# DeepBioisostere: Deep Learning-based Bioisosteric Replacements for Optimization of Multiple Molecular Properties

**Woo Youn Kim**

wooyoun@kaist.ac.kr

KAIST https://orcid.org/0000-0001-7152-2111

**Hyeongwoo Kim**

KAIST https://orcid.org/0009-0002-0376-8616

**Seokhyun Moon**

KAIST https://orcid.org/0000-0002-5197-0096

**Wonho Zhung**

KAIST https://orcid.org/0000-0001-7499-2063

**Jaechang Lim**

HITS Incorporation https://orcid.org/0000-0001-7342-4283

**Article**

**Keywords:**

**Posted Date:** July 9th, 2024

**DOI:** https://doi.org/10.21203/rs.3.rs-4225249/v1



**Additional Declarations:** There is **NO** Competing Interest.

# DeepBioisostere: Deep Learning-based Bioisosteric Replacements for Optimization of Multiple Molecular Properties

Hyeongwoo Kim[1†], Seokhyun Moon[1†], Wonho Zhung[1], Jaechang Lim[2], Woo Youn Kim[1,2,3*]

[1]Department of Chemistry, KAIST, 291 Daehak-ro, Yuseong-gu, 34141, Daejeon, Republic of Korea.
[2]HITS Inc., 124 Teheran-ro, Gangnam-gu, 06234, Seoul, Republic of Korea.
[3]AI Institute, KAIST, 291 Daehak-ro, Yuseong-gu, 34141, Daejeon, Republic of Korea.

*Corresponding author(s). E-mail(s): wooyoun@kaist.ac.kr;
†These authors contributed equally to this work.

**Abstract**

Optimizing molecules to improve their properties is a fundamental challenge in drug design. For a fine-tuning of molecular properties without losing bio-activity validated in advance, the concept of bioisosterism has emerged. Many in silico methods have been proposed for discovering bioisosteres, but they require expert knowledge for their applications or are restricted to known databases. Here, we introduce DeepBioisostere, a deep generative model to design suitable bioisosteric replacements. Our model allows a fully automated bioisosteric replacement by intelligently selecting fragments for both removal and insertion. Through various scenarios of multiple property control, we showcase the model's capability to delicately tune specific properties, addressing the challenge in molecular optimization. Our model's innovation lies in its capacity to design a bioisosteric replacement reflecting the compatibility with the surroundings of the modification site, facilitating the control of sophisticated properties that depend on the overall molecular structures such as drug-likeness. DeepBioisostere can also provide previously unseen bioisosteric replacements, highlighting its capability for exploring diverse chemical modifications rather than just mining them from known databases. Lastly, we used DeepBioisostere to improve the sensitivity of a known SARS-CoV-2 main protease inhibitor to the E166V mutant that exhibits

drug resistance to the inhibitor, demonstrating its potential application in drug discovery.

# 1 Introduction

Optimizing molecules to fulfill multiple properties is a challenging yet fundamental goal of drug design. A drug should be bioactive, bio-compatible, and synthetically accessible,[1] taking an enormous time and cost to bring a compound from a lab to a patient.[2] One promising approach to achieving this multi-objective molecular design problem is modifying a chemical structure from an acclaimed molecule with satisfied properties. Namely, chemical modifications, such as scaffold hopping[3, 4] or R-group replacement,[5] aim to delicately enhance a target property while maintaining the other properties of the original structure, where a comprehensive consideration of both structural and biochemical analogies is necessary. To develop a more efficient way of chemical modification, many chemists attempted to standardize useful rules that are frequently adopted in molecular optimization, leading to the emergence of the concept of bioisosteric replacement.[6–8] Numerous chemical moiety pairs involved in bioisosteric replacements, namely the bioisosteres, have been reported for the last decades and facilitated more proper chemical modifications.[9–11]

Traditional bioisosteric replacements heavily relied on the intuition and experience of medicinal chemists.[12] Identification of proper bioisosteric replacements has required extensive trial and error, leading to the rise of in silico approaches.[13, 14] There have been two mainstream conventional approaches to finding bioisosteres—similarity-based and database mining methods. Similarity-based approaches operate on the assumption that structural analogs would exhibit similar biological or chemical properties.[15] They prioritize potential bioisosteres by comparing their structural, steric, or electronic features to the original moiety of interest.[16–19] Diverse molecular descriptors such as fingerprints[20, 21] or structural alignments[22] were often employed to assess the similarity between chemical moieties. The similarity-based methods facilitated the expectation of biochemical property changes during a particular chemical modification before proceeding to experimental validation.

On the other hand, database mining approaches identify potential bioisosteres through the extraction from large chemical databases.[23–25] (see Fig. 1) They systematically enumerate pairs of molecules with only a small structural difference, the so-called matched molecular pairs(MMPs),[26] from a public biochemical library such as ChEMBL.[27] Then, molecular fragments corresponding to the difference in each MMP are paired and identified as potential bioisosteric replacements along with their occurrences in the parent library. The differences in chemical properties and bioactivity in biochemical assays are also used to expect the effect of each bioisosteric replacement, which is essential for subsequent bioisostere recommendations. This statistical evidence enables the identification of a new bioisosteric replacement, even if the involved moieties are structurally unrelated at first glimpse.

Despite each approach's advantages, limitations still exist and hamper the identification of a proper chemical modification. The similarity-based methods highly depend on the correlation between carefully crafted molecular descriptors and a specific property to be controlled. Yet, finding suitable descriptors from a variety of available options[28] inevitably relies on expertise and is often arbitrary. The similarity-based methods may also struggle to identify bioisosteres that are structurally dissimilar but able to modulate particular molecular properties in a desired way. The database mining methods, on the other hand, cannot assess bioisosterism between unseen chemical moiety pairs that do not appear in the parent database due to their statistical absence. Also, the expectation of property changes based on the mined data may not work for more complicated properties that depend on complex relationships between internal moieties, such as synthesizability or drug-likeness. Moreover, although both methods propose potential bioisosteres once a target molecular fragment is assigned, the decision on which substructure to modify still relies on the experts' assessment of target properties and molecular structures. Since the design of bioisosteric replacements involves selecting which molecular fragment to modify and determining its proper bioisosteres, neither the similarity-based nor database mining approaches can fully address a fully automated process based on bioisosteric replacements.

Recently, deep learning-based chemical modification approaches have stood out to featurize data-intrinsic information that contains chemical and biological contexts without handcrafted descriptors.[29–36] Ertl first used a deep learning approach in identifying potential bioisosteres.[29] In this work, substructural information constructed from ChEMBL was used to train a neural network on the inherent relationship between analogous fragments. Remarkably, this model could propose bioisosteric relationships not present in the training dataset, serving as novel bioisostere candidates. In parallel, a number of approaches that directly model the chemical modification task with deep generative models have appeared. They regard the chemical modification problem as a language translation[30–32] or a graph-editing task[33–35] by representing a molecule as a text or a molecular graph, respectively. The advantage of applying deep generative modeling is that multiple property changes can be accounted for by designing chemical modifications, e.g., changing solubility while maintaining molecular similarity. Furthermore, they could identify novel modified structures with improved biochemical properties by learning the intrinsic relationships between molecular structures and their analogs. Based on these advantages, deep generative modeling is expected to significantly contribute to overcoming the limitations of the conventional approaches for identifying and prioritizing bioisosteres.

Here, we propose DeepBioisostere, a deep generative model for multi-property control with designing suitable bioisosteric replacements in a fully automated fashion. Our model optimizes a molecular structure in the following strategy: selecting a target substructure to be removed and inserting a proper bioisostere to achieve target properties. With the sequential modeling of bioisosteric replacement, our model automates the overall process of chemical modification while the previous similarity and database mining approaches aim to provide bioisosteres for a moiety selected by experts in advance. Our step-wise strategy enables more flexible suggestions of novel bioisosteric replacements even if they are not included in training data. By utilizing a

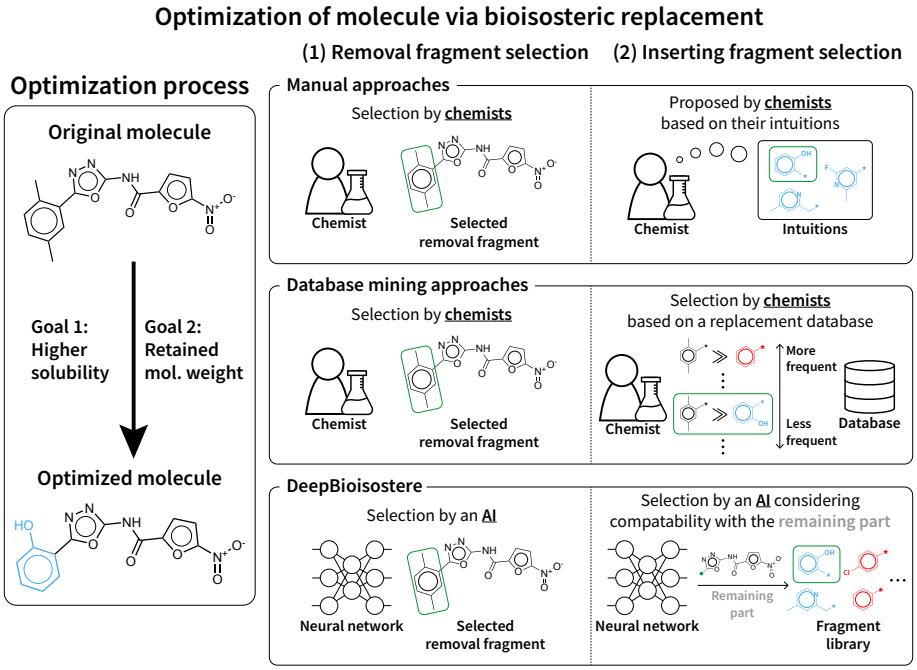


**Fig. 1 An illustration of various bioisosteric replacement approaches.** The manual approaches rely on the knowledge of experts throughout the entire optimization process. With an appropriate removal fragment selected by experts, one can employ conventional database mining approaches to search various potential bioisosteres. Still, the intuition and experience of experts are applied to selecting a proper replacement moiety to achieve the property optimization goals. Unlike the previous approaches, which inevitably depend on experts, DeepBioisostere automates the sequential decision-making process based on bioisosterism using deep learning. By taking into account the entire molecular structure, DeepBioisostere can flexibly suggest modified molecules that satisfy the delicate property control conditions.

deep learning-based approach for chemical modification, DeepBioisostere could control complicated properties such as synthetic accessibility and drug-likeness simultaneously despite the difficulty due to their dependency on an overall molecular structure. We demonstrate with a case study that the compatibility between an insertion moiety and its surroundings to be attached is crucial for multiple property control, which can be comprehensively considered by DeepBioisostere. Lastly, we adopted DeepBioisostere to the challenging problem of treating drug resistance due to a point mutation of SARS-CoV-2 main protease, implying its potential application in the lead optimization process. To our best knowledge, this is the first work that proposes an in silico method to design the complete procedure of bioisosteric replacement without any aid of expertise, facilitating an efficient lead optimization process. (see Fig. 1)

# 2 Results

## 2.1 Brief introduction of DeepBioisostere

In this work, we regard chemical modification based on bioisosterism as a sequential process comprising three essential steps: (1) removal fragment selection, (2) insertion fragment selection, and (3) attachment orientation prediction. These three steps are sequentially applied to molecular optimization. To achieve desirable structural modification, target biochemical properties dictated by the molecular structure should be considered during the three steps. For this purpose, we devised a deep generative model named DeepBioisostere to design a bioisosteric replacement for the multi-property control of a particular molecule.

The DeepBioisostere model is trained on an MMP dataset with structural analogs with small differences, which we constructed from the ChEMBL database. Since we designed the model with three main modules corresponding to each step for the chemical modification, it learns proper ways of removing and then inserting fragments to meet the given condition of multiple biochemical properties. This sequential modification facilitates the optimization of a molecule while retaining the overall properties except the target ones to be improved. The DeepBioisostere model comprises two embedding networks at different levels—atom and fragment. This unique feature, especially the latter network, allows the model to learn the compatibility between an insertion fragment and the remaining parts, enabling the fine control of complex properties depending on a whole molecular structure. For synthetically accessible chemical modification, the DeepBioisostere model enumerates molecular fragments inside a given molecule based on the breaking retrosynthetically interesting chemical substructures(BRICS)[37] rules and selects one of them as a removal fragment. Then, the model selects an insertion fragment from a pre-defined fragment library by examining all of its items in terms of their compatibility with the remaining part of a given molecule and the target property control condition. More details about the training and generation processes of the DeepBioisostere model can be found in Section 4.

## 2.2 Multi-property Control with DeepBioisostere

We first apply our DeepBioisostere model to multi-property control tasks. During the optimization of a potent molecule in drug developments, its structure is modified and elaborated to improve biochemical properties, for example, lipophilicity, water solubility, synthetic accessibility, and drug-likeness. This so-called multi-property control is more challenging than controlling a single property since, for example, the lipophilicity can be increased simply by elongating a carbon chain, but the modified structure is often not drug-like.[38] Thus, a comprehensive consideration of the whole ligand structure and its properties is required to modulate multiple properties simultaneously.

Here, the goal is to provide modified molecular structures to meet a given multi-property condition. We tested our model for three scenarios that practitioners might encounter in a lead optimization process: (1) increasing or decreasing ligand lipophilicity while maintaining its molecular size, (2) increasing drug-likeness while maintaining

its molecular weight, and (3) alleviating the synthetic complexity of a ligand with fair bio-activities but low synthetic accessibility. Obviously, the second and third scenarios are more challenging than the first one, which requires more comprehensive considerations about the relationship between the properties of interest. Throughout Section 2.2, the following criteria are commonly employed for selecting test molecules: (1) a single neutral molecule, (2) absence of more than three rings, (3) inclusion of at least one BRICS bond, and (4) molecular weight falling within the range of 250 to 500. In addition, to examine our model's capability to find novel bioisosterism, we only used fragments from the test dataset for insertion, which were never adopted as ground truths nor negative samples for the insertion step in the training process.

The overall property control results are exhibited in Table 1. For scenario 1, we selected 100 unique molecules whose log$P$ values are within the range of 1 to 5, regarded as moderate lipophilicity. Whether one should increase a molecule's log$P$ or not depends on its specific application; for example, the biological target under consideration or the drug delivery plan are critical factors in making such a decision. Thus, we tested our model for both cases with two property control conditions: increasing log$P$ by 1 and decreasing log$P$ by 1 while maintaining overall molecule size in terms of molecular weight(MW). log$P$ is a logarithm of the octanol-water partition coefficient, so the difference by 1 corresponds to a 10 times larger or smaller solubility ratio.[39] We generated 100 unique molecular structures for each original molecule according to the two property control conditions. For a few test molecules, DeepBioisostere generated less than 100 molecular structures for each one since the BRICS-allowed insertion fragments were less than 100. The results are summarized in the first row on Table 1. We noticed that the mean values of the molecular weight differences between an original molecule and its modified ones were +6.60 and +3.34, respectively. These values suggest that the molecular weights of most modified structures are different from the original ones by less than one heavy atom. (see Supplementary Figure 4 for more details of property distributions) Meanwhile, log$P$ was modulated with averaged changes of +0.81 and -0.86, falling slightly behind the requested property values. These statistics indicate that our model successfully adjusted the log$P$ values of the given molecular structures with barely changing their molecular weights, which requires careful selection of both removal and insertion fragments.

Second, we demonstrated an optimization process where a molecular structure is modified to improve its drug-likeness in terms of QED.[40] (scenario 2 in Table 1) In this scenario, only molecules whose QED values lie in the range of 0.4 to 0.6 were used as test molecules. Property control conditions were set to increase the QED value by 0.1 or 0.2 while keeping molecular weight constant. By employing the DeepBioisostere model, we were able to obtain novel molecules with enhanced drug-likeness; the mean values of QED changes were +0.064 and +0.087, respectively. QED is calculated based on eight physicochemical descriptors, including the molecular weight. Thus, finely modulating QED while preserving molecular weight is challenging since the other descriptors, such as the number of hydrogen bond donors and acceptors, rotatable bonds, and aromatic rings, should be adjusted in the desired way. The result in scenario 2 implies that our model can capture information to modulate those descriptors for increasing the QED value. Although the property control condition

| Scenario | Requested Optimization | | Property Change | |
|---|---|---|---|---|
| | Property | ΔValue | Average | Std. Dev. |
| Scenario 1 | $MW$ | 0 | +6.60 | 14.0 |
| | log$P$ | +1 | +0.81 | 0.35 |
| | $MW$ | 0 | +3.34 | 13.0 |
| | log$P$ | -1 | -0.86 | 0.37 |
| Scenario 2 | $MW$ | 0 | +2.73 | 12.7 |
| | QED | +0.1 | +0.064 | 0.098 |
| | $MW$ | 0 | +1.90 | 12.4 |
| | QED | +0.2 | +0.087 | 0.101 |
| Scenario 3 | QED | 0 | -0.049 | 0.092 |
| | SAscore | -0.5 | -0.294 | 0.242 |
| | QED | 0 | -0.050 | 0.09 |
| | SAscore | -1.0 | -0.501 | 0.224 |

**Table 1** Results of three multi-property control scenarios. The requested optimization is the property control condition given to our model. For each scenario, we control two properties simultaneously: Scenario 1 - modulating log$P$ while maintaining molecular weight(MW). Scenario 2 - increasing QED while keeping MW unchanged. Scenario 3 - decreasing SAscore without losing QED. ΔValue is the requested property change in each property. Average and Std. Dev. denote the mean value of property changes and its standard deviation, respectively.

specifying a larger QED increase yielded molecules with higher QED values, deviation from the given condition also became larger. It is not surprising that increasing QED by 0.2 is more challenging since its occurrence within the training dataset is rarer. (see Supplementary Figure 2)

Lastly, we tested our model to improve the synthetic accessibility of a molecule that has a high QED value but low synthetic accessibility. (scenario 3 in Table 1) Here, we adopted the SAscore introduced by Ertl and Schuffenhauer [41], a well-known estimator of the synthetic accessibility. We collected test molecules for this scenario with criteria that QED value larger than 0.7 and SAscore larger than 4. Note that SAscore of a simple catalog molecule is about 2 or 3, whereas for a natural product, it is mostly greater than 5.[41] Property control conditions were set to decrease the SAscores of molecules by 0.5 or 1.0 while retaining their QED values. In this scenario, our model provided novel molecules with lower SAscores, although several of them exhibited slightly decreased QED values. This result suggests that DeepBioisostere can be adopted to supplement many in silico methods for drug design whose outcomes are often questioned in their synthetic accessibility.[42, 43]

## 2.3 Detailed Analysis on Modification with DeepBioisostere

In Section 2.2, we demonstrated that DeepBioisotere can control multiple molecular properties simultaneously while modifying molecular structures. Next, we analyze the modification process in detail in terms of two major steps: the removal fragment selection and the insertion fragment selection. Careful decisions in these steps are essential to control molecular properties, where the overall molecular structure is to be

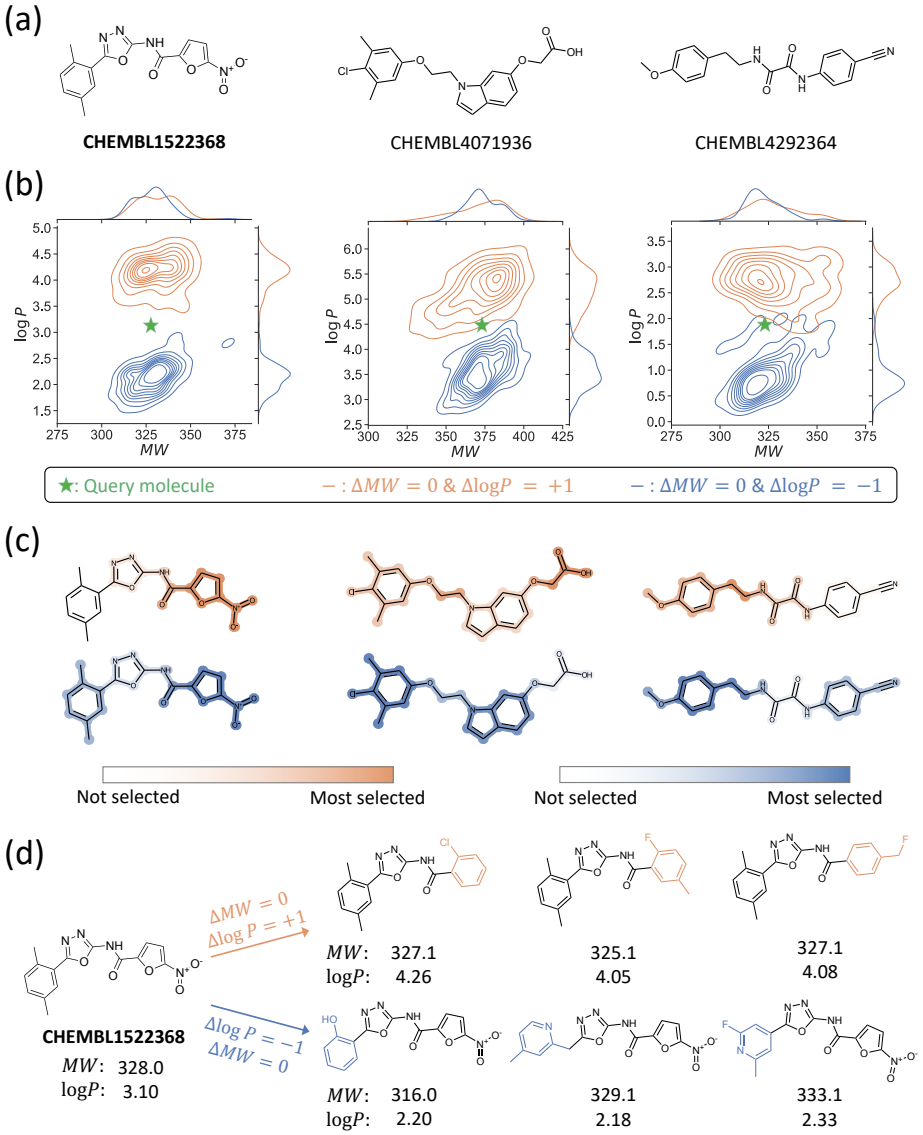


**Fig. 2 Case studies on the multi-property control with DeepBioisostere for scenario 1, described in Table 1.** The orange-colored ones denote results for increasing log$P$ by +1, while the blue-colored ones denote results for decreasing log$P$ by -1. (a) Three original molecules to be modified. (b) The two-dimensional distribution plots of 100 unique molecules modified from each original molecule. (c) Removal fragment selection results. The thicker a fragment is shaded, the more frequently it was chosen as a removal fragment for property control. The upper orange-colored ones visualize the results for increasing log$P$, while the lower blue-colored ones visualize the results for decreasing log$P$. (d) Two series of the top-ranked structures were derived from the first test molecule, CHEMBL1522368

determined. Here, we carry out case studies for multi-property control and will discuss the removal fragment selection and the insertion fragment selection sequentially.

We randomly selected three molecules from the ChEMBL database with moderate molecular weight and lipophilicity, as visualized in Fig. 2(a). We supposed scenario 1 introduced in Section 2.2, where log$P$ is to be modulated while a molecular weight is kept unchanged. We generated 100 unique modified molecular structures for each test molecule and property condition with the DeepBioisostere model. Again, only the fragments from the test dataset for insertion were adopted, the same as in the previous section. We exhibited their log$P$ values and molecular weights in Fig. 2(b). For every test molecule, the generated structures under different property conditions showed distributions that were completely distinct in terms of log$P$ values but very similar in molecular weights. log$P$ was increased or decreased by about 1 while most changes in molecular weights were less than 20, corresponding to the weight of one or two heavy atoms.

### 2.3.1 Removal Fragment Selection

We looked into the generated molecules and analyzed the removal fragment selection step. We counted how many times each substructure was selected as a removal fragment and visualized them in Fig. 2(c). For the molecule in the middle column, CHEMBL4071936, the carboxylic acid group was most frequently selected as a removal fragment to increase log$P$. In contrast, to decrease log$P$, it was hardly substituted, but the chlorodimethyl benzene moiety was preferred as a removal fragment. These choices of removal fragments align well with the known chemical knowledge; introducing a carboxylic acid group enhances hydrophilicity, whereas incorporating an aryl halide moiety increases hydrophobicity. The other two test molecules also showed similar differences in the removal fragment selection under different property control conditions, as exhibited in the 1st and 3rd columns in Fig. 2.

For the first test molecule in Fig. 2(a), CHEMBL1522368, we visualized top-3 ranked structures generated by DeepBioisostere in Fig. 2(d). When required to increase the log$P$ value (the orange arrow), our model substituted a nitrofuran ring with another fragment. All the selected three insertion fragments are aryl groups with halogen elements, which are apparently more hydrophobic than the nitrofuran ring. As a result, the provided molecular structures exhibited log$P$ values higher than the original one (3.10): 4.26, 4.05, and 4.08. Meanwhile, when required to decrease the log$P$ value (the blue arrow), our model selected a 2,5-dimethyl phenyl group to replace with other aryl substituents. The alternative fragments contain a hydroxyl group or a pyridine moiety, resulting in the decreased log$P$ values of the modified structures: 2.20, 2.18, and 2.33. These results suggest that, as discussed in Section 2.1, our model selects insertion fragments by considering the given property control objectives and the removal fragment chosen earlier.

### 2.3.2 Insertion Fragment Selection - Effect of Chemical Environment on Property Control

Another essential factor that governs the selection of insertion fragments is the chemical environment. Here, the chemical environment refers to the substituents surrounding a chosen removal substructure to be alternated by a new insertion fragment. We demonstrate its importance in molecular optimization by two single property control tasks: increasing log$P$ by 1 and increasing QED by 0.2. Obviously, the latter task is more complicated to achieve than the former since QED is calculated based on eight individual molecular features, including log$P$.

Fig. 3 illustrates the overall experimental procedure and its results. First, we selected two molecules that contain a pyrazole group, which is a frequently considered aromatic group in bioisosteric replacement.[11] Their molecular structures are shown in Fig. 3(a), and we will refer to them as molecules A and B, respectively. The two molecules exhibit great differences in log$P$(4.65 and -1.39) while showing moderate levels of QED values(0.58 and 0.60). Notably, the pyrazole moiety in the two molecules is surrounded by totally different chemical environments, particularly in terms of the number of aromatic rings, hydrogen bond donors and acceptors, and rotatable bonds. Here, we forced our model to choose only the pyrazole group as a removal fragment to modulate a given target property, yielding 100 unique modified structures, respectively. (A→A and B→B in Fig. 3(a)) For comparison, we retrieved the insertion fragments selected for molecule A and adopted them into molecule B, resulting in another set of 100 unique molecules. (A→B in Fig. 3(a)) In the first two cases, A→A and B→B, the individual chemical environments directly engaged in the selection of insertion fragments. However, in the case of A→B, the insertion fragments were selected based solely on the chemical environment of molecule A, without any involvement of molecule B. By comparing those cases, we examine the significance of the chemical environment in molecular property control, especially during the insertion fragment selection.

Fig. 3(b) and (c) show the property distributions of the modified molecules, where both cases of A→A and B→B almost achieved the objective values. However, in the case of A→B, the observed property changes in log$P$ and QED conditioning tasks were distinct. In the former task, insertion fragments selected to increase log$P$ of molecule A also increased log$P$ of molecule B, despite the difference between the chemical environments that the pyrazole moiety was involved in. Conversely, in the QED conditioning task, the case of A→B resulted in molecular structures with decreased QED values while those from A→A and B→B exhibited increased QED values compared to the original ones. This apparently shows that the bioisosteric replacements selected to improve the QED value of molecule A failed to improve that of molecule B. We attribute this distinction to the intrinsic difference between the two properties; log$P$ is readily calculated by summing over the estimated contributions of each chemical moiety in a molecule, while QED is calculated with a non-linear function of the molecule-level structural features such as the numbers of hydrogen bond donors and acceptors. This example study suggests that the surrounding parts, in addition to the removal fragment, should be considered when choosing a suitable bioisosteric

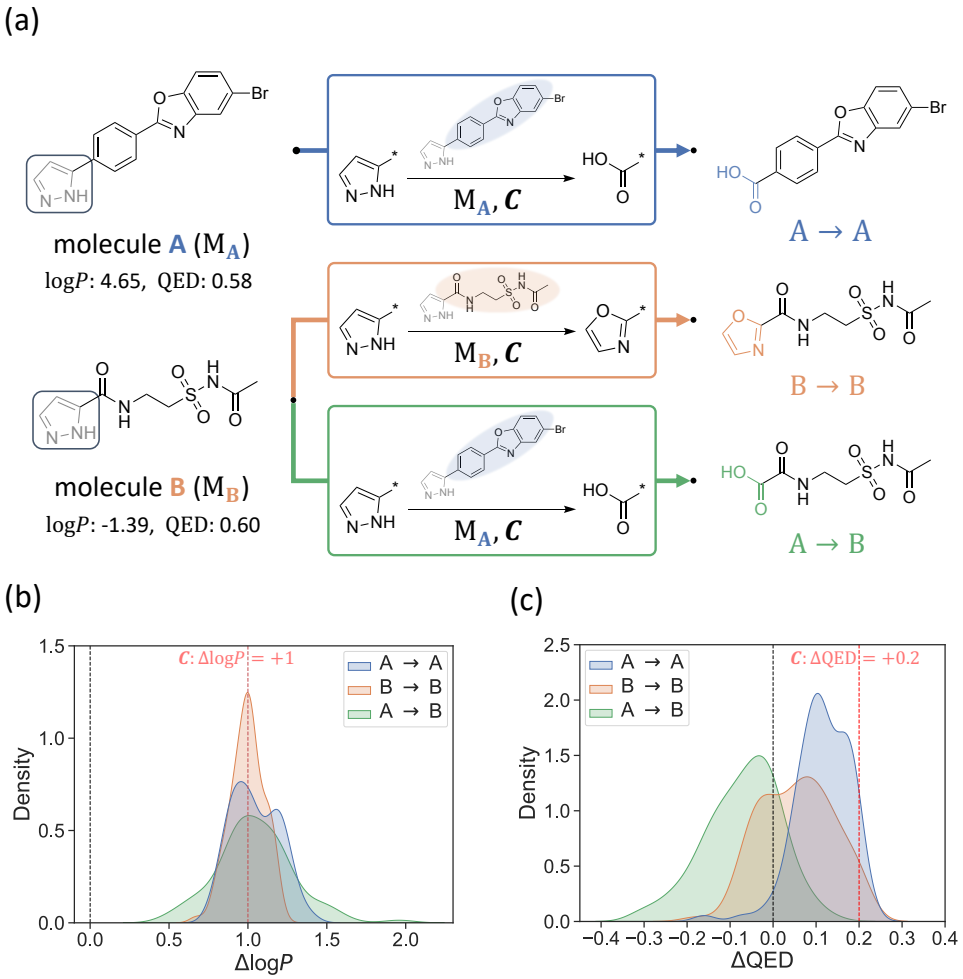


**Fig. 3 An example showing the importance of chemical environment during insertion fragment selection.** (a) The three cases of selecting and applying replacements – within molecule A (A→A), within molecule B (B→B), and selecting for molecule A and applying to molecule B (A→B). (b) $\Delta \log P$ distributions of molecules modified from each original molecule. (c) ΔQED distributions of molecules modified from each original molecule. In both (b) and (c), the black and red dashed lines denote the baselines implying the original molecules' property values and the target property value given as the condition, respectively.

replacement to modulate a target molecular property dependent on the whole molecular structure in a broad sense. It is especially important in drug design, where crucial properties such as synthetic accessibility and ligand binding affinity depend on the non-linear relationships between internal substructures or structural features.

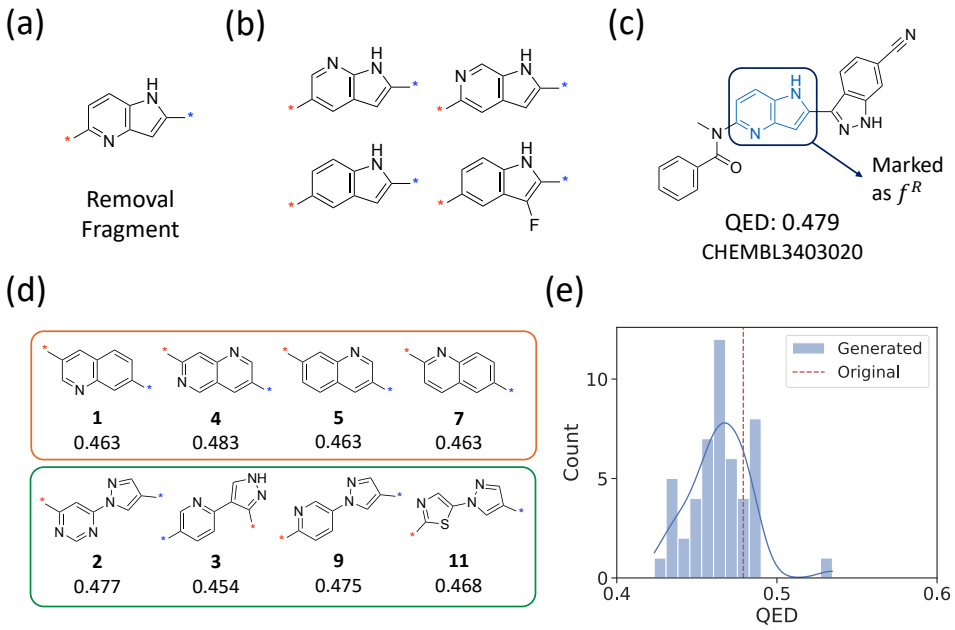


**Fig. 4 Example to find the bioisosteres of an uncommon molecular fragment.** (a) A 4-azaindole moiety, an example of an uncommon moiety. (b) Bioisosteres of the 4-azaindole linker identified from our MMP database. (c) A test molecule from the ChEMBL database, which was modified by our model with its azaindole moiety marked as a removal fragment. (d) Eight top-ranked fragments proposed by our model as bioiosteres. The fragments fall into mainly two types, grouped by different colored boxes for each type. The integer placed below each fragment indicates the order ranked by the DeepBioisostere model, and the decimals are the QED values of the resultant molecules. In (a), (c), and (d), the red and blue stars (*) indicate the corresponding connection sites, respectively. (e) The QED distribution of 50 molecules generated by our model. The red dashed line denotes the QED value of the original molecule.

## 2.4 Finding Novel Bioisosteres for Uncommon Molecular Fragment

Next, we examine whether our model can find novel bioisosterism. Identification of novel bioisosteres becomes necessary when modification of an uncommon molecular fragment is required. In this case, a typical database mining approach is likely to be insufficient since only a few bioisosteres from the database might be available for that fragment. Indeed, among 206,008 fragments in the fragment library we created, 102,467 fragments are found to have no more than ten bioisosteres, respectively. Fig. 4 illustrates such an example of modifying an uncommon fragment. The molecular fragment in Fig. 4(a), a 4-azaindole moiety, appears dozens of times in the ChEMBL database as a linker in the visualized fashion. Utilizing our MMP database, only four fragments were found as bioisosteres, as visualized in Fig. 4(b). All the bioisosteres are fused aromatic heterocycles, the same as the original fragment except for one heavy

atom, which can form similar non-covalent interactions with biological targets. However, by solely depending on these fragments identified from the MMP analysis, one has a limitation in exploring diverse chemical moieties to find potential bioisosteres. More statistics about the number of bioisosteres for each fragment in the fragment library can be found in supplementary figure 3.

To find more bioisosteres beyond them, we employed our model on a molecule from the ChEMBL database by specifying the azaindole moiety as a removal fragment, as in Fig. 4(c). We allowed all insertion fragments in our fragment library regardless of whether they were used in the training process. Here, to explore bioisosteres of the azaindole moiety that exhibits similar drug-likeness, we set the property control objective as $\Delta$QED=0. With this setting, we generated 50 unique molecular structures from the original molecule with DeepBioisostere.

Fig. 4(d) illustrates eight highly ranked insertion fragments proposed by our model. We noticed that they fall into mainly two types: fused aromatic rings (the orange box) and covalently bonded aromatic rings (the green box). The fragments in the orange box consist of fused aromatic heterocycles, each containing one or two nitrogen atoms, similar to those identified from our MMP database shown in Fig. 4(b). On the other hand, the fragments in the green box have substantially different molecular structures, where each of them contains covalently bonded aromatic rings. Still, all the highly ranked fragments are inserted into the test molecule to exhibit QED values similar to that of the original molecule. (see Fig. 4(d)) Indeed, they contain a few hydrogen bond donors and acceptors and two aromatic rings, the same as the original fragment, and they have a comparable level of lipophilicity. These structural features allow us to expect that they possess the potential to form similar non-covalent interactions as the original one. Fig. 4(e) illustrates the QED distribution of 50 unique modified structures. The peak near the original QED value suggests that most of the 50 insertion fragments are similar to the azaindole moiety in terms of diverse structural features crucial for bio-activity such as the number of hydrogen bond donors and acceptors, aromatic rings, and polar surface area. This example study demonstrates that one can adopt DeepBioisostere for an uncommon chemical moiety to explore potential bioisosteres with a similar level of drug-likeness or QED.

## 2.5 Targeting SARS-CoV-2 Main Protease Mutant

Lastly, we demonstrate DeepBioisostere in a more practical example, where modifying a known drug to inhibit mutants of the original biological target. The ongoing coronavirus 2019 (COVID-19) pandemic still threatens public health and safety. Due to the limited therapeutic options, a few oral COVID-19 therapeutics have been recently developed, especially for non-hospitalized patients.[44] Ensitrelvir is the first oral noncovalent and non-peptidic SARS-CoV-2 3CL protease ($3CL^{pro}$) inhibitor that received emergency regulatory approval from the Ministry of Health, Labour and Welfare in Japan.[45, 46] Ensitrelvir shows remarkable biochemical activity and antiviral activity against several mainstream SARS-CoV-2 variants,[45] but recent studies have reported that several naturally occurring mutations of $3CL^{pro}$ show drug resistance to Ensitrelvir and other oral antiviral drugs.[47–49] Since these mutations still hold

the possibility to undergo changes that show more increased drug resistance, the development of specific therapeutics for them is potentially needed.

We selected the E166V mutant as a target here since it is associated with about 10-fold reduced Ensitrelvir inhibitory activity relative to the wild type. [50] We retrieved and exhibited the complex structures of the wild type(PDB ID: 8HBK) and the E166V mutant(PDB ID: 8H3G) with Ensitrelvir in Fig. 5(a), reported by Duan et al. [50] We analyzed their structures and noticed that they were not much different in their binding pockets. Based on this observation, we hypothesized that a more potent inhibitor on the mutant could be derived from a modification of Ensitrelvir rather than discovering a new lead series.

In the binding structure of the wild type and Ensitrelvir, the E166 residue is known to form multiple H-bondings with another region of $3CL^{pro}$ protein, contributing to the stability of the overall binding structure. Thus, the change of the E166 residue to Valine (E166V) must increase the hydrophobicity nearby and interrupt the H-bonding network, resulting in the decreased binding stability of Ensitrelvir or increased drug resistance. [50] In this regard, we hypothesized that modifying the methyl-triazole moiety, circled in gray at Fig. 5(b), of Ensitrelvir to a hydrophobic one would increase sensitivity to the E166V mutant. A large change in molecular size must be avoided during modification since the pocket volume of the E166V mutant is not adequate to contain bulkier fragments. (see Fig. 5(b)) Since Ensitrelvir has been synthesized and marketed as an emergency drug, we expect the synthetic accessibility and drug-likeness to be maintained during bioisosteric replacement. Hence, we set a property control condition as increasing log$P$ by 1 while sparing the molecular weight, QED, and SAscore to identify an appropriate bioisostere of the methyl-triazole moiety with increased hydrophobicity.

We designated the methyl-triazole as a removal fragment to be modified. Adopting the entire fragment library, we obtained 500 modified molecular structures with DeepBioisostere. We scored their binding affinities to the E166V mutant with PIGNet2, a physics-informed deep learning model for predicting protein-ligand binding affinities that exhibits promising performance in comparing binding affinities of structural analogs to the same biological target.[53] We generated conformers of the generated molecules by fixing the conformations of the remaining part, followed by local structural optimization with SMINA,[54] a docking program based on AutoDock Vina.[55] (See supplementary information for details about the overall experimental process.)

The PIGNet2 score distribution of the 500 generated molecules is shown in Fig. 5(c). The average and median scores of all generated molecules were -7.80 kcal/mol and -7.68 kcal/mol, respectively, which are better than the score of the reference ligand, -7.05 kcal/mol. Also, 25.8% of the generated molecules were expected to show 10 times better binding affinity against the E166V mutant than Ensitrelvir in terms of half maximal inhibitory concentration, regarding the difference of 1.36 kcal/mol in energy scale as the 10-fold difference in the inhibitory concentration. We visualized 3D interaction patterns of Ensitrelvir in the binding structure in Fig. 5(d) and that of one generated molecule with a highly ranked affinity score in Fig. 5(e). Notably, the alternative cyclohexanol fragment possesses a broader hydrophobic surface so as to

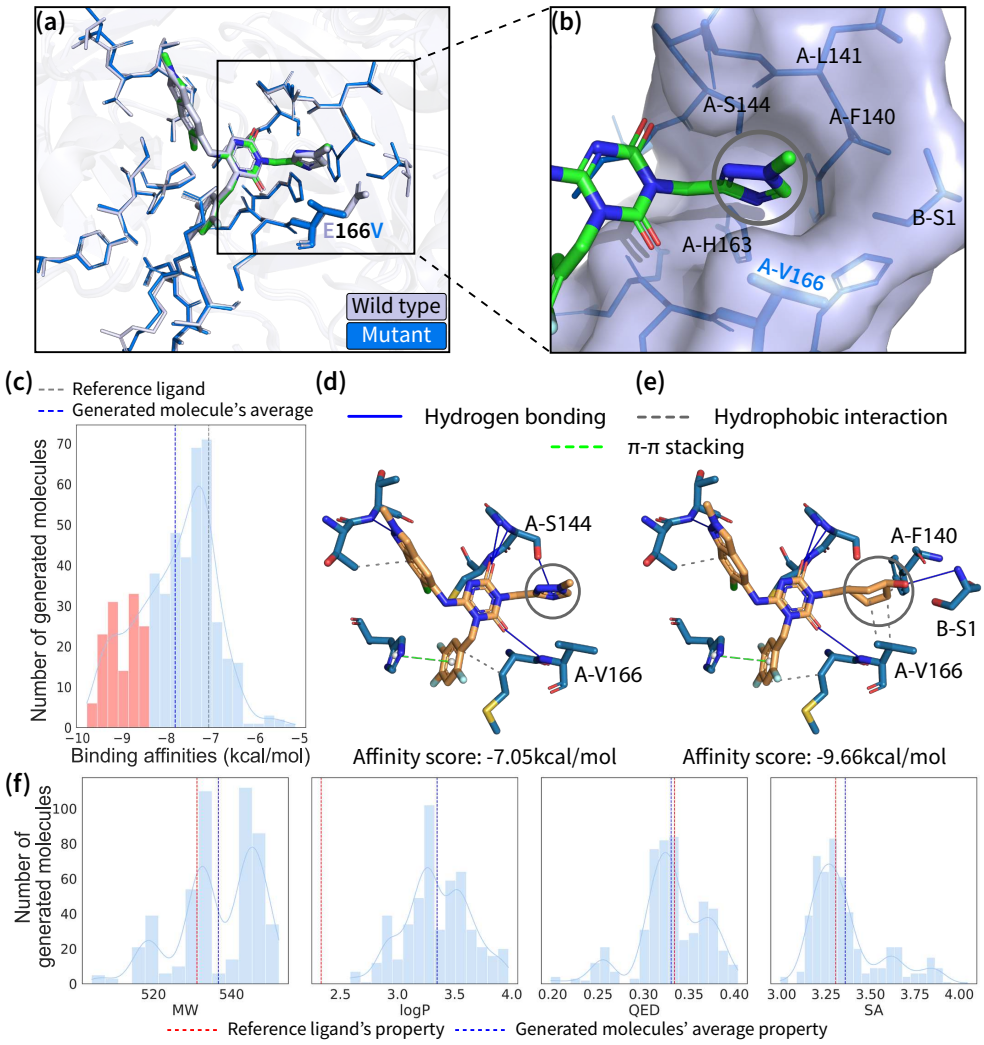


**Fig. 5 Case study on the inhibitor design of 3CL$^{pro}$ E166V mutant with DeepBioisostere.** (a) An overlay of the aligned pocket-ligand binding structures of the wild type and the mutant. The wild-type pocket and the corresponding ligand structure are colored white, while the mutant pocket is colored blue. The ligand in the mutant is represented with various colors. (b) A focused view of the mutated residue V166 and its nearby. The gray circle highlights the methyltriazole moiety to be modified. (c) A binding affinity histogram of generated molecules for the 3CL$^{pro}$ E166V mutant for statistical analysis. Here, the red boxes imply molecules with over 10-fold better binding affinities than that of the reference ligand. (d), (e) 2D diagrams illustrating the intermolecular interactions of Ensitrelvir (d) and a molecule generated by DeepBioisostere (e) with their estimated affinity scores with the E166V mutant. The gray-circled moieties are removal and insertion fragments, respectively. Both 2D diagrams of protein-ligand interactions are drawn with a protein-ligand interaction profiler(PLIP).[51, 52] (f) Property distributions of molecules generated DeepBioisostere.

establish extensive hydrophobic interactions with the Val166 residue, while its terminal hydroxyl group also forms an additional hydrogen bond with the Ser1 in another chain. This resulted in a better affinity score against the E166V mutant compared to Ensitrelvir, as intended. Similarly, many molecules with binding affinities better than -9 kcal/mol exhibit extensive hydrophobic interactions with Val166 in addition to an extra hydrogen bond with the Ser1 or Ser144.(see Supplementary Figure 6) Remarkably, along with the enhanced binding affinity scores, the four target properties of the generated molecules consistently satisfy our property control objectives as displayed in Fig. 5(f). This highlights the potential of DeepBioisostere on lead transformation, utilizing bioisosteric replacements to discover better binders while preserving crucial properties in drug developments.

# 3 Discussion

In this work, we propose a deep generative model named DeepBioisostere for designing appropriate bioisosteric replacements to control multiple molecular properties. Our model proceeds a chemical modification in an end-to-end manner, choosing a fragment to remove or insert along with its attachment orientation. We demonstrate diverse scenarios of dual property control, where one property is changed while the other is maintained. DeepBioisostere, in every scenario, successfully modulated their properties in the desired manner by replacing fragments from the given molecules. We further show the adjustability of DeepBioisostere; it dynamically selects which fragment to remove depending on the given property control condition. It is practically advantageous where previous database mining approaches heavily rely on the help of experts for choosing a substructure to be modified. Moreover, we justify the significance of substituents near the replacement site to control comprehensive molecular properties such as drug-likeness, especially for the insertion fragment selection. Our model can design novel bioisosteric replacements that are not included in the training data via flexible selections of fragment pairs. We emphasize that it is particularly important when the removal fragment is uncommon, providing a wide variety of potential chemical modifications. Finally, we applied DeepBioisostere to a practical example of encountering drug-resistant mutation of SARS-CoV-2 main protease. We could obtain potentially improved molecules that strongly interact with the mutant residue while retaining drug-likeness, synthetic accessibility, and size, showcasing the strength of our model.

# 4 Methods

## 4.1 Related works

One of the representative works in database mining approaches is SwissBioisostere.[14, 56] Its bioisostere database is constructed through the matched molecular pair analysis (MMPA) of the ChEMBL database, which conducts the statistical analysis to count the occurrences of each bioisosteric replacement and the resultant changes in biochemical properties, such as log$P$, topological polar surface area (TPSA), and bioactivity in biochemical assays. SwissBioisostere recommends bioisosteric replacements for a

queried molecular fragment based on their frequency in the parent database and expected shifts in the desired properties. The database is easily accessible on the website.

Deep generative modeling has been adopted in molecule optimization, mainly focusing on multiple-property control—the essential goal of the lead optimization process. One major category utilizes a language representation of a molecular structure(SMILES),[57] formulating molecular optimization as a language translation problem.[30–32] He et al. [30] have utilized natural language as a condition with two machine translation models to optimize absorption, distribution, metabolism, excretion, and toxicity(ADMET). In parallel, Yang et al. [32] have used a constrained transformer along with the state-of-the-art ADMET prediction model, demonstrating successful modifications achieving desired property enhancements. Although they showed attractive performance in various multi-property control tasks, the overall structure of an original molecule is not always retained in the resultant structures, which is not desirable in a typical lead optimization process. Also, since the language representation of a molecular structure assumes a particular order of atoms in a molecule, it is difficult to assess the compatibility between a newly introduced fragment and its surroundings, which is essential to handle complicated molecular properties such as drug-likeness and synthesizability.

An important reference to this work is Chen et al. [35], where molecular optimization is redefined as a graph-editing task via single bond disconnection. Their generative model, Modifier-with-one-fragment(Modof), performs an R-group replacement to achieve a desired property by selecting a bond to be cleaved from an original molecule, detaching a single fragment, and attaching another fragment. Since the Modof model sequentially generates the structural change and applies it to the original structure, it can modify a molecular structure while not changing the residual substructure.

In this work, we attempt to model the overall process of chemical modification through bioisosteric replacements with deep generative modeling, contributing to the autonomous optimization of molecular structures. While database mining approaches, including the SwissBioisostere database, provide bioisosteres for a molecular fragment selected by a medicinal chemist, DeepBioisostere selects a removal fragment from a given molecule and replaces it with a proper alternative automatically (see Fig. 1). Through the fragment-level modification, DeepBioisostere can replace not only terminal R-groups but also internal substructures of target molecules with their bioisosteres, while the remaining structure is unchanged.

## 4.2 Finding molecular pairs with bioisosterism: dataset construction

In this section, we present our scheme for constructing a training dataset. It is a pivotal step in our work, where the dataset should be precisely tailored to align with the objective of DeepBioisostere. To obtain a dataset with potential bioisosteric replacements, we begin from the ChEMBL database[27] and go through two steps of training dataset construction: pre-filtration and MMP identification.

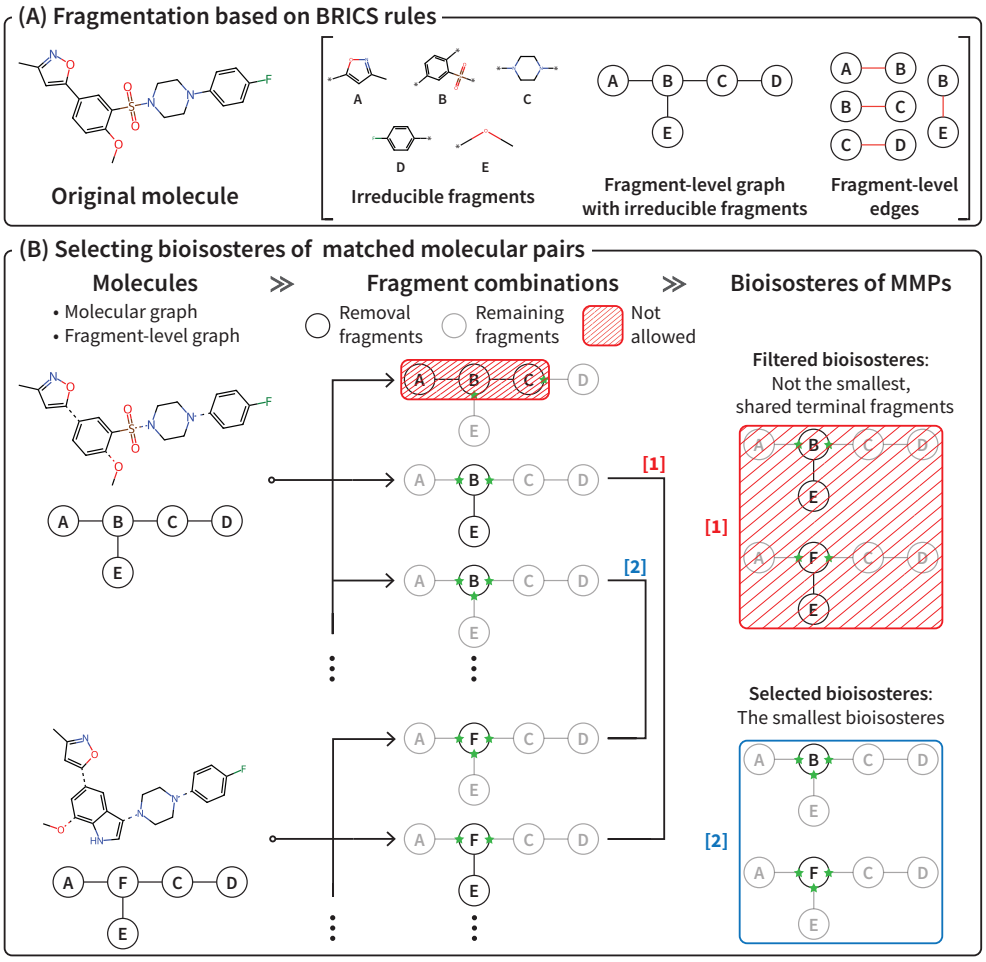


**Fig. 6 The overall dataset generation scheme.** (A) The original molecule is fragmented according to the BRICS rules to derive irreducible fragments, resulting in a fragment-level graph with the corresponding edges, where each edge corresponds to a single BRICS bond. (B) The dataset is constructed by identifying bioisosteres. First, allowed fragment combinations are extracted for each molecule. Then, the smallest bioisosteres are selected for each matched pair of molecules that share the remaining fragments. The dataset incorporates all selected bioisosteres, along with the original molecule, the modified molecule, and the bioisostere pair representing the removal and insertion fragments.

### 4.2.1 Pre-filtration

We adopted a permissive pre-filtration strategy to maximally retain the data. Simply, we excluded heavy molecules with molecular weights exceeding 800 Da and molecules consisting of salts. Additionally, molecules with an inhibitory concentration greater than 10 $\mu$M are also excluded so that the remaining molecules can be regarded as possessing sufficient bioactivities. This pre-filtration process provides a refined dataset

appropriate to the next stage of MMP identification. From the total of 3,849,629 ChEMBL molecules, we were able to obtain 964,999 molecules by the pre-filtration.

### 4.2.2 Identifying MMPs based on the BRICS rules

MMP is a pair of molecules in which all parts are identical except for a small fragment in each molecule. From the pre-filtered ChEMBL database, we identified MMPs by fragmenting each molecule and then selecting molecule pairs that share common fragments. It is crucial to adopt a proper fragmentation rule in this process since it determines the fragments corresponding to the difference between each MMP, which are to be considered potential bioisosteres. To consider more synthetically accessible modifications than the conventional strategy of simply cleaving rotatable bonds in the fragmentation process, we employed BRICS[37] rules for molecular fragmentation, implemented in the RDKit[58] package.

Using the BRICS rules, the ChEMBL molecules are represented as fragment-level graphs (see Fig. 6(A)). Each fragment-level node is an irreducible fragment, a molecular subgraph that cannot be further divided using the fragmentation rules. Edges between nodes correspond to BRICS bonds. Then, we define a removal fragment as a connected subgraph of a fragment-level molecular graph. With these definitions, we enumerate fragment combinations that involve a single removal fragment and remaining fragments for all BRICS-fragmented ChEMBL molecules. Then, the fragment combinations are further filtered with two criteria regarding their removal fragments: (1) the removal fragment contains no more than 12 heavy atoms and (2) the removal fragment should have fewer heavy atoms than those present in the remaining fragments.

The aforementioned process results in an enormous number of fragment combinations derived from diverse molecules. The obtained fragment combinations are subsequently compared, where any molecular pairs with the same remaining fragments are classified as MMPs along with their removal fragments. The majority of the resultant MMPs contain more than two matched fragment combinations, respectively. In these cases, the smallest removal fragments are exclusively selected as bioisosteres for each MMP, preventing the accumulation of redundant data in the training data (see Fig. 6(B)). The final training dataset consists of a total of 24,163,198 MMPs designated about their corresponding bioisosteres. Note that their bioisosteres may not necessarily be irreducible fragments since, by its definition, a bioisostere does not need to be a BRICS-irreducible fragment. Thus, a generative model trained with these data would modify a molecular structure at the bioisostere level rather than at the BRICS-irreducible fragment level.

Based on the final dataset, we collect the insertion fragments to build a fragment library. The fragment library is necessary for both the training and inference phases of DeepBioisostere as a source of insertion fragment candidates, which our model utilizes to optimize a given molecule. For more details on the use of the fragment library of DeepBioisostere, refer to the Sections 4.4 and 4.5. Note that gathering all removal fragments will result in the same fragment library since each identified MMP results in two training data entries with reverse relationships in their removal and insertion fragments. The resulting fragment library is composed of 206,008 diverse

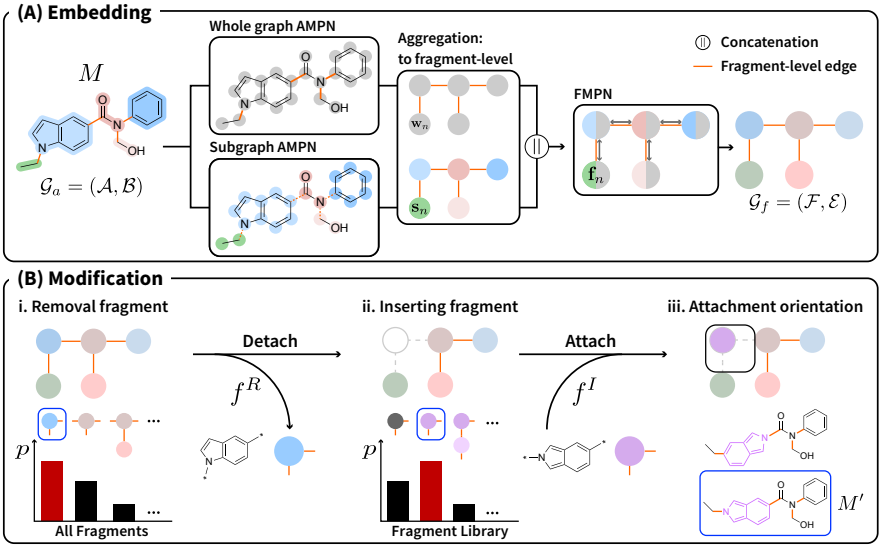


**Fig. 7 Operation scheme of DeepBioisostere.** (A) Fragment-level feature embedding process. The original molecule, $M$, initially propagates through the whole graph and subgraph AMPN(atom message passing network). Subsequently, the atom-level embedding vectors are aggregated into a single fragment-level embedding vector, followed by an FMPN(fragment message passing network). (B) Based on the fragment-level vectors, DeepBioisostere sequentially selects the removal fragment, the insertion fragment, and the attachment orientation of the insertion fragment. The removal fragment is selected from all possible fragments of $M$, and the insertion fragment is chosen from our fragment library.

fragments. Furthermore, we randomly split the fragment library at a ratio of 8:1:1, leading to 164,813, 20,604, and 20,591 fragments for training, validation, and test datasets, respectively.

## 4.3 DeepBioisostere: Deep generative modeling of bioisosterism

Chemical modification aims to refine a molecular structure to achieve desirable chemical properties in respective drug developments. In this work, we define the chemical modification task as learning the likelihood of a modified molecular structure, $M'$, for a given original molecule, $M$, and property control condition, $\boldsymbol{C}$ — $p\,(M'|M, \boldsymbol{C})$. We devise a generative model, named DeepBioisostere, to learn the conditional probability of $p\,(M'|M, \boldsymbol{C})$. The DeepBioisostere model optimizes a molecule through three essential steps involving the selection of (1) the removal fragment, (2) the insertion fragment, and (3) the attachment orientation. This sequential modeling was inspired by the traditional chemical modification process based on bioisosterism.

The probability distribution of modified molecules depends on the fragmentation rule, $\mathbf{R}$, which defines what bond in a starting molecule can be cleaved to enhance the synthetic accessibility of the final molecules. For example, if we do not allow the cleavage of the covalent bond connecting aromatic rings, we cannot obtain phenyl aniline moiety by modifying a biphenyl compound. Thus, we denote the objective

conditional probability of our model as $p_{\mathbf{R}}\left(M'|M, \boldsymbol{C}\right)$, reflecting the dependence of the chemical modification task on the bond cleavage rule, $\mathbf{R}$. For an MMP according to $\mathbf{R}$, $(M,\ M')$, their conditional probability can be reformulated by factorization with the conditional probabilities of corresponding three modification components as follows:

$$\begin{aligned} p_{\mathbf{R}}\left(M'|M, \boldsymbol{C}\right) &= p_{\mathbf{R}}\left(f^R, f^I, A\middle|M, \boldsymbol{C}\right) \\ &= \underbrace{p_{\mathbf{R}}\left(f^R\middle|M, \boldsymbol{C}\right)}_{(1)} \cdot \underbrace{p_{\mathbf{R}}\left(f^I\middle|M, f^R, \boldsymbol{C}\right)}_{(2)} \cdot \underbrace{p_{\mathbf{R}}\left(A\middle|M, f^R, f^I, \boldsymbol{C}\right)}_{(3)}, \end{aligned} \tag{1}$$

where $f^R$ is a removal fragment, $f^I$ is an insertion fragment, and $A$ is an attachment orientation. Note that for a modification from $M$ to $M'$, there might be more than one set of the three modification components, but we uniquely defined $f^R$ and $f^I$ to be only the smallest removal and insertion fragments in the training data construction procedure, respectively. In this way, the ground truth for $p_{\mathbf{R}}\left(f^R, f^I, A\middle|M, \boldsymbol{C}\right)$ is defined to be zero if $f^R$ and $f^I$ are not the smallest ones to modify $M$ to another molecular structure, $M'$. Since the smallest structural change can always be identified from any MMP, this choice does not impose any restriction on the probability distribution of $p_{\mathbf{R}}\left(M'|M, \boldsymbol{C}\right)$. The details about how we model each conditional probability of the Eq. (1) are discussed in the method section.

### 4.3.1 Fragment-level molecular encoding

The overall framework of DeepBioisostere is depicted in Fig. 7. First, we encode an original molecule to be modified into fragment-level embedding vectors. We adopt a hierarchical architecture to facilitate probability modelings on fragment-level information afterward. An original molecule is represented in two ways: an atom-level graph, $G_a$, and a fragment-level graph, $G_f$. Largely inspired by Chen et al. [35], we employ two message-passing neural networks in different levels named an atom message passing network(AMPN) and a fragment message passing network(FMPN), respectively. An original molecule is first encoded into atom-level embedding vectors by AMPN. As illustrated in Fig. 7(A), AMPN runs within the whole graph and subgraphs of a molecule, which adopts respective atom-level embedding vector update mechanisms and uses differently initialized weights. An initial atom feature vector, $\hat{\mathbf{a}}$, is concatenated with the condition vector, $\boldsymbol{C}$, to yield an input atomic embedding vector by $\mathbf{a}_i^0 = [\hat{\mathbf{a}}_i \| \boldsymbol{C}_i]$. Details about constructing the initial atom feature vectors are summarized in the supplementary information. We summarized key notations for model architecture in Table 2. Embedding vectors for both atoms and bonds are propagated through the AMPN layers as follows:

$$\mathbf{m}_i^t = \sum_{a_j \in \mathcal{N}(a_i)} W_1^a \mathbf{a}_i^t + W_2^a \mathbf{a}_j^t + W_3^a \mathbf{b}_{ij}, \tag{2}$$

$$\mathbf{a}_i^{t+1} = \mathrm{GRU}(\mathbf{a}_i^t, \mathbf{m}_i^t), \tag{3}$$

where $\mathcal{N}(a_i)$ is the neighboring atoms of $a_i$, GRU is the gated recurrent unit introduced in Chung et al. [59] The whole graph AMPN operates on the entire molecular graph $\mathcal{G}_a$, updating atom-level embeddings based on the collective information of all atoms and bonds in the molecule. In contrast, the subgraph AMPN works locally within each irreducible fragment, focusing its update on atoms only within the specific fragment, denoted as $\mathcal{A}(\hat{f})$. This subgraph-based update strategy allows the model to learn the comprehensive features of each irreducible fragment for further hierarchical message passing.

| **Notation** | **Description** |
|---|---|
| $\mathcal{G}_a = (\mathcal{A}, \mathcal{B})$ | Atom-level molecular graph |
| $\mathcal{A}$, $\mathcal{B}$ | Set of atoms and bonds in $\mathcal{G}_a$, respectively |
| $a \in \mathcal{A}$ | An atom in $\mathcal{G}_a$ |
| $b_{ij} = \{a_i, a_j\} \in \mathcal{B}$ | A bond connecting atoms $a_i$ and $a_j$ in $\mathcal{G}_a$ |
| $\mathbf{a}_i^t$ | An embedding vector of $a_i$ in the $t^{\text{th}}$ layer |
| $\mathbf{b}_{ij}$ | An embedding vector of $b_{ij}$ |
| $\mathcal{G}_f = (\mathcal{F}, \mathcal{E})$ | Fragment-level molecular graph |
| $\mathcal{F}$, $\mathcal{E}$ | Set of irreducible fragments and edges in $\mathcal{G}_f$, respectively |
| $f$ | A subgraph, fragmented by the BRICS rules |
| $\hat{f}$ | An irreducible fragment |
| $f^R$ | A *removal* fragment |
| $f^I$ | An *insertion* fragment |
| $e_{nm} \in \mathcal{E}$ | An edge between irreducible fragments $\hat{f}_n$ and $\hat{f}_m$ in $\mathcal{G}_f$, corresponding to a BRICS bond in $\mathcal{G}_a$ |
| $\mathcal{A}(f)$ | A set of atoms included in a subgraph $f$ |
| $\mathcal{A}(e_{nm})$ | A set of atoms composing $e_{nm}$ |
| $\mathcal{F}(f)$ | A set of irreducible fragments included in a subgraph $f$ |
| $\mathbf{f}_n^t$ | An embedding vector of $\hat{f}_n$ in the $t^{\text{th}}$ layer |
| $\mathbf{e}_{nm}$ | An embedding vector of $e_{nm}$ |
| $\mathbf{h}_f$ | An aggregated vector for $f$ |
| $A$ | An attachment orientation, a set of edges enumerated based on the BRICS rules |
| $[\cdot\|\|\cdot]$ | Element-wise concatenation |
| $\sigma$ | Sigmoid activation function |
| $W$ | Learnable weights |

**Table 2** Notations and the corresponding descriptions used in the formulation of DeepBioisostere

The updated atomic embedding vectors, $\{\mathbf{a}_i\}$, are aggregated into fragment-level embedding vectors for each irreducible fragment as follows:

$$\mathbf{w}_n = \sum_{a_i \in \mathcal{A}(\hat{f}_n)} \mathbf{a}_i, \tag{4}$$

$$\mathbf{e}_{nm} = \sum_{a_i \in \mathcal{A}(e_{nm})} \mathbf{a}_i, \tag{5}$$

where $\mathbf{w}_n$ denotes a fragment-level embedding vector aggregated from the whole graph AMPN embeddings. Atom-level embeddings from the subgraph AMPN are aggregated in the same way, which is denoted as $\mathbf{s}_n$. Given that the subgraph AMPN updates the vectors without considering the information of adjacent atoms connected by the BRICS bonds, we exclusively employ the whole graph AMPN to aggregate the edge vectors of fragment-level graphs. We concatenate the two aggregated fragment-level embedding vectors for further propagation in FMPN, which incorporates the feature vector of each fragment and intrinsically works like AMPN:

$$\mathbf{m}_n^t = \sum_{\hat{f}_m \in \mathcal{N}(\hat{f}_n)} W_1^f \mathbf{f}_n^t + W_2^f \mathbf{f}_m^t + W_3^f \mathbf{e}_{nm}, \tag{6}$$

$$\mathbf{f}_n^{t+1} = \mathrm{GRU}(\mathbf{f}_n^t, \mathbf{m}_n^t). \tag{7}$$

Here, $\mathbf{f}_n^0 = [\mathbf{w}_n \| \mathbf{s}_n]$, is an initial feature vector of an irreducible fragment. The resulting fragment-level embedding vectors, $\{\mathbf{f}_n\}$, are utilized for the modification of $M$. Among the overall embedding process, the subgraph AMPN is significant since it allows DeepBioisostere to generate embedding vectors for previously unseen fragments.

With the fragment-level embedding vectors obtained by AMPN and FMPN, we model each conditional probability of Eq. (1). During the substitution, both $f^R$ and $f^I$ can be a combination of multiple irreducible fragments. Thus, we obtain the final embedding vector for each individual fragment through the aggregation of the involved irreducible fragments as follows:

$$\mathbf{h}_f = \sum_{\hat{f}_n \in \mathcal{F}(f)} \mathbf{f}_n. \tag{8}$$

We note that $\mathbf{h}_f$ contains comprehensive information about the chemical structure of $f$, its relationships with surrounding fragments, and the given property control condition.

### 4.3.2 Modeling sequential modification steps

With the aggregated embedding vectors, the DeepBioisostere model sequentially modifies the original molecule, $M$ (see Fig. 7(B)). First, the removal fragment selection module estimates the probability of removing each fragment. This exclusively considers the allowed removal fragments, satisfying the criteria used during the dataset construction, as detailed in Section 4.2. For a potential removal fragment $f^R$, the likelihood is formulated as follows:

$$p\left(f^R \middle| M, \boldsymbol{C}\right) \propto \sigma\big(\phi\left(\mathbf{h}_{f^R}; \theta_R\right)\big), \tag{9}$$

where $\phi(\cdot; \theta_R)$ is a removal fragment selection module parameterized with $\theta_R$. After the removal fragment is determined, our insertion fragment selection module chooses an appropriate fragment from the predefined library established during the dataset

generation. The following likelihood guides this selection:

$$p\left(f^I \middle| M, f^R, \boldsymbol{C}\right) \propto \sigma\left(\phi\left(\left[\mathbf{h}_{f^R} \middle\| \mathbf{h}_{f^I}\right]; \theta_I\right)\right). \tag{10}$$

Similarly, $\phi(\cdot;\theta_I)$ is an insertion fragment selection module parameterized with $\theta_I$. Thus, the insertion fragment depends on the pre-selected removal fragment, its surroundings, and the given property control condition, which are included in $\mathbf{h}_{f^R}$.

For an insertion fragment with two or more attachment sites, more than one attachment orientations might be allowed. For example, isoindole in the left side of Fig. 7(B) can be attached to the remaining substructure in two different orientations. In these cases, a proper attachment orientation should be determined to insert the selected fragment for desired property control. We enumerate all attachment orientations that are allowed by the BRICS rules, and then the last module estimates their likelihood:

$$p\left(A \middle| M, f^R, f^I, \boldsymbol{C}\right) \propto \prod_{e_{nm} \in A} \sigma\big(\phi\left(\mathbf{e}_{nm}; \theta_A\right)\big), \tag{11}$$

where $A$ is an attachment orientation, and $\phi\left(\cdot;\theta_A\right)$ is an attachment orientation selection module parameterized with $\theta_A$. Eq. (11) indicates that each attachment orientation is evaluated based on the respective attaching sites (or edges) under the assumption that the compatibility of joining two fragments mainly depends on the local relationship.

## 4.4 Training DeepBioisostere

As explained in section 4.3.2, DeepBioisostere modifies a given molecule with the three main modules. To learn the conditional distribution $p\left(M'|M, \boldsymbol{C}\right)$ with them, we devised three loss functions for each module, $\mathcal{L}_{\text{remove}}$, $\mathcal{L}_{\text{insert}}$, and $\mathcal{L}_{\text{attach}}$, respectively. Then, we trained our model with their summation as the final training loss function.

The dataset constructed in Section 4.2 includes only positive data of the ground-truth removal fragment, insertion fragment, and attachment orientation. To make the removal fragment selection module likely select the positive removal fragment over others, we regard other allowed fragments of the original molecule as negative samples for each pair of bioisosteres in the dataset. We employed the binary cross-entropy loss to formulate $\mathcal{L}_{\text{remove}}$ as follows:

$$-\mathcal{L}_{\text{remove}} = \log p^R_{pos} + \frac{1}{N_R} \sum_{f^R_{neg}}^{N_R} \log\left(1 - p^R_{neg}\right), \tag{12}$$

where $p^R_{pos} = \sigma(\phi(\mathbf{h}_{f^R_{pos}}; \theta_R))$, $p^R_{neg} = \sigma(\phi(\mathbf{h}_{f^R_{neg}}; \theta_R))$, $f^R_{neg}$ is an allowed but negative fragment from the original molecule, and $N_R$ is the total number of allowed but negative fragments from the original molecule.

The selection of insertion fragments is more complicated since our model should be able to retrieve the ground truth among a vast number of potential candidates from the fragment library. To avoid the inefficiency in evaluating tens of thousands of fragments

for every data point, we employed the negative sampling strategy inspired by Seo et al. [60] This strategy samples only a few fragments (20 fragments in this work) with the probability weakly proportional to their occurrence as negative samples. With this negative sampling strategy, we formulated $\mathcal{L}_{\text{insert}}$ as follows:

$$-\mathcal{L}_{\text{insert}} = \log p^I_{pos} + \frac{1}{N_I} \sum_{f^I_{neg}}^{N_I} \log\left(1 - p^I_{neg}\right), \tag{13}$$

where $p^I_{pos} = \sigma(\phi([\mathbf{h}_R \| \mathbf{h}_{f^I_{pos}}]; \theta_I))$, $p^I_{neg} = \sigma(\phi([\mathbf{h}_R \| \mathbf{h}_{f^I_{neg}}]; \theta_I))$, $N_I$ is the number of negative samples for each positive data. By including negative samples in Eq. (13), we counteracted the bias towards always preferring more frequently occurring fragments over less frequent ones.

According to Eq. (11), we estimate the probability of each attachment orientation as the product of the likelihood of the potential edges to be formed from the attachment. Hence, we devised an edge-based objective function to train the attachment orientation selection module. For each positive data, we enumerate all the potential edges allowed by the BRICS rules and classify them into positive and negative ones according to whether they are included in the ground-truth attachment orientation. Then, we again set the binary cross-entropy loss in this task:

$$\begin{aligned} -\mathcal{L}_{\text{attach}} = & \frac{1}{N_{pos}} \sum_{e^{pos}_{nm}}^{N_{pos}} \log p^A_{pos} \\ & + \frac{1}{N_{neg}} \sum_{e^{neg}_{nm}}^{N_{neg}} \log\left(1 - p^A_{neg}\right), \end{aligned} \tag{14}$$

where $p^A_{pos} = \sigma(\phi(\mathbf{e}^{pos}_{nm}; \theta_A))$ and $p^A_{neg} = \sigma(\phi(\mathbf{e}^{neg}_{nm}; \theta_A))$. $N_{pos}$ and $N_{neg}$ are the number of potential edges within the positive and negative orientations, respectively.

For training and validation, we split the whole dataset constructed in Section 4.2 into training, validation, and test sets with a ratio of 8:1:1. The dataset was split according to their insertion fragments so that the three sub-sets share no insertion fragment. Thus, each dataset contains distinct bioisosteric replacements. For the multi-property control, we used a total of four properties that can be readily computed based on the structure of a molecule: molecular weight, log$P$, quantitative estimate of drug-likeness(QED),[40] and SAscore.[41] The training was finished in no more than 17 hours with a single NVIDIA RTX A4000 GPU for each property control condition. Details about the training process can be found in the supplementary information.

## 4.5 Molecular Optimization Process with DeepBioisostere

DeepBioisostere performs optimization with two inputs: a molecule to be improved and a property control condition. The molecule is first analyzed with the BRICS rules to obtain its fragment-level graph, and all its substructures are enumerated. Then, we filter the substructures based on the criteria used to identify MMPs in Section 4.2.2,

allowing the modification to occur in a small part of the molecule. Thus, a substructure that contains more than 12 heavy atoms or more heavy atoms than the other parts of the molecule is not considered in the removal fragment selection. All the allowed fragments are examined by the removal fragment selection module, yielding the likelihood of each fragment. Based on the multinomial distribution, removal fragments are sampled according to the desired number of modified structures.

For each selected removal fragment, DeepBioisostere explores the fragment library to find suitable insertion fragments for property control. Since the chemical modification of a certain part should retain its surroundings, molecular fragments with different numbers of fragmented sites are not considered. In addition, among the corresponding fragments, only those that satisfy the BRICS rules at the respective removal site of the original molecule are considered. All the passed insertion fragments are then evaluated by the DeepBioisostere model, and top insertion fragments are greedily selected. We note that, during the insertion fragment selection, either their occurrence data from the MMP database we created or from public chemical libraries like ChEMBL is not utilized because the differences in their occurrences are implicitly captured during the training process.

Lastly, the selected insertion fragments are examined in their attachment orientation. The combinations between fragmented sites of the insertion fragment and the surroundings are enumerated and again filtered with the BRICS rules. The passed combinations are then prioritized by the DeepBioisostere model, and the insertion fragments are attached following the selected attachment orientations. More details about the generation process can be found in the supplementary information.

## Data Availability

The code for processing data from the ChEMBL database is available at Github: https://github.com/ACE-KAIST/DeepBioisostere.git

## Code Availability

The code for training DeepBioisostere, sampling molecules, and evaluating the generated structures is available at Github: https://github.com/ACE-KAIST/DeepBioisostere.git

## References

[1] Di, L., Kerns, E.H.: Chapter 1 - introduction. In: Di, L., Kerns, E.H. (eds.) Drug-Like Properties (Second Edition), Second edition edn., pp. 1–3. Academic Press, Boston (2016). https://doi.org/10.1016/B978-0-12-801076-1.00001-0 . https://www.sciencedirect.com/science/article/pii/B9780128010761000010

[2] Wouters, O.J., McKee, M., Luyten, J.: Estimated research and development investment needed to bring a new medicine to market, 2009-2018. Jama **323**(9), 844–853 (2020)

[3] Schneider, G., Neidhart, W., Giller, T., Schmid, G.: "scaffold-hopping" by topological pharmacophore search: a contribution to virtual screening. Angewandte Chemie International Edition **38**(19), 2894–2896 (1999)

[4] Hu, Y., Stumpfe, D., Bajorath, J.: Recent advances in scaffold hopping. J. Med. Chem. **60**(4), 1238–1246 (2016) https://doi.org/10.1021/acs.jmedchem.6b01437

[5] Takeuchi, K., Kunimoto, R., Bajorath, J.: Systematic mapping of r-group space enables the generation of an r-group replacement system for medicinal chemistry. European Journal of Medicinal Chemistry **225**, 113771 (2021)

[6] Langdon, S.R., Ertl, P., Brown, N.: Bioisosteric replacement and scaffold hopping in lead generation and optimization. Mol. Inf. **29**(5), 366–385 (2010) https://doi.org/10.1002/minf.201000019

[7] Dick, A., Cocklin, S.: Bioisosteric replacement as a tool in anti-HIV drug design. Pharmaceuticals **13**(3), 36 (2020) https://doi.org/10.3390/ph13030036

[8] Cao, X., Yang, H., Liu, C., Zhang, R., Maienfisch, P., Xu, X.: Bioisosterism and scaffold hopping in modern nematicide research. J. Agric. Food Chem. **70**(36), 11042–11055 (2022) https://doi.org/10.1021/acs.jafc.2c00785

[9] Zou, Y., Liu, L., Liu, J., Liu, G.: Bioisosteres in drug discovery: focus on tetrazole. Future Science (2020)

[10] Malik, M.A., Wani, M.Y., Al-Thabaiti, S.A., Shiekh, R.A.: Tetrazoles as carboxylic acid isosteres: chemistry and biology. J Incl Phenom Macrocycl Chem **78**(1-4), 15–37 (2013) https://doi.org/10.1007/s10847-013-0334-x

[11] Subbaiah, M.A.M., Meanwell, N.A.: Bioisosteres of the phenyl ring: Recent strategic applications in lead optimization and drug design. Journal of Medicinal Chemistry **64**(19), 14046–14128 (2021) https://doi.org/10.1021/acs.jmedchem.1c01215

[12] Li, T., Liu, M., Liu, Y., Zhang, Y.: Identification of isosteric replacements of glycosyl domain of ligands by data mining (2021) https://doi.org/10.21203/rs.3.rs-955132/v1

[13] Papadatos, G., Brown, N.: In silico applications of bioisosterism in contemporary medicinal chemistry practice. WIREs Computational Molecular Science **3**(4), 339–354 (2013) https://doi.org/10.1002/wcms.1148

[14] Wirth, M., Zoete, V., Michielin, O., Sauer, W.H.: Swissbioisostere: a database of molecular replacements for ligand design. Nucleic acids research **41**(D1), 1137–1143 (2013)

[15] Cai, C., Gong, J., Liu, X., Gao, D., Li, H.: Molecular similarity: Methods and

performance. Chin. J. Chem. **31**(9), 1123–1132 (2013) https://doi.org/10.1002/cjoc.201300390

[16] Ertl, P.: World wide web-based system for the calculation of substituent parameters and substituent similarity searches. Journal of Molecular Graphics and Modelling **16**(1), 11–13 (1998)

[17] Schuffenhauer, A., Gillet, V.J., Willett, P.: Similarity searching in files of three-dimensional chemical structures: analysis of the BIOSTER database using two-dimensional fingerprints and molecular field descriptors. J. Chem. Inf. Comput. Sci. **40**(2), 295–307 (1999) https://doi.org/10.1021/ci990263g

[18] Thormann, M., Klamt, A., Hornig, M., Almstetter, M.: Cosmosim: Bioisosteric similarity based on cosmo-rs $\sigma$ profiles. Journal of Chemical Information and Modeling **46**(3), 1040–1053 (2006) https://doi.org/10.1021/ci050464m

[19] Devereux, M., Popelier, P.L.A., McLay, I.M.: Quantum isostere database: A web-based tool using quantum chemical topology to predict bioisosteric replacements for drug design. J. Chem. Inf. Model. **49**(6), 1497–1513 (2009) https://doi.org/10.1021/ci900085d

[20] Wagener, M., Lommerse, J.P.M.: The quest for bioisosteric replacements. J. Chem. Inf. Model. **46**(2), 677–685 (2006) https://doi.org/10.1021/ci0503964

[21] Wood, D.J., Vlieg, J.d., Wagener, M., Ritschel, T.: Pharmacophore fingerprint-based approach to binding site subpocket similarity and its application to bioisostere replacement. Journal of chemical information and modeling **52**(8), 2031–2043 (2012)

[22] Krier, M., Hutter, M.C.: Bioisosteric similarity of molecules based on structural alignment and observed chemical replacements in drugs. Journal of chemical information and modeling **49**(5), 1280–1297 (2009)

[23] Ujváry, I.: Extended summary: Bioster—a database of structurally analogous compounds. Pesticide Science **51**(1), 92–95 (1997)

[24] Desaphy, J., Rognan, D.: sc-PDB-frag: A database of protein–ligand interaction patterns for bioisosteric replacements. J. Chem. Inf. Model. **54**(7), 1908–1918 (2014) https://doi.org/10.1021/ci500282c

[25] Lešnik, S., Škrlj, B., Eržen, N., Bren, U., Gobec, S., Konc, J., Janežič, D.: Bober: web interface to the base of bioisosterically exchangeable replacements. Journal of Cheminformatics **9**(1), 1–8 (2017)

[26] Griffen, E., Leach, A.G., Robb, G.R., Warner, D.J.: Matched molecular pairs as a medicinal chemistry tool. J. Med. Chem. **54**(22), 7739–7750 (2011) https://doi.org/10.1021/jm200452d

[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., *et al.*: Chembl: towards direct deposition of bioassay data. Nucleic acids research **47**(D1), 930–940 (2019)

[28] Timmerman, H., Todeschini, R., Consonni, V., Mannhold, R., Kubinyi, H.: Handbook of molecular descriptors. Weinheim: Wiley-VCH (2002)

[29] Ertl, P.: Identification of bioisosteric substituents by a deep neural network. Journal of Chemical Information and Modeling **60**(7), 3369–3375 (2020) https://doi.org/10.1021/acs.jcim.0c00290

[30] He, J., You, H., Sandström, E., Nittinger, E., Bjerrum, E.J., Tyrchan, C., Czechtizky, W., Engkvist, O.: Molecular optimization by capturing chemist's intuition using deep neural networks. Journal of Cheminformatics **13**(1) (2021) https://doi.org/10.1186/s13321-021-00497-0

[31] He, J., Nittinger, E., Tyrchan, C., Czechtizky, W., Patronov, A., Bjerrum, E.J., Engkvist, O.: Transformer-based molecular optimization beyond matched molecular pairs. Journal of Cheminformatics **14**(1) (2022) https://doi.org/10.1186/s13321-022-00599-3

[32] Yang, L., Jin, C., Yang, G., Bing, Z., Huang, L., Niu, Y., Yang, L.: Transformer-based deep learning method for optimizing ADMET properties of lead compounds. Physical Chemistry Chemical Physics **25**(3), 2377–2385 (2023) https://doi.org/10.1039/d2cp05332b

[33] Jin, W., Barzilay, R., Jaakkola, T.: Hierarchical graph-to-graph translation for molecules. arXiv preprint arXiv:1907.11223 (2019)

[34] Jin, W., Barzilay, R., Jaakkola, T.: Hierarchical generation of molecular graphs using structural motifs. In: International Conference on Machine Learning, pp. 4839–4848 (2020). PMLR

[35] Chen, Z., Min, M.R., Parthasarathy, S., Ning, X.: A deep generative model for molecule optimization via one fragment modification. Nature Machine Intelligence **3**(12), 1040–1049 (2021) https://doi.org/10.1038/s42256-021-00410-2

[36] Green, H., Koes, D.R., Durrant, J.D.: Deepfrag: a deep convolutional neural network for fragment-based lead optimization. Chemical Science **12**(23), 8036–8047 (2021)

[37] Degen, J., Wegscheid-Gerlach, C., Zaliani, A., Rarey, M.: On the art of compiling and using 'drug-like' chemical fragment spaces. ChemMedChem **3**(10), 1503–1507 (2008) https://doi.org/10.1002/cmdc.200800178

[38] Zhou, Z., Kearnes, S., Li, L., Zare, R.N., Riley, P.: Optimization of molecules via

deep reinforcement learning. Scientific reports **9**(1), 10752 (2019)

[39] Wildman, S.A., Crippen, G.M.: Prediction of physicochemical parameters by atomic contributions. Journal of chemical information and computer sciences **39**(5), 868–873 (1999)

[40] Bickerton, G.R., Paolini, G.V., Besnard, J., Muresan, S., Hopkins, A.L.: Quantifying the chemical beauty of drugs. Nature chemistry **4**(2), 90–98 (2012)

[41] Ertl, P., Schuffenhauer, A.: Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. Journal of cheminformatics **1**, 1–11 (2009)

[42] Walters, W.P.: Virtual chemical libraries: miniperspective. Journal of medicinal chemistry **62**(3), 1116–1124 (2018)

[43] Gao, W., Coley, C.W.: The synthesizability of molecules proposed by generative models. Journal of chemical information and modeling **60**(12), 5714–5723 (2020)

[44] WHO Target product profiles for COVID-19 Therapeutics for non-hospitalized patients. https://www.who.int/publications/m/item/who-target-product-profiles-for-covid-19-therapeutics-for-non-hospitalized-patients. (Accessed: 2023-11-21)

[45] Unoh, Y., Uehara, S., Nakahara, K., Nobori, H., Yamatsu, Y., Yamamoto, S., Maruyama, Y., Taoda, Y., Kasamatsu, K., Suto, T., *et al.*: Discovery of s-217622, a noncovalent oral sars-cov-2 3cl protease inhibitor clinical candidate for treating covid-19. Journal of medicinal chemistry **65**(9), 6499–6512 (2022)

[46] Japan grants emergency regulatory approval for Shionogi's Covid-19 drug. https://www.pharmaceutical-technology.com/news/japan-approval-shionogi-covid-19-drug/?cf-view. (Accessed: 2023-11-21)

[47] Kiso, M., Yamayoshi, S., Iida, S., Furusawa, Y., Hirata, Y., Uraki, R., Imai, M., Suzuki, T., Kawaoka, Y.: In vitro and in vivo characterization of sars-cov-2 resistance to ensitrelvir. Nature Communications **14**(1), 4231 (2023)

[48] Ip, J.D., Chu, A.W.-H., Chan, W.-M., Leung, R.C.-Y., Abdullah, S.M.U., Sun, Y., To, K.K.-W.: Global prevalence of SARS-CoV-2 3CL protease mutations associated with nirmatrelvir or ensitrelvir resistance. eBioMedicine **91**, 104559 (2023)

[49] Iketani, S., Mohri, H., Culbertson, B., Hong, S.J., Duan, Y., Luck, M.I., Annavajhala, M.K., Guo, Y., Sheng, Z., Uhlemann, A.-C., *et al.*: Multiple pathways for sars-cov-2 resistance to nirmatrelvir. Nature **613**(7944), 558–564 (2023)

[50] Duan, Y., Zhou, H., Liu, X., Iketani, S., Lin, M., Zhang, X., Bian, Q., Wang,

H., Sun, H., Hong, S.J., et al.: Molecular mechanisms of sars-cov-2 resistance to nirmatrelvir. Nature, 1–3 (2023)

[51] Salentin, S., Schreiber, S., Haupt, V.J., Adasme, M.F., Schroeder, M.: Plip: fully automated protein–ligand interaction profiler. Nucleic acids research **43**(W1), 443–447 (2015)

[52] Adasme, M.F., Linnemann, K.L., Bolz, S.N., Kaiser, F., Salentin, S., Haupt, V.J., Schroeder, M.: Plip 2021: Expanding the scope of the protein–ligand interaction profiler to dna and rna. Nucleic acids research **49**(W1), 530–534 (2021)

[53] Moon, S., Hwang, S.-Y., Lim, J., Kim, W.Y.: Pignet2: a versatile deep learning-based protein–ligand interaction prediction model for binding affinity scoring and virtual screening. Digital Discovery (2024)

[54] Koes, D.R., Baumgartner, M.P., Camacho, C.J.: Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise. Journal of chemical information and modeling **53**(8), 1893–1904 (2013)

[55] Trott, O., Olson, A.J.: Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of computational chemistry **31**(2), 455–461 (2010)

[56] Cuozzo, A., Daina, A., Perez, M.A.S., Michielin, O., Zoete, V.: SwissBioisostere 2021: updated structural, bioactivity and physicochemical data delivered by a reshaped web interface. Nucleic Acids Research **50**(D1), 1382–1390 (2021) https://doi.org/10.1093/nar/gkab1047

[57] Weininger, D.: Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. Journal of chemical information and computer sciences **28**(1), 31–36 (1988)

[58] Landrum, G.: RDKit: open-source cheminformatics. http://www.rdkit.org

[59] Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv (2014). https://doi.org/10.48550/ARXIV.1412.3555 . https://arxiv.org/abs/1412.3555

[60] Seo, S., Lim, J., Kim, W.Y.: Molecular generative model via retrosynthetically prepared chemical building block assembly. Advanced Science **10**(8), 2206674 (2023)

## Acknowledgments

This work was supported by Basic Science Research Programs through the National Research Foundation of Korea funded by the Ministry of Science and ICT (Grant No. 2023R1A2C2004376 and RS-2023-00257479) to H.K., S.M., W.Z., and W.Y.K.

## Author contributions

H.K. and S.M. equally contributed to this work. They designed the methodology and carried out the experiments. All authors designed the experiments and analyzed the results. All authors wrote the manuscript together, and W.Y.K. supervised the project.

## Competing interests

The authors declare no competing interests.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CodeandSoftwaresubmissionchecklistDeepBioisostere.pdf
- DeepBioisostereESI.pdf
- EditorialPolicyChecklistDeepBioisostere.pdf
- ReportingSummaryDeepBioisostere.pdf