. 4. Four of them evolved from the same branch and shared a common ancestor GEN\_2\_M\_002106. Fig. 5 shows the poses and RMSD of rediscovered ligands compared to crystal conformation (6RJ3, 6RJ5, and 6RIH), respectively.

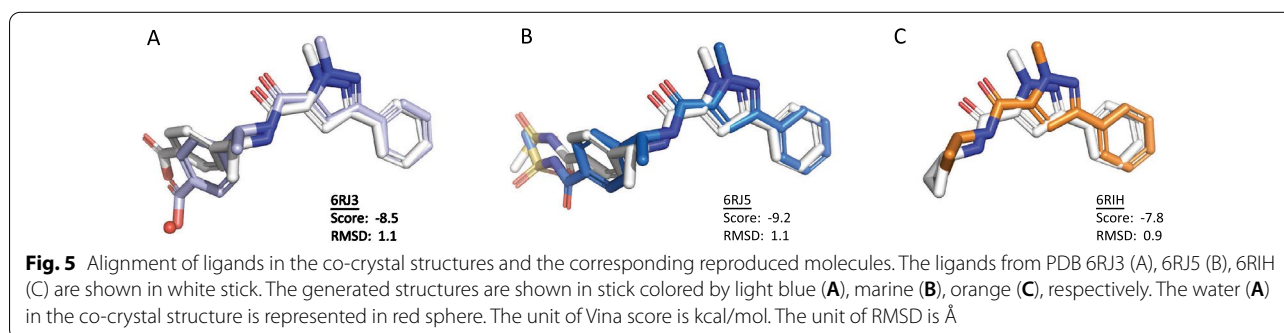
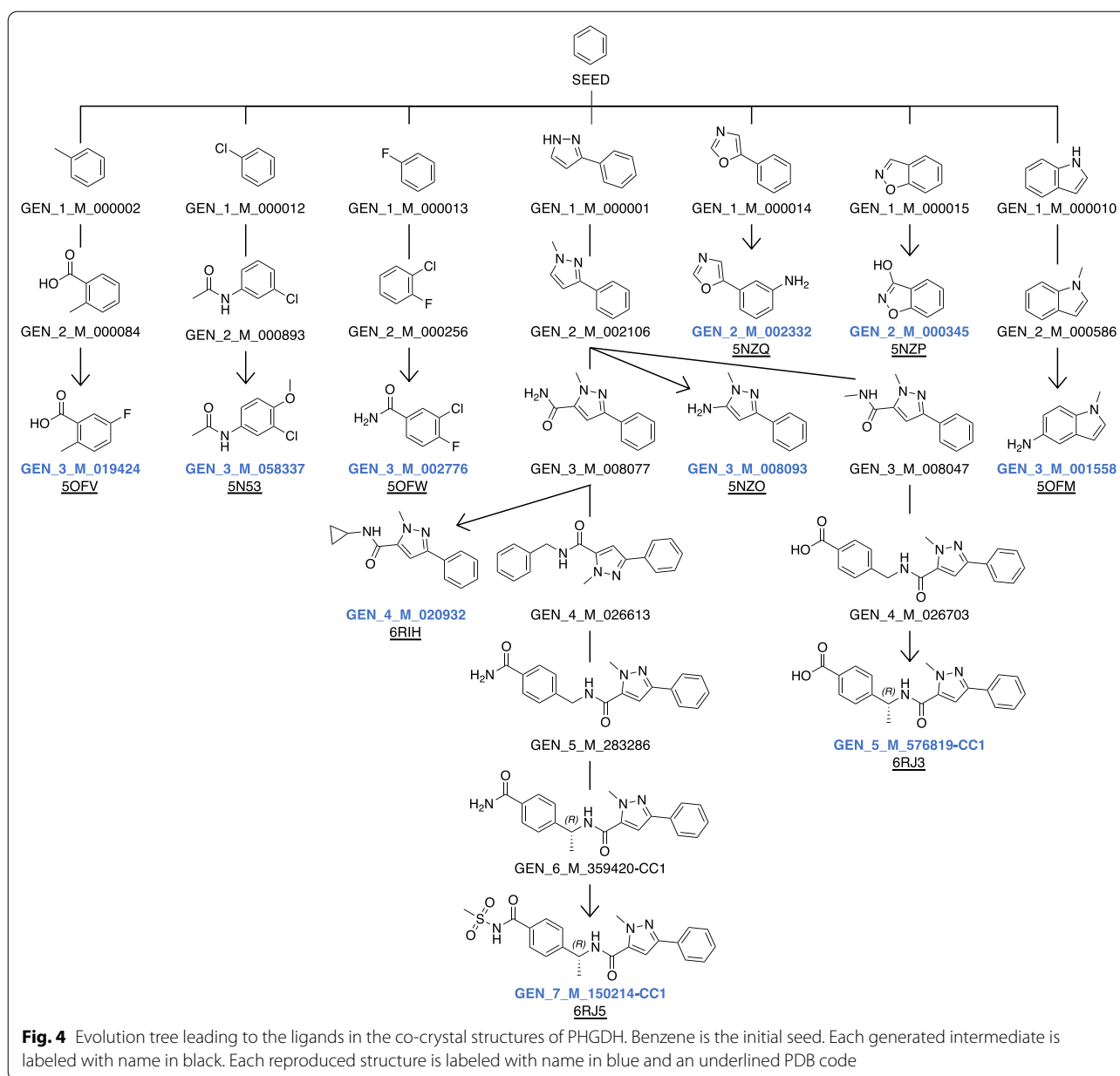
The three RMSD values are all less than 2.0 Å. Compared with the ligand in 6RJ3, the left carboxyl and phenyl of GEN\_5\_M\_576819-CC1 are shifted, resulting in a RMSD of 1.1 Å. The reason for less constrained conformation in this part is due to removal of the crystal water at this position (shown in Fig. 5A). In addition, the other four crystallographic ligands (ligands from 6RJ2, 6PLF, 6PLG, 6RJ6) were not regenerated. The ligand from 6RJ2 was not reproduced because of the low ligand efficiency ranking of its parent molecule. In addition, ligands from 6PLF, 6PLG and 6RJ6 were not regenerated since their parent molecules did not continue to evolve as a result of insufficient

sampling. The raw data can be found in Figshare (see Data Availability section).

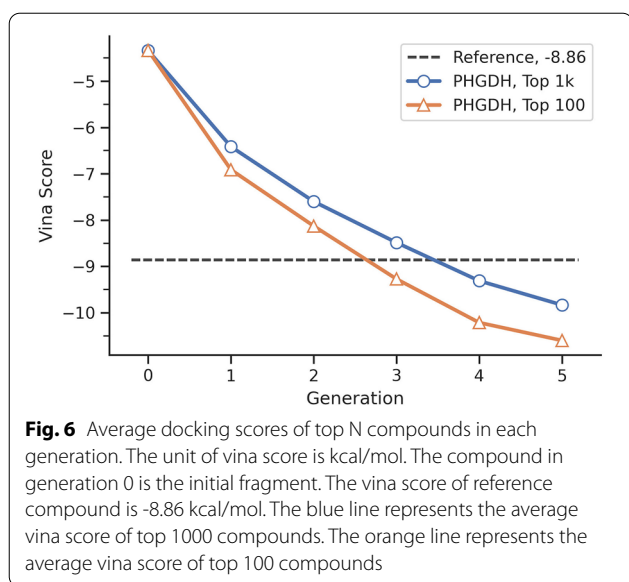
From the test run, SECSE has proven to be powerful in the rediscovery of known ligands against PHGDH. Next, we sought to demonstrate its value in the generation of novel compounds, the transferable process of which can be applied in hit identification against novel targets with limited ligand-bound information available.

#### Generation of novel compounds

The upper limitation of molecular weight was set to 500 Da, and the starting fragment was benzene. The crystal structure of PHGDH (PDB code: 6RJ3) was prepared using previous descriptions in Implementation. Together, 502,226 poses were collected after five generations. The docking scores were gradually decreased (Fig. 6). Not surprisingly, compared with the average docking score of the top 1000 of each generation, that of the top 100 molecules was improved more rapidly. After three generations, the average docking scores of either the top 100 or top 1000 compounds were better than that of the reference compound (− 8.8 kcal/mol). Additionally, it was observed that the scores started to converge at later







generations indicating the pocket occupancy was quickly approaching its optimum. It is plausible that the converging rate for different targets might be different. More rounds of iteration can be performed. Yet in this case, we stopped here for further analysis.

Finally, 14,413 poses with AutoDock Vina score less than  $-9$  kcal/mol were obtained. Then, the similarity distance cutoff was set to 0.15 to cluster these molecules according to the RDKit fingerprint. The one with the lowest docking score of each cluster was chosen for further binding pose inspection. Afterwards, we retrieved analogs of molecules of interest from the original docking pose pool. To keep the long-range electrostatic (LRE) interactions in the phosphate channel [62], the generated molecules with electron-rich functional groups are preferably selected using substructure filters. The final list of compounds is compiled manually. Table 1 (see Additional file 2) below includes some selected examples.

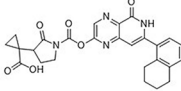
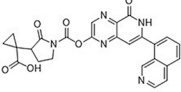
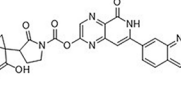
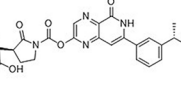
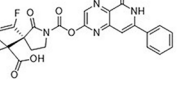
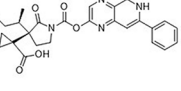
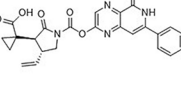
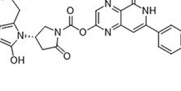
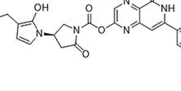
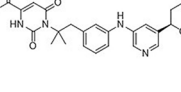
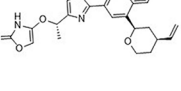
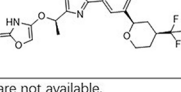
All the molecules shared similar binding modes with the reference compound from 6RJ3. Compounds 1–7 share a common topology. Similar to the phenylpyrazole-5-carboxamide part of the reference compound, the phenylpyridopyrazin-5-one occupies the same position of the adenine pocket. The nitrogen atom in the pyrazine ring forms a polar interaction with the side chain of D174 to stabilize the lipophilic aromatic fragment. The cyclopropanecarboxylic acid motif, which mimics the benzoic acid in the reference compound, has long-range Van der Waals interactions with the basic residues in the phosphate channel. The hydroxy-pyrrol-2-yl acetic acid moiety of compounds 8 and 9 act as the same role for Van der Waals interactions, as well as oxazolone of compounds 11 and 12.

The step-by-step elaboration of compound 2 from the benzene ring in the NADH/NAD<sup>+</sup> binding pocket was presented in Fig. 7A to F. The final pose of compound 2 was aligned to the reference compound. The common structure between current and previous generations showed nearly identical orientations. The AutoDock Vina docking scores decreased from  $-4.3$  kcal/mol to  $-10.8$  kcal/mol, while the MW increased from 78 Da to 485 Da. Despite the decent docking scores, conformation of compounds in D, E, F generated by AutoDock Vina may not be energetically favorable. The oxygen atom of the carbonyl group is near the nitrogen atom of pyrazine, whereas they should stay away in the lowest energy conformations. More accurate docking programs are needed for better outcomes.

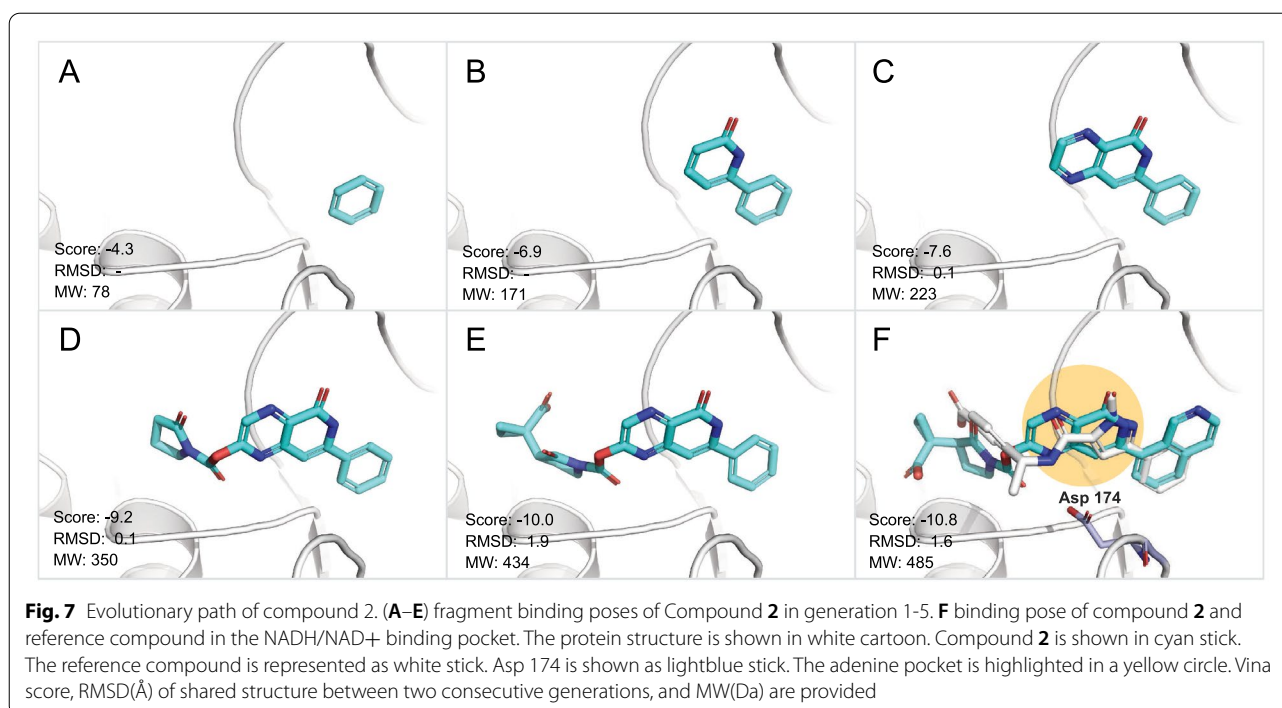
Subsequently, the Synthetic Accessibility Module from Chemical.AI was used to estimate accessibility of these proposed molecules. The Synthetic Accessibility Module provides a primary estimation for many organic compounds under restricted computing resources. The predicted routes may not be the best choice, but it gives a quick estimate that can be used to assess whether the compound is easy to make or not. Generally speaking, majority of compounds can be made within 15 synthetic steps with no more than 7 linear steps. Unfortunately, no synthetic routes of compound 5 are suggested under the default setting of Synthetic Accessibility Module. In such cases, or when dealing with a short list of candidates that are of high interest, more accurate predictions can be done using the Synthesis Plan Module, which performs extensive searches for all possible synthetic routes.

To address the sampling limitation of generated molecules before fitness evaluation, a new selection protocol combined deep learning method was developed. As described previously in the Implementation section, the clean mode was used to build the DL model. Fig. 8 demonstrates the details of the model performance of each generation (A-E) and combined set (F), which includes data from all previous generations. The value of  $R^2$  from Generation 1 to Generation 5 was gradually improved from 0.66 to 0.85. Furthermore, the  $R^2$  of the DL model from the combined set was 0.85, slightly better than other models trained only by a single generation. It is reasonable to believe the model performance is sufficient for prediction [64–66]. The model F was then used for the prediction of the 66,687,173 molecules and 2,094 molecules with predicted scores less than  $-10.5$  kcal/mol were subjected for redocking. The structures, MW, LogP, and synthetic accessibility analysis of representative molecules were listed in Table 2 (see Additional file 3). It was pleased to see compounds that share similar scaffold with compound 2. Compounds that have completely different scaffolds were also identified to provide diverse and

**Table 1** List of selected candidates

ID	Name	Structure	MW	LogP	Vina Score	Total Steps	Liner Steps	SAScore
1	GEN_5_M_00769405-CC0		488	3.08	-11.51	11	7	3.32
2	GEN_5_M_00769403-CC0		485	2.75	-10.76	9	7	3.00
3	GEN_5_M_00769397-CC0		485	2.75	-10.25	10	7	3.16
4	GEN_5_M_00768747-CC3		490	3.71	-10.11	11	5	3.32
5	GEN_5_M_00769103-CC1		492	3.44	-10.21	N/A*	N/A	5.00
6	GEN_5_M_00768755-CC3		490	3.61	-10.08	7	5	2.65
7	GEN_5_M_00768685-CC3		460	2.61	-10.06	11	7	3.32
8	GEN_5_M_01041792-CC1		489	2.09	-10.35	8	5	2.83
9	GEN_5_M_01041787-CC0		489	2.09	-9.91	8	5	2.83
10	GEN_4_M_05182502-CC1		464	3.59	-9.54	14	10	3.74
11	GEN_5_M_67166829-CC1		433	4.09	-10.00	12	7	3.46
12	GEN_5_M_00132975-CC7		479	5.74	-9.20	13	9	3.61

\* N/A means the predicted synthetic routes are not available.



valuable hypotheses for further validation. The superior performance of the deep learning model in the SECSE platform was speculated to result from the intrinsic logic. The 3D structural information of the parent compound was inherited in its child compounds, while the child generation would feedback rich structure-activity relationship (SAR) for model training. It also explained the model performance was improved in later generations while the combined set yielded the best result.

In this case, we used an 80-core computer, which took 40.5 hours in total. The Fig. 9 shows the calculation time of each generation. As the molecules grow larger and their complexity increases, the running time of each generation would gradually increase. Deep learning modeling significantly reduces search time and makes it possible to obtain estimated fitness scores for all generated molecules.

## Discussion

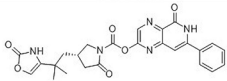
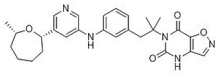
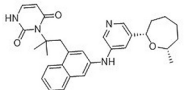
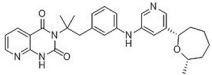
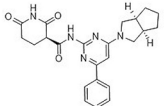
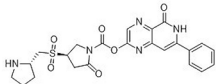
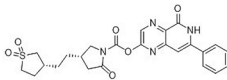
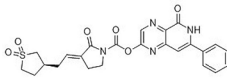
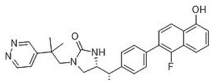
To our knowledge, the previously reported de novo design tools are not widely used in hit-finding programs. A potential reason is the inadequate exploration of chemical space. Moreover, limitation in chemical space coverage of current chemical libraries is a common problem the field is facing. Intensively expanding the chemical space via brute-force exploration, which leads to ultra-large chemical libraries such as GDB [36, 67, 68] is explored. A more widespread attempt is the

make-on-demand library [4], which comprises of structures from the enumeration of commercial building blocks based on reliable reaction schemes. The commercial providers also claim to have a relatively high synthesis success rate (at least 30%). The success of ultra-large compounds virtual screening contributes to the vigor of the make-on-demand library [2, 3]. Furthermore, people train machine learning models to accelerate the speed of virtual screening to balance the tradeoff between accuracy and speed [64]. However, it is still a very tough task to do virtual screening of ultra-large libraries directly on the present hardware.

All these factors are considered and balanced in our own platform. It is probably unrealistic to enumerate all druglike molecules, but an exhaustive enumeration of fragments with less than 13 heavy atoms is doable. Compared with previous de novo design softwares, reaction rules are a systemically curated to enrich the diversity of structures. Many heuristic filters from medicinal chemists' knowledge were incorporated in SECSE to maintain the beauty and practicality of molecules. To avoid the combinatorial explosion, protein pockets are constraints to direct the evolution. In addition, the unexpected accuracy of the deep learning model allows us to evaluate a large amount of compounds with minimal false positives or false negatives.

Recently, deep generative neural networks have become a promising approach for molecular generation. Many seminal reviews [10, 69] have summarized

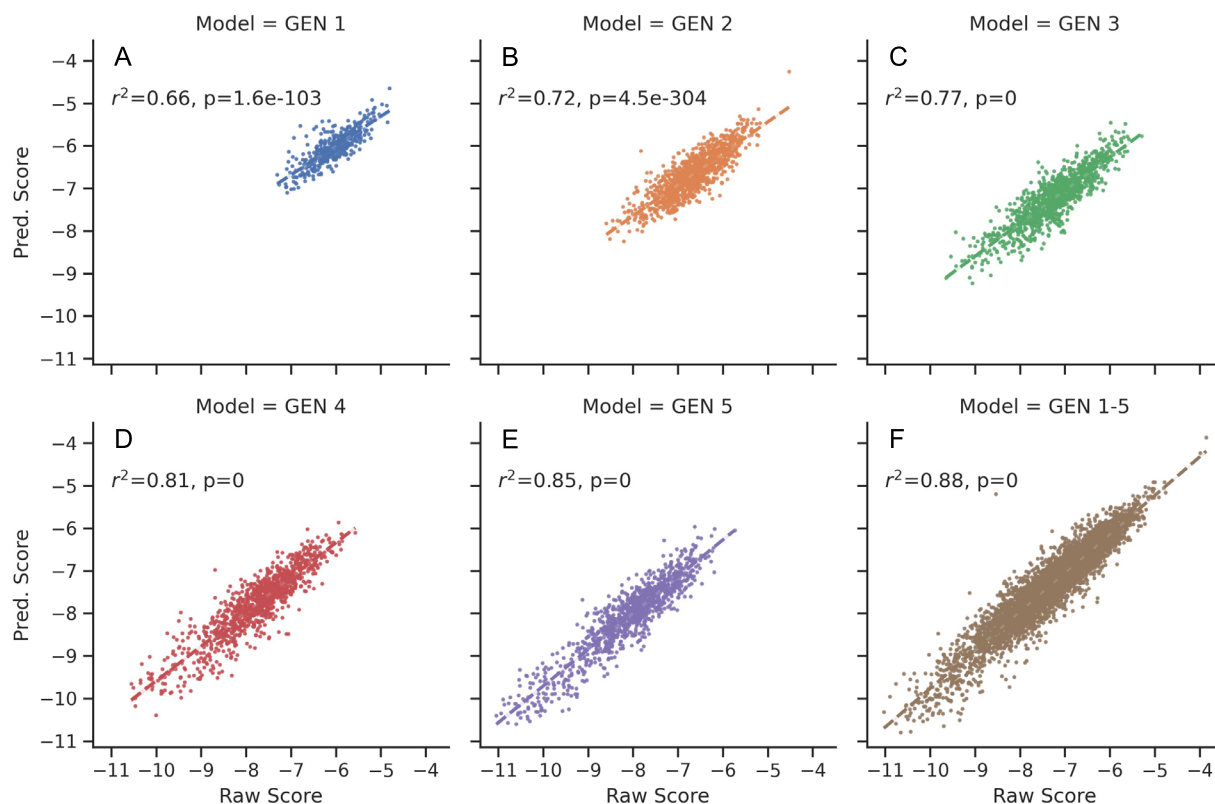
**Table 2** List of selected candidates based on the evaluation of DL models

ID	Name	Structure	MW	LogP	Predicted Score	Vina Score	Total Steps	Linear Steps	SAScore
13	GEN_4_M_00490117-CC1		489	2.98	-10.60	-11.34	14	9	3.74
14	GEN_5_M_02805023-CC3		489	4.81	-10.53	-10.12	12	10	3.46
15	GEN_5_M_02804877-CC0		498	5.83	-10.53	-10.12	16	8	4.00
16	GEN_5_M_02804883-CC2		499	5.22	-10.73	-10.10	15	8	3.87
17	GEN_3_M_26077238-CC1		419	2.37	-10.65	-10.06	5	4	2.24
18	GEN_4_M_00456833-CC1		498	0.23	-10.58	-10.06	10	7	3.16
19	GEN_4_M_00471055-CC0		496	2.55	-10.58	-10.21	12	7	3.46
20	GEN_4_M_00476893-CC7		494	2.47	-10.62	-10.21	17	7	4.12
21	GEN_5_M_05971075-CC1		484	5.62	-10.58	-10.18	15	13	3.87

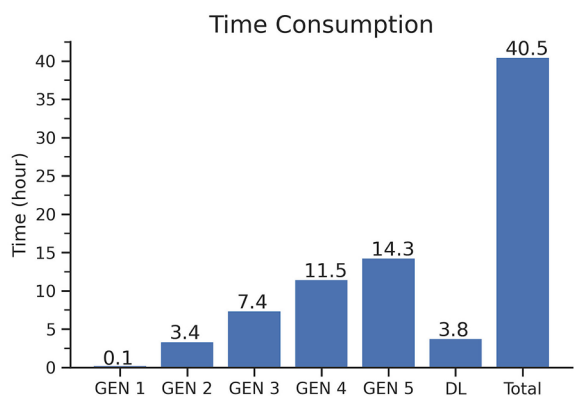
the development of these deep generative models with different generative architectures (like recurrent neural networks, autoencoders, and generative adversarial networks) based on various molecular representations (SMILES, molecular graph). Despite the limitations of these generative models and the inaccuracy of current evaluation techniques for these models [70], they are indeed one choice for de novo molecular generation.

In parallel, rule-based molecular generation is also very popular such as AbbVie's project Drug Guru [34, 35], the abbreviation of drug generation using rules. A

data-driven method called matched molecular pairs (MMPs) [71–73] is another way to collect the experts' knowledge from literature. Indeed, the rules of Drug Guru and MMPs are essentially the same method and nearly from the same source, that is molecular design thoughts of human beings. They can be stored as lines of reaction SMARTS code in RDKit. Scientists from Shanghai Institute of Materia Medica constructed DrugSpaceX [74], a virtual compound library, using Nova and BIOSTER rules from StarDrop. Scientists from GSK did a Turing test [75, 76] for molecular generators



**Fig. 8** Test data  $R^2$  of deep learning models. All the docking data are randomly split into three parts: training (80%), test (10%), and validation (10%) datasets. **A–E** shows the performance of models using docking data set from generation 1–5, respectively. **F** shows the performance of the aggregate dataset, including data from the previous five generations



**Fig. 9** Time consumption. Running time of each generation (GEN 1–5). DL represents the time cost of a final search by deep learning model. With 80 cores, a five-generation computing cost was 40.5 hours

by comparing three molecular generators in-house. The first one is BioDig, an MMPs-based algorithm. The second one is BRICS, a molecular generator by fragment recombination. The last one is RG2Smi, a deep generative model for generating molecules, which translates a molecule into a pharmacophore-based graph representation, then generates smiles string by deconvolution algorithm trained using natural language processing architecture. BioDig performed better than the other two methods across all tests in their report. Despite the fact that rule-based methods are somewhat limited to human knowledge or bias, we prefer rule-based methods for practical considerations.

Another challenge in computational de novo drug design is that compounds proposed by these tools are often hard to synthesize. Therefore, synthetic accessibility is a critical assessment for meaningful output. Previous retrosynthesis analysis tools were usually incapable

of handling complex synthetic routes. Recently, with the development of deep learning and the growing number of available reaction collections, several new algorithms [77–79] have been developed to improve the capability of synthetic route planning. We introduce synthetic accessibility evaluator of Chemical.AI to analyze the feasibility of generated structures. Yet increasing the evaluation throughput is challenging since it may take a few minutes to find practical routes for a single molecule. A batch mode that can evaluate thousands to millions of molecules at an affordable cost within a given timeframe is urgently needed.

SECSE mainly relies on structure-based computational design tools. Different tools will lead to different search directions, which might result in different chemical structure output. SECSE platforms are built to be compatible with various tools as fitness evaluators, like molecular docking, shape-based screening, and pharmacophore alignment, even ligand-based screening methods. Until now, docking has been the primary choice of SECSE because of its tradeoff between accuracy and speed. Despite the excellent performance of SECSE docking mode, there are still some inherent shortcomings in molecular docking methods, such as simplistic scoring with empirical energy function, rigid protein structures, ill-modeled poses. To enhance the prediction power, theoretically more accurate methods need to be introduced into the fitness assessment module. Claudio *et al.* [80] proposed a new QM-based docking program to replace the current docking methods based on molecular mechanics (MM) force fields. The new scoring function has achieved excellent performance in most cases. However, their QM docking scoring function is ten times slower than traditional MM-based scoring functional per core. To explicitly consider the dynamic nature of proteins, molecular dynamics (MD) simulation is the best choice. Hugo Guterres *et al.* [81] reported a high-throughput molecular dynamics (HTMD) simulations method to refine the docking results from AutoDock Vina. They calculated the RMSD of ligand by aligning protein structure from the initial docking pose and all protein structures in MD trajectory. They used a large set of 56 diverse target proteins and 560 ligands from the DUD-E dataset. The results show that short time MD simulations increase the area under the curve (AUC) of 0.8 from a value of 0.68 from AutoDock Vina. Enabled by the increasing computational power, attempts to add QM and MD concepts to the current docking program will be a promising way to improve the fitness evaluation module of SECSE. Results of the application of SECSE in our internal research projects will be reported in due course.

## Conclusion

We have developed a de novo design platform SECSE that integrates human intelligence for systemic evolutionary chemical space exploration against a specific protein pocket. The platform incorporated design rules of medicinal chemistry, computational evaluation methods, and deep learning models to efficiently speed up the search process of virtual hit compounds. The application in a demo target PHGDH proved its utility in finding diverse, potent and novel drug-like chemotypes. Further optimization considering high-precision evaluation methods and protein dynamics is currently underway. SECSE is released as an open-source project under the Apache License, Version 2.0. Any efforts and suggestions to improve its performance are welcomed.

## Abbreviations

HTS: High-throughput screening; AS-MS: Affinity selection-mass spectrometry; FBDD: Fragment-based drug design; DELT: DNA-encoded library technology; VS: Virtual screening; MW: Molecular weight; HBD: Number of hydrogen bond donors; HBA: Number of hydrogen bond acceptor; RB: Number of rotatable bonds; TPSA: Topological polar surface area; RMSD: Root Mean Square Deviation; DL: Deep learning; PDB: Protein Data Bank; Fsp3: Fraction of *sp*<sup>3</sup> hybridized carbons; PMI: Principal moments of inertia; PHGDH: Phosphoglycerate Dehydrogenase; NADH: Reduced nicotinamide adenine dinucleotide; NAD<sup>+</sup>: Oxidized nicotinamide adenine dinucleotide; LRE: Long-range electrostatic; MMP: Matched molecular pair; MM: Molecular mechanics; MD: Molecular dynamics; QM: Quantum mechanics; HTMD: High-throughput molecular dynamics; AUC: Area under curve; SAR: Structure-activity relationship; SA: Synthetic accessibility.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13321-022-00598-4>.

**Additional file 1:** Fragment library generation and clustering algorithms.

**Additional file 2: Table S1.** SMILES and related information of molecules from Table 1.

**Additional file 3: Table S2.** SMILES and related information of molecules from Table 2.

**Additional file 4:** Rules for regeneration test in SMARTS format.

## Acknowledgements

We thank all of our colleagues at Keen Therapeutics, who have contributed to this work. We appreciate Chemical.AI for their help. Finally, we thank Dr. Yingxiao Cai for assistance with SA evaluation.

## Availability and requirements

Project name: Systemic Evolutionary Chemical Space Explorer.

Project home page: <https://github.com/KeenThera/SECSE>.

Operating system: Linux.

Programming language: Python, Shell.

License: Apache License, Version 2.0.

## Authors' contributions

CL and SL are co-first authors and contributed equally to this work. YW initiated this project. WS, JY, and YW collected and designed the growing rules

and filters. CL and SL wrote the code, tested its performance, and finished the demo case. NX evaluated the SA score of selected structures. All authors were involved in the analysis and discussion of results. CL and SL prepared the first version of this manuscript. ZZ, XZ, XL, FC, and NX assisted in the preparation of the manuscript. Y.W. revised this manuscript. All authors approved the final manuscript.

#### Funding

Not applicable.

#### Data availability

Fragment Library: <https://doi.org/10.6084/m9.figshare.17142236>.  
Random Library: <https://doi.org/10.6084/m9.figshare.19142624>.  
Raw data of regeneration test : <https://doi.org/10.6084/m9.figshare.19235217>.  
Raw data of demo case PHGDH: <https://doi.org/10.6084/m9.figshare.17141879>.

#### Declarations

#### Competing interests

The authors declare that they have no competing interests.

#### Author details

<sup>1</sup>Keen Therapeutics Co., Ltd., Shanghai, China. <sup>2</sup>Chemical.AI, Shanghai, China.

Received: 13 December 2021 Accepted: 11 March 2022

Published online: 01 April 2022

#### References

- MAGE Building blocks from Enamine. <https://enamine.net/building-blocks/make-on-demand-building-blocks>. Accessed 1 Dec 2021
- Lyu J, Wang S, Balias TE, Singh I, Levit A, Moroz YS, O'Meara MJ, Che T, Alga E, Tolmachova K, Tolmachev AA, Shoichet BK, Roth BL, Irwin JJ (2019) Ultra-large library docking for discovering new chemotypes. *Nature* 566(7743):224–229. <https://doi.org/10.1038/s41586-019-0917-9>
- Bender BJ, Gahbauer S, Luttens A, Lyu J, Webb CM, Stein RM, Fink EA, Balias TE, Carlsson J, Irwin JJ, Shoichet BK (2021) A practical guide to large-scale docking. *Nat protoc* 16:1–34
- Warr W (2021). Report on an NIH Workshop on Ultralarge Chemistry Databases. *ChemRxiv*. <https://doi.org/10.26434/chemrxiv.14554803.v1>
- BioSolveIT: Efficient 3D exploration of multi-billion compound spaces. *BioSolveIT*. [https://cactus.nci.nih.gov/presentations/NIHBigDB\\_2020-12/ChristianLemmen4NIHWorkshop.pdf](https://cactus.nci.nih.gov/presentations/NIHBigDB_2020-12/ChristianLemmen4NIHWorkshop.pdf). Accessed 01 Dec 2021
- Bohacek RS, McMartin C, Guida WC (1996) The art and practice of structure-based drug design: a molecular modeling perspective. *Med Res Rev* 16(1):3–50. [https://doi.org/10.1002/\(SICI\)1098-1128\(199601\)16:1<3::AID-MED1>3.0.CO;2-6](https://doi.org/10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6)
- Schneider G, Fechner U (2005) Computer-based de novo design of drug-like molecules. *Nat Rev Drug Discov* 4(8):649–663. <https://doi.org/10.1038/nrd1799>
- Hartenfeller M, Schneider G, Bajorath J (2011) De novo drug design. *Humana Press, Totowa*, pp 299–323. [https://doi.org/10.1007/978-1-60761-839-3\\_12](https://doi.org/10.1007/978-1-60761-839-3_12)
- Schneider G, Clark DE (2019) Automated de novo drug design: are we nearly there yet? *Angew Chem Int Ed* 58(32):10792–10803. <https://doi.org/10.1002/anie.201814681>
- Mouchlis VD, Afantitis A, Serra A, Fratello M, Papadiamantis AG, Aidinis V, Lynch I, Greco D, Melagraki G (2021) Advances in de novo drug design: From conventional to machine learning methods. *Int J Mol Sci* 22(4):1–22. <https://doi.org/10.3390/ijms22041676>
- Dollar O, Joshi N, Beck DAC, Pfaendtner J (2021) Attention-based generative models for: de novo molecular design. *Chem Sci* 12(24):8362–8372. <https://doi.org/10.1039/d1sc01050f>
- Böhm HJ (1992) LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J Comput Aided Mol Des* 6(6):593–606. <https://doi.org/10.1007/BF00126217>
- Wang R, Gao Y, Lai L (2000) LigBuilder: a multi-purpose program for structure-based drug design. *J Mol Model* 6(7–8):498–516. <https://doi.org/10.1007/s0089400060498>
- Chen J, Lai L (2006) Pocket vol 2: further developments on receptor-based pharmacophore modeling. *Journal of Chem Inform Model* 46(6):2684–2691. <https://doi.org/10.1021/ci600246s>
- Yuan Y, Pei J, Lai L (2020) LigBuilder V3: a multi-target de novo drug design approach. *Front Chem* 8(5):1083–1091. <https://doi.org/10.3389/fchem.2020.00142>
- Li Y, Zhao Y, Liu Z, Wang R (2011) Automatic tailoring and transplanting: a practical method that makes virtual screening more useful. *J Chem Inform Model* 51(6):1474–1491. <https://doi.org/10.1021/ci200036m>
- Li Y, Zhao Z, Liu Z, Su M, Wang R (2016) AutoT&T vol 2: an efficient and versatile tool for lead structure generation and optimization. *J Chem Inform Model* 56(2):435–453. <https://doi.org/10.1021/acs.jcim.5b00691>
- Chéron N, Jasty N, Shakhnovich EI (2016) OpenGrow: an automated and rational algorithm for finding new protein ligands. *J Med Chem* 59(9):4171–4188. <https://doi.org/10.1021/acs.jmedchem.5b00886>
- Durrant JD, Amaro RE, McCammon JA (2009) AutoGrow: a novel algorithm for protein inhibitor design. *Chem Biol Drug Design* 73(2):168–178. <https://doi.org/10.1111/j.1747-0285.2008.00761.x>
- Spiegel JO, Durrant JD (2020) AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization. *Cheminform* 12(1):1–16. <https://doi.org/10.1186/s11862-020-00429-4>
- Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2012) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 64(Suppl):4–17. <https://doi.org/10.1016/j.addr.2012.09.019>
- Baell JB, Holloway GA (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J Med Chem* 53(7):2719–2740. <https://doi.org/10.1021/jm901137j>
- Polishchuk P (2020) CReM: chemically reasonable mutations framework for structure generation. *J Cheminform* 12(1):28. <https://doi.org/10.1186/s13321-020-00431-w>
- Nigam A, Pollice R, Krenn M, Gomes GDP, Aspuru-Guzik A (2021) Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (STONED) algorithm for molecules using SELFIES. *Chem Sci* 12(20):7079–7090. <https://doi.org/10.1039/d1sc00231g>
- Steinmann C, Jensen JH (2021) Using a genetic algorithm to find molecules with good docking scores. *PeerJ Phys Chem* 3:18. <https://doi.org/10.7717/peerj-pchem.18>
- Bai Q, Tan S, Xu T, Liu H, Huang J, Yao X (2021) MolAICal: A soft tool for 3D drug design of protein targets by artificial intelligence and classical algorithm. *Brief Bioinform* 22(3):161. <https://doi.org/10.1093/bib/bbaa161>
- Ma B, Terayama K, Matsumoto S, Isaka Y, Sasakura Y, Iwata H, Araki M, Okuno Y (2021) Structure-based de novo molecular generator combined with artificial intelligence and docking simulations. *J Chem Inform Model* 61(7):3304–3313. <https://doi.org/10.1021/acs.jcim.1c00679>
- Yang X, Zhang J, Yoshizoe K, Terayama K, Tsuda K (2017) ChemTS: an efficient python library for de novo molecular generation. *Sci Technol Adv Mater* 18(1):972–976. <https://doi.org/10.1080/14686996.2017.1401424>
- Li Y, Pei J, Lai L (2021) Structure-based de novo drug design using 3D deep generative models. *Chem Sci* 12(41):13664–13675. <https://doi.org/10.1039/d1sc04444c>
- Gebauer NWA, Gastegger M, Schütt KT (2019) Symmetry-adapted generation of 3D point sets for the targeted discovery of molecules. *Adv Neural Inform Process Syst*. 32 (2019). <https://arxiv.org/abs/1906.00957> <https://arxiv.org/abs/1906.00957>
- Imrie F, Bradley AR, Van Der Schaar M, Deane CM (2020) Deep generative models for 3D linker design. *J Chem Inform Model* 60(4):1983–1995. <https://doi.org/10.1021/acs.jcim.9b01120>
- Green H, Koes DR, Durrant JD (2021) DeepFrag: a deep convolutional neural network for fragment-based lead optimization. *Chemical Science* 12(23):8036–8047. <https://doi.org/10.1039/d1sc00163a>
- Nesterov V, Wieser M, Roth V (2020) 3DMolNet: A generative network for molecular structures. *arXiv*. <https://arxiv.org/abs/2010.06477>

34. Stewart KD, Shiroda M, James CA (2006) Drug Guru: a computer software program for drug design using medicinal chemistry rules. *Bioorgan Med Chem* 14(20):7011–7022. <https://doi.org/10.1016/j.bmc.2006.06.024>
35. Stewart KD, Shanley J, Ahmed KBA, Bowen JP (2012) The drug guru project, Chap. 11. vol. 54, pp 183–198. West Sussex: John Wiley. <https://onlinelibrary.wiley.com/doi/abs/10.1002/9783527654307.ch11>
36. Blum LC, Reymond JL (2009) 970 Million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J Am Chem Soc* 131(25):8732–8733. <https://doi.org/10.1021/ja902302h>
37. RDKit: Open-source cheminformatics. <http://www.rdkit.org>. Accessed 13 Oct 2021
38. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open babel: an open chemical toolbox. *J Cheminform* 3(10):1–14. <https://doi.org/10.1186/1758-2946-3-33>
39. Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, Christie CH, Dalenberg K, Di Costanzo L, Duarte JM, Dutta S, Feng Z, Ganesan S, Goodsell DS, Ghosh S, Green RK, Guranovic V, Guzenko D, Hudson BP, Lawson CL, Liang Y, Lowe R, Namkoong H, Peisach E, Persikova I, Randle C, Rose A, Rose Y, Sali A, Segura J, Sekharan M, Shao C, Tao YP, Voigt M, Westbrook JD, Young JY, Zardecki C, Zhuravleva M (2021) RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res* 49(D1):437–451. <https://doi.org/10.1093/NAR/GKAA1038>
40. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Židek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596(7873):583–589. <https://doi.org/10.1038/s41586-021-03819-2>
41. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Dustin Schaeffer R, Millán C, Park H, Adams C, Glassman CR, DeGiovanni A, Pereira JH, Rodrigues AV, Van Dijk AA, Ebrecht AC, Opperman DJ, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy MK, Dalwadi U, Yip CK, Burke JE, Christopher Garcia K, Grishin NV, Adams PD, Read RJ, Baker D (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373(6557):871–876. <https://doi.org/10.1126/science.abj8754>
42. Ravindranath PA, Forli S, Goodsell DS, Olson AJ, Sanner MF (2015) AutoDockFR: advances in protein-ligand docking with explicitly specified binding site flexibility. *PLoS Comput Biol* 11(12):1–28. <https://doi.org/10.1371/journal.pcbi.1004586>
43. Ravindranath PA, Sanner MF (2016) AutoSite: an automated approach for pseudo-ligands prediction-from ligand-binding sites identification to predicting key ligand atoms. *Bioinformatics* 32(20):3142–3149. <https://doi.org/10.1093/bioinformatics/btw367>
44. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL (2012) Quantifying the chemical beauty of drugs. *Nat Chem* 4(2):90–98. <https://doi.org/10.1038/nchem.1243>
45. Trott O, Olson AJ (2009) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*. <https://doi.org/10.1002/jcc.21334>
46. Eberhardt J, Santos-Martins D, Tillack AF, Forli S (2021) AutoDock Vina 1.2.0: new docking methods, expanded force field, and python bindings. *J Chem Inform Model* 61(8):3891–3898. <https://doi.org/10.1021/acs.jcim.1c00203>
47. Kuntz ID, Chen K, Sharp KA, Kollman PA (1999) The maximal affinity of ligands. *Tech Rep*. <https://doi.org/10.1073/pnas.96.18.9997>
48. How is Ligand Efficiency calculated? Schrödinger, Inc. <https://www.schrodinger.com/kb/1622>. Accessed October 13, 2021
49. Chemical.AI. Wuhan Zhihua Technology Co., Ltd. <https://chemical.ai>. Accessed October 13 Oct 2021
50. Goh GK-m, Foster JA (1999) Evolving molecules for drug design using genetic algorithms via molecular trees. In: Proceedings of the Genetic and Evolutionary Computation Conference. pp. 27–33
51. Goldberg DE, Holland JH (1988) Genetic algorithms and machine learning. *Mach Learn* 3(2):95–99. <https://doi.org/10.1023/A:1022602019183>
52. Rocke DM, Michalewicz Z (2000) Genetic Algorithms + Data Structures = Evolution Programs. vol. 95, p. 347. New York; Springer. <https://doi.org/10.2307/2669583>
53. Yang K, Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., Barzilay, R (2019) Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling* 59(8), 3370–3388. <https://doi.org/10.1021/acs.jcim.9b00237>. arXiv:1904.01561
54. Tange O (2011) GNU parallel - the command-line power tool. In: login. The USENIX Magazine, vol 36, pp. 42–47. <https://doi.org/10.5281/zenodo.16303>
55. Lovering F, Bikker J, Humblet C (2009) Escape from flatland: increasing saturation as an approach to improving clinical success. *J Med Chem* 52(21):6752–6756. <https://doi.org/10.1021/jm901241e>
56. Sauer WHB, Schwarz MK (2003) Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity. *J Chem Inform Comput Sci* 43(3):987–1003. <https://doi.org/10.1021/ci025599w>
57. Zhang B, Zheng A, Hydrbring P, Ambrose G, Ouchida AT, Goiny M, Vakifahmetoglu-Norberg H, Norberg E (2017) PHGDH defines a metabolic subtype in lung adenocarcinomas with poor prognosis. *Cell Rep* 19(11):2289–2303. <https://doi.org/10.1016/j.celrep.2017.05.067>
58. Rathore R, Schutt CR, van Tine BA (2020) PHGDH as a mechanism for resistance in metabolically-driven cancers. *Cancer Drug Resist* 3(4):762–774. <https://doi.org/10.20517/cdr.2020.46>
59. Zhao JY, Feng KR, Wang F, Zhang JW, Cheng JF, Lin GQ, Gao D, Tian P (2021) A retrospective overview of PHGDH and its inhibitors for regulating cancer metabolism. *Eur J Med Chem*. <https://doi.org/10.1016/j.ejmech.2021.113379>
60. Reid MA, Allen AE, Liu S, Liberti MV, Liu P, Liu X, Dai Z, Gao X, Wang Q, Liu Y, Lai L, Locasale JW (2018) Serine synthesis through PHGDH coordinates nucleotide levels by maintaining central carbon metabolism. *Nat Commun* 9(1):1–11. <https://doi.org/10.1038/s41467-018-07868-6>
61. Mullarky E, Xu J, Robin AD, Huggins DJ, Jennings A, Noguchi N, Olland A, Lakshminarasimhan D, Miller M, Tomita D, Michino M, Su T, Zhang G, Stamford AW, Meinke PT, Kargman S, Cantley LC (2019) Inhibition of 3-phosphoglycerate dehydrogenase (PHGDH) by indole amides abrogates de novo serine synthesis in cancer cells. *Bioorg Med Chem Lett* 29(17):2503–2510. <https://doi.org/10.1016/j.bmcl.2019.07.011>
62. ...Weinstabl H, Treu M, Rinnenthal J, Zahn SK, Etmayer P, Bader G, Dahmann G, Kessler D, Rumpel K, Mischerikow N, Savarese F, Gerstberger T, Mayer M, Zoepfel A, Schnitzer R, Sommergruber W, Martinelli P, Arnhof H, Peric-Simov B, Hofbauer KS, Garavel G, Scherbantini Y, Mitzner S, Fett TN, Scholz G, Bruchhaus J, Burkard M, Kousek R, Ciftci T, Sharps B, Schrenk A, Harrer C, Haering D, Wolkerstorfer B, Zhang X, Lv X, Du A, Li D, Li Y, Quant J, Pearson M, McConnell DB (2019) Intracellular trapping of the selective phosphoglycerate dehydrogenase (PHGDH) inhibitor BI-4924 disrupts serine biosynthesis. *J Med Chem* 62(17):7976–7997. <https://doi.org/10.1021/acs.jmedchem.9b00718>
63. Unterlass JE, Baslé A, Blackburn TJ, Tucker J, Cano C, Noble MEM, Curtin NJ, Unterlass JE, Baslé A, Blackburn TJ, Tucker J, Cano C, Noble MEM, Curtin NJ (2016) Validating and enabling phosphoglycerate dehydrogenase (PHGDH) as a target for fragment-based drug discovery in PHGDH-amplified breast cancer. *Oncotarget* 9(17):13139–13153. <https://doi.org/10.18632/oncotarget.11487>
64. Yang Y, Yao K, Repasky MP, Leswing K, Abel R, Shoichet BK, Jerome SV (2021) Efficient exploration of chemical space with docking and deep learning. *J Chem Theory Comput*. <https://doi.org/10.1021/acs.jctc.1c00810>
65. Gentile F, Agrawal V, Hsing M, Ton AT, Ban F, Norinder U, Gleave ME, Cherkasov A (2020) Deep docking: a deep learning platform for augmentation of structure based drug discovery. *ACS Cent Sci* 6(6):939–949. <https://doi.org/10.1021/acscentsci.0c00229>
66. Choi J, Lee J (2021) V-dock: fast generation of novel drug-like molecules using machine-learning-based docking score and molecular optimization. *Int J Mol Sci*. <https://doi.org/10.3390/ijms222111635>
67. Fink T, Bruggesser H, Reymond JL (2005) Virtual exploration of the small-molecule chemical universe below 160 daltons. *Angew Chem Int Ed* 44(10):1504–1508. <https://doi.org/10.1002/anie.200462457>
68. Ruddigkeit L, Van Deursen R, Blum LC, Reymond JL (2012) Enumeration of 166 billion organic small molecules in the chemical universe database



- GDB-17. *J Chem Inform Model* 52(11):2864–2875. <https://doi.org/10.1021/ci300415d>
69. Sousa T, Correia J, Pereira V, Rocha M (2021) Generative deep learning for targeted compound design. *J Chem Inform Model*. <https://doi.org/10.1021/acs.jcim.0c01496>
70. Renz P, Van Rompaey D, Wegner JK, Hochreiter S, Klambauer G (2019) On failure modes in molecule generation and optimization. *Drug Discov Today Technol* 32–33:55–63. <https://doi.org/10.1016/j.ddtec.2020.09.003>
71. Warner DJ, Griffen EJ, St-Gallay SA (2010) WisePairZ: A novel algorithm to identify, encode, and exploit matched molecular pairs with unspecified cores in medicinal chemistry. *J Chem Inform Model* 50(8):1350–1357. <https://doi.org/10.1021/ci100084s>
72. Hussain J, Rea C (2010) Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J Chem Inform Model* 50(3):339–348. <https://doi.org/10.1021/ci900450m>
73. Awale M, Hert J, Guasch L, Riniker S, Kramer C (2021) The playbooks of medicinal chemistry design moves. *J Chem Inform Model* 61(2):729–742. <https://doi.org/10.1021/acs.jcim.0c01143>
74. Yang T, Li Z, Chen Y, Feng D, Wang G, Fu Z, Ding X, Tan X, Zhao J, Luo X, Chen K, Jiang H, Zheng M (2021) DrugSpaceX: a large screenable and synthetically tractable database extending drug space. *Nucleic Acids Res* 49(D1):1170–1178. <https://doi.org/10.1093/nar/gkaa920>
75. Green DVS, Pickett S, Luscombe C, Senger S, Marcus D, Meslamani J, Brett D, Powell A, Masson J (2020) BRADSHAW: a system for automated molecular design. *J Comput Aided Mol Design* 34(7):747–765. <https://doi.org/10.1007/s10822-019-00234-8>
76. Bush JT, Pogany P, Pickett SD, Barker M, Baxter A, Campos S, Cooper AWJ, Hirst D, Inglis G, Nadin A, Patel VK, Poole D, Pritchard J, Washio Y, White G, Green DVS (2020) A turing test for molecular generators. *J Med Chem* 63(20):11964–11971. <https://doi.org/10.1021/acs.jmedchem.0c01148>
77. Coley CW, Rogers L, Green WH, Jensen KF (2017) Computer-assisted retrosynthesis based on molecular similarity. *ACS Cent Sci* 3(12):1237–1245. <https://doi.org/10.1021/acscentsci.7b00355>
78. Segler MHS, Preuss M, Waller MP (2018) Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* 555(7698):604–610. <https://doi.org/10.1038/nature25978>
79. Genheden S, Thakkar A, Chadimová V, Reymond JL, Engkvist O, Bjerrum E (2020) AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J Cheminform* 12(1):1–9. <https://doi.org/10.1186/s13321-020-00472-1>
80. Cavasotto CN, Aucar MG (2020) High-throughput docking using quantum mechanical scoring. *Front Chem* 8:246. <https://doi.org/10.3389/fchem.2020.00246>
81. Guterres H, Im W (2020) Improving protein-ligand docking results with high-throughput molecular dynamics simulations. *J Chem Inform Model* 60(4):2189–2198. <https://doi.org/10.1021/acs.jcim.0c00057>

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

