Journal of Computer-Aided Molecular Design (2023) 37:107–115
https://doi.org/10.1007/s10822-022-00493-y

DeepCubist: Molecular Generator for Designing Peptidomimetics based on Complex three-dimensional scaffolds

Kohei Umedera[1,2] · Atsushi Yoshimori[3] · Hengwei Chen[2] · Hiroyuki Kouji[4] · Hiroyuki Nakamura[1,5] · Jürgen Bajorath[2]

Received: 7 October 2022 / Accepted: 23 November 2022 / Published online: 3 December 2022

Abstract

Mimicking bioactive conformations of peptide segments involved in the formation of protein-protein interfaces with small molecules is thought to represent a promising strategy for the design of protein-protein interaction (PPI) inhibitors. For compound design, the use of three-dimensional (3D) scaffolds rich in sp3-centers makes it possible to precisely mimic bioactive peptide conformations. Herein, we introduce DeepCubist, a molecular generator for designing peptidomimetics based on 3D scaffolds. Firstly, enumerated 3D scaffolds are superposed on a target peptide conformation to identify a preferred template structure for designing peptidomimetics. Secondly, heteroatoms and unsaturated bonds are introduced into the template via a deep generative model to produce candidate compounds. DeepCubist was applied to design peptidomimetics of exemplary peptide turn, helix, and loop structures in pharmaceutical targets engaging in PPIs.

Keywords 3D scaffold · Peptidomimetics · Deep learning · Generative modeling

Introduction

While small molecules have been essential sources for drugs targeting enzymes or G-protein-coupled receptors (GPCRs), lead compounds for specifically interfering with protein-protein interactions (PPIs) have been difficult to obtain using conventional small molecular design approaches [1]. Many contemporary compound collections are strongly enriched with aromatic and other unsaturated compounds, due to the preferential use of efficient synthetic approaches such as palladium coupling reactions, giving rise to planar compounds with limited or absent 3D features [2]. Especially for developing PPI inhibitors, such predominantly "flat" molecular templates are typically unsuitable. Hence, there is a growing interest in utilizing scaffolds for design that are rich in sp^3 hybridized centers, thus having pronounced three-dimensional (3D) character and the ability to adopt compact molecular shapes.

Peptidomimetics are compounds designed to mimic the bioactive conformation of isolated peptides or specific peptide segments in proteins. While peptidomimetic design has a long tradition in drug discovery, it continues to be challenging, depending on the particular target and its ligand binding characteristics. This especially applies to the design of PPI inhibitors that are required to disrupt large protein-protein interfaces because in such cases, binding of a single small molecule must compensate for the free energy gained by large complementary protein surfaces forming many specific interactions.

Grossmann proposed four different classes of peptidomimetics: Class A-modified peptides (formed by α-amino acids with small backbone and side chain differences) ; Class B-modified peptides/foldamers (formed by amino

✉ Hiroyuki Nakamura
hiro@res.titech.ac.jp

✉ Jürgen Bajorath
bajorath@bit.uni-bonn.de

[1] School of Life Science and Technology, Tokyo Institute of Technology, 4259, Nagatsuta-cho, Midori-ku, 226-8503 Yokohama, Japan

[2] Department of Life Science Informatics, LIMES Program Unit Chemical Biology and Medicinal Chemistry, B-IT, Rheinische Friedrich-Wilhelms-Universität, Friedrich-Hirzebruch-Allee 5/6, D-53115 Bonn, Germany

[3] Institute for Theoretical Medicine, Inc, 26-1, Muraoka-Higashi 2-chome, 251-8555 Fujisawa, Kanagawa, Japan

[4] Oita University Institute of Advanced Medicine, Inc, 17-20, Higashi Kasuga-machi, 870-0037 Oita City, Oita, Japan

[5] Laboratory for Chemistry and Life Science, Institute of Innovative Research, Tokyo Institute of Technology, 4259, Nagatsuta-cho, Midori-ku, 226-8503 Yokohama, Japan

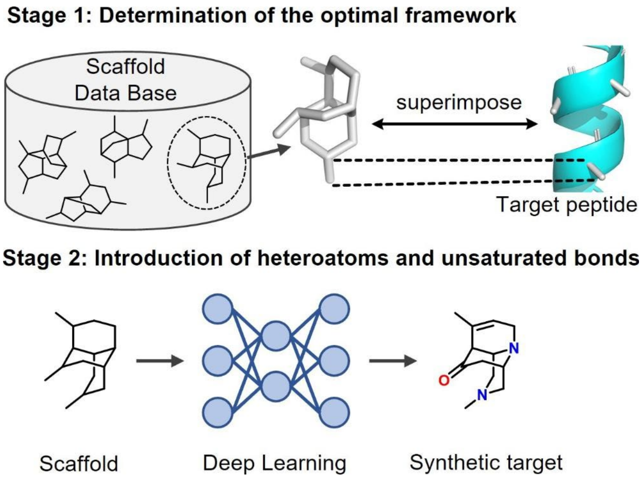
Stage 1: Determination of the optimal framework
Scaffold Data Base
superimpose
Target peptide
Stage 2: Introduction of heteroatoms and unsaturated bonds
O
N
N
Scaffold
Deep Learning
Synthetic target

Fig. 1 An overview of DeepCubist. The two major stages of the computational design approach are illustrated

acids with varying backbone and side chain iterations); Class C-structural mimetics (synthetic scaffolds with substitution sites corresponding to peptide side chain vectors); and Class D-mechanistic mimetics (small molecules mimicking the mode of action of peptides in the absence of side chain resemblance) [3]. Class C peptidomimetics completely replace the backbone of the parent peptide with a small molecule and reproduce the original side chain arrangements with its substituents. As a design premise, the development of peptidomimetics based on highly saturated bridged scaffolds, which can adopt complex shapes and enable the placement of functional groups in a variety of spatial arrangements for forming specific interactions can be expected to improve target selectivity and also potency compared to more planar scaffolds with less 3D character. This is the case because substitution sites in planar scaffolds have very limited ability to match the geometry of side chain arrangements across multiple amino acids in protein secondary structure elements. By contrast, 3D scaffolds increase the potential to precisely mimic side chains in peptides and match pharmacophores resulting from bioactive peptide conformations.

Scaffolds with new topology can in principle be obtained by computational enumeration of ring systems. For example, construction of a database called GDB4c containing 916,130 possible ring systems composed of up to four individual rings has been reported, enabling the discovery of new kinase inhibitors with previously unobserved chirality and shape [4]. For peptidomimetic design, 3D scaffolds must also be shape-diverse. Moreover, they must be capable of closely matching and replacing different peptide secondary structure motifs. At the same time, individual scaffolds should preferably be rigid to minimize entropic penalties upon binding.

Herein, we introduce DeepCubist, a molecular generator relying on deep learning for designing peptidomimetics based on previously unobserved 3D scaffolds and report initial proof-of-concept applications. In practice, the best fitting 3D scaffolds can be identified for turns, loops, or helical segments in structures of target proteins of interest and chemically diversified to obtain peptidomimetic candidates for interfering with PPIs. Hence, the DeepCubist approach complements and further extends structure-based design of PPI inhibitors by generating peptidomimetics with varying chemical features.

Development of DeepCubist

Methodological Concept

DeepCubist is conceptualized to include two design stages, as illustrated in Fig. 1. At the first stage, a preferred scaffold for reproducing spatial side chain arrangements of a target peptide is determined. Therefore, a database of 3D scaffolds with methyl groups initially placed at three substituent positions is constructed, enabling initial superposition of the scaffold and Cα-Cβ bond of the target peptide. At the second stage, heteroatoms and unsaturated bonds are introduced into selected frameworks to provide further functionalities and support synthetic accessibility.

Template scaffolds

Construction of DeepCubist's scaffold database began with defining a qualifying 3D scaffold as a *tricyclic or tetracyclic bridged ring system consisting of 5- and/or 6-membered rings*. This scaffold definition can be modified for different applications depending on the specific requirements. Our definition ensured that scaffold structures could be chemically diversified compared to, for example, bicyclic systems while restricting theoretically possible chemical complexity and hence increasing the likelihood of achieving synthetic accessibility. For our proof-of-concept investigation, so-defined scaffolds consisting of 10 to 14 carbon atoms were then systematically generated as illustrated in Fig. 2.

1) Six fused or bridged bicyclic systems consisting of 5- and/or 6-membered rings were computationally constructed as starting points (the number can be varied).

2) Tricyclic ring systems were then exhaustively generated by extensions of bicyclic systems with fragments comprising *m* carbon atoms added to any pair of ring atoms. From the resulting tricyclic ring systems, tetracyclic structures were obtained by addition of fragments with *n* carbon atoms to every atom pair of the tricyclic systems. Hence, (*m, n*) fragment combination were defined to obtain target

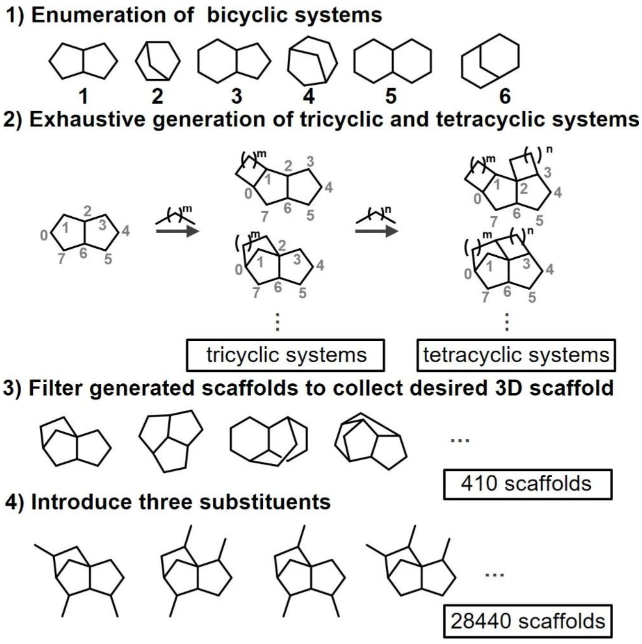
1) Enumeration of bicyclic systems
1
2
3
4
5
6
2) Exhaustive generation of tricyclic and tetracyclic systems
m
n
tricyclic systems
tetracyclic systems
3) Filter generated scaffolds to collect desired 3D scaffold
...
410 scaffolds
4) Introduce three substituents
...
28440 scaffolds

Fig. 2 Generation of a 3D scaffold database

scaffolds with 10 to 14 carbon atoms, depending on the size of the original bicyclic system.

For example, for bicyclic ring system **1** consisting of eight carbon atoms, $(m, n) = \{(2, 0), (1,1)\}$ were used to exhaustively construct scaffolds with 10 carbons. As a result of these operations, a total of 1347 bridged ring systems were obtained at this stage.

3) The generated tri- and tetracyclic candidate structures were then filtered to collect chemically feasible 3D scaffolds with limited strain energy. Scaffold conformers were generated using the "ligand preparation" option of Discovery Studio 2020 [5] and conformers with a "clean energy" value of no more than 100 kcal/mol were collected, yielding 405 different 3D scaffolds with no chiral information.

4) Finally, combinations of three substituents were added to each 3D scaffold, in each case permitting the presence of at most one quaternary carbon for ease of synthesis (the number of substituents can vary). The introduction of substituent combinations resulted in a total of 28,440 unique carbon atom scaffolds with no chiral information. These carbon atom scaffolds can be classified as *3D cyclic skeletons*, following the hierarchical scaffold definition of Bemis & Murcko [6]. These skeletons served as input for the design of final 3D scaffolds containing heteroatoms and unsaturated bonds, as further described below.

Generative model

Once 3D carbon skeletons are generated, they must be converted into chemically meaningful scaffolds. For this purpose, DeepCubist employs a deep generative model based on SMILES strings [7] as a standard text-based molecular representation. Such generative models have been applied, for example, to construct target-focused virtual libraries [8] or natural product-like compounds [9], demonstrating the ability to generate chemical structures of varying complexity. For training such models, SMILES of existing compounds are often augmented with randomized SMILES [10] to support learning of the chemical language encoded by string representations. As a deep learning architecture, a *transformer* model from natural language processing was selected [11]. Different from other sequence-to-sequence models, transformer models operate on the basis of *attention* mechanisms that identify and highly weight the most important representation elements for achieving accurate predictions during the training phase [11]. As further discussed below, the transformer model was trained to convert 3D carbon scaffolds into compounds containing heteroatoms and unsaturated bonds, that is, candidate compounds with chemical features amenable to synthesis.

Source and target structures for training

Drug- and natural product-like compounds were retrieved from ChEMBL version 30 [12] and COCONUT [13], a database of natural products, respectively. A total of 1,914,739 ChEMBL and 406,919 COCONUT compounds were obtained, referred to as original compounds. For model derivation, all possible *target* (output) structures were extracted from the original ChEMBL and COCONUT compounds by removing all exocyclic atoms from primary ring substituents and replacing removed fragments with a hydrogen atom (including, for example, ester, amide, or sulfone moieties), as illustrated in Fig. 3. Thus, target structures represented consistently defined scaffolds with primary substituents for deep learning and candidate structures for further chemical modifications. *Source* (input) structures were then obtained by converting target structures into cyclic skeletons through replacement of all heteroatoms with carbons and conversion of all bond orders to 1 (single bonds), as also illustrated in Fig. 3. After original compounds were decomposed, target structures with no more than eight atoms in individual rings and {C, N, O, S, F, Cl, Br, I} elements were collected for modeling. A total of 53,075 pairs of target and corresponding source structures were obtained. The use of these pairs of corresponding source and target structures for model derivation provided the basis for the generation of 3D scaffolds containing heteroatoms and unsaturated bonds from our newly generated database of 3D carbon skeletons described above. The 53,075 target structures were found to contain 268 of the total of 405 enumerated 3D scaffolds; hence, the remaining 137 scaffolds were novel.

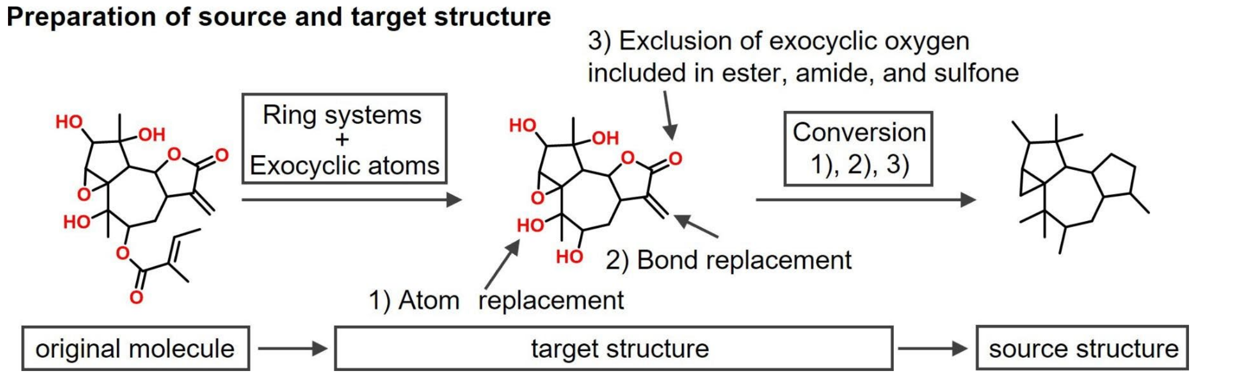

Fig. 3 Source and target structures for training the transformer model

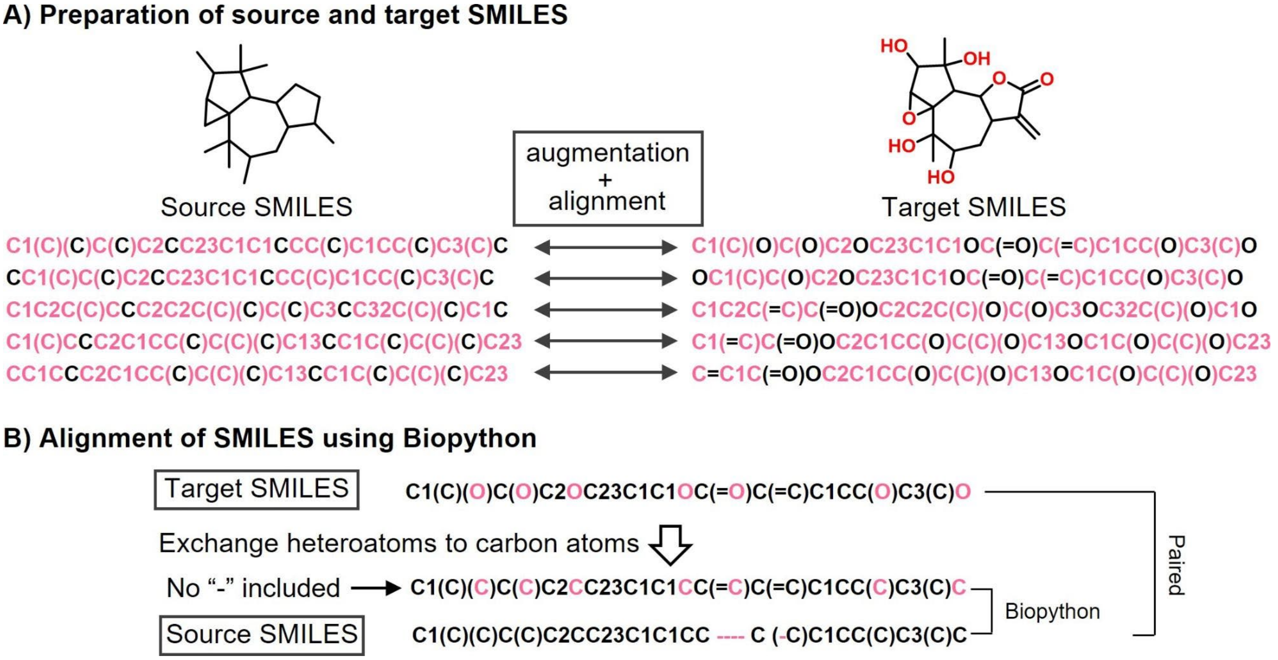
A) Preparation of source and target SMILES
Source SMILES
augmentation
+
alignment
HO
OH
O
O
O
HO
HO
Target SMILES
C1(C)(C)C(C)C2CC23C1C1CCC(C)C1CC(C)C3(C)C
CC1(C)C(C)C2CC23C1C1CCC(C)C1CC(C)C3(C)C
C1C2C(C)CCC2C2C(C)(C)C(C)C3CC32C(C)(C)C1C
C1(C)CCC2C1CC(C)C(C)(C)C13CC1C(C)C(C)(C)C23
CC1CCC2C1CC(C)C(C)(C)C13CC1C(C)C(C)(C)C23
C1(C)(O)C(O)C2OC23C1C1OC(=O)C(=C)C1CC(O)C3(C)O
OC1(C)C(O)C2OC23C1C1OC(=O)C(=C)C1CC(O)C3(C)O
C1C2C(=C)C(=O)OC2C2C(C)(O)C(O)C3OC32C(C)(O)C1O
C1(=C)C(=O)OC2C1CC(O)C(C)(O)C13OC1C(O)C(C)(O)C23
C=C1C(=O)OC2C1CC(O)C(C)(O)C13OC1C(O)C(C)(O)C23
B) Alignment of SMILES using Biopython
Target SMILES
C1(C)(O)C(O)C2OC23C1C1OC(=O)C(=C)C1CC(O)C3(C)O
Exchange heteroatoms to carbon atoms
No “-” included
C1(C)(C)C(C)C2CC23C1C1CC(=C)C(=C)C1CC(C)C3(C)C
Source SMILES
C1(C)(C)C(C)C2CC23C1C1CC ---- C (-C)C1CC(C)C3(C)C
Biopython
Paired

Fig. 4 Molecular representations. (A) illustrates the generation and (B) the alignment of source and target SMILES for transformer training

String representations for training

For converting scaffolds into compounds using SMILES-based deep generative models, substitution sites in input structures are often marked as wild-card sites such as “*” to enable chemical diversification [14–16]. Furthermore, transformer-based retrosynthetic predictions have been improved by minimizing the edit distance between augmented input and output SMILES strings compared to unique canonical SMILES [17]. The edit distance between two SMILES strings is defined as the number of editing operations consisting of insertion, deletion, and substitution for transforming one string into the other. Corresponding SMILES representations with minimized edit distance closely link these representations for learning, which tends to reduce errors rates. In our study, this strategy was applied for model derivation, as illustrated in Fig. 4 A. After source and target SMILES were augmented by generating additional SMILES rooted at each atom using RDkit [18], newly generated SMILES with smallest edit distance were paired using the sequence alignment module implemented in Biopython [19], as shown in Fig. 4B. In accordance with the DeepCubist design strategy, heteroatoms in target SMILES were replaced with carbon atoms to obtain corresponding source SMILES. Then, the additional SMILES strings were aligned with the original source SMILES using the “pairwise2.align.globalxx” function of Biopython. In the alignment, identical characters obtain a score of 1, otherwise the score is 0. Since source structures were generated from target structures, gaps (“-”) in aligned SMILES strings can only occur in source SMILES.

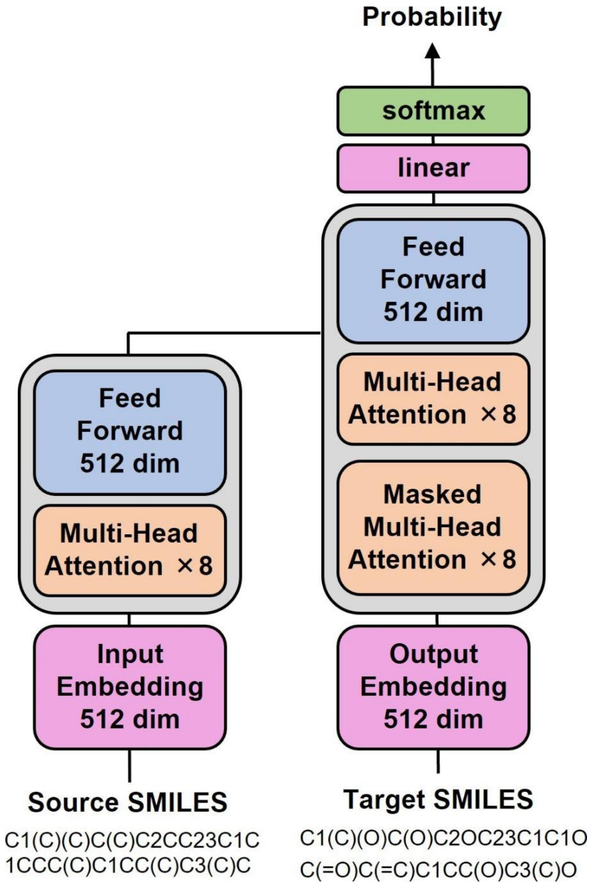

Fig. 5 Transformer architecture and parameter settings

Model derivation

Pairs of source and target structures were randomly divided into 42,990 training (90%) and 4777 validation set (10%) instances. Following data separation, the SMILES augmentation and alignment steps were carried out. Original SMILES were iteratively augmented with randomized SMILES to obtain a total number of 168,137 pairs for training and 18,624 pairs for validation. A multi-head attention transformer model was constructed using Pytorch [20]. SMILES tokens were embedded in 512 dimensions, the number of heads was set as 8, the number of sub-layers in both encoder and decoder units was set to 3, and the dimensionality of the feed-forward network model was set to 512. For all remaining parameters, default settings were used. The model architecture including parameter settings is schematically illustrated in Fig. 5. For structure generation, SMILES tokens were sampled according to the learned probability distribution.

Scripts for the calculations and the data can be obtained via the following link:

https://www.dropbox.com/s/4gdhew9xjit43e4/DeepCubist_Materials.zip?dl=0.

Exemplary applications

Scaffold retention

Initially, the effect of minimizing the edit distance between source and target SMILES was assessed. For transformer models trained using pairs of canonical SMILES (Fig. 6 A) or augmented and aligned SMILES (Fig. 6B), the loss value during validation was lower for augmented and aligned SMILES. Furthermore, for canonical SMILES, input structures were only poorly retained in target structures. After generating 100 unique structures using skeleton **7** as input with the model trained for 30 epochs, only nine target structures were found to completely retain the ring systems and the positions of substituents of source structures. In many cases, output structures with different ring sizes were obtained. By contrast, when the model was trained over 30 epochs with augmented and aligned SMILES, structure retention significantly increased. Using the same skeleton **7** as input, 99 of 100 newly generated structures exactly matched the composition and topology of the source structures.

The model was further evaluated using 100 randomly selected skeletons as input. In this case, sampling of 100 new structures yielded ~ 94% string validity, ~ 99% input structure retention, and 100% output structure novelty (Fig. 6 C).

Designing peptidomimetics

Having confirmed that the trained transformer model could effectively retain input structures during scaffold generation, DeepCubist was used to design exemplary peptidomimetics. To assess the general applicability of the approach, three types of peptide secondary structure were selected as starting points including peptide turns, helices, and loops. For each of the examples discussed below, a consistent number of 100 unique scaffolds was sampled. Furthermore, as an initial assessment of synthetic feasibility, the synthetic accessibility (SA) score according to Ertl & Schuffenhauer [21] were calculated using RDkit. The SA score ranges from 1 (easy synthesis) to 10 (very difficult) [21].

Turn mimetics

The tripeptide Glu-Asp-Leu is an inhibitor of HIV-1 protease. X-ray crystallography revealed that this tripeptide adopts a turn-like bioactive conformation in the active site of the enzyme [22]. The Cα-Cβ bonds of the tripeptide were superimposed on attachment points of 3D skeletons stored in DeepCubist's database using the "rdAlignment.GetAlignmentTransform" module of RDkit. Tetracyclic skeleton **8** was discovered to best reproduce the side chain orientations

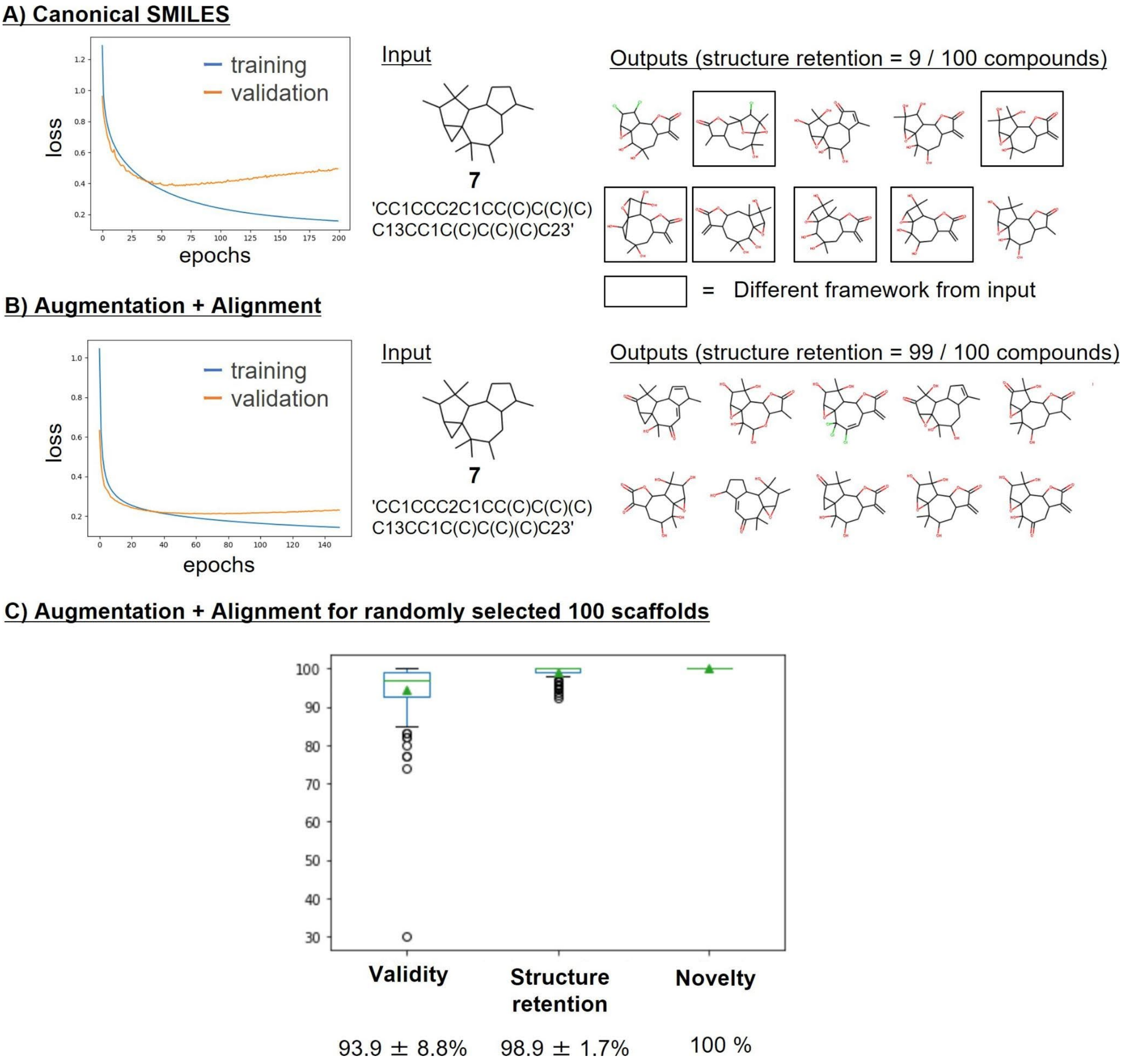

Fig. 6 Model derivation. Results of transformer training and validation are shown including the evolution of the loss function (left) and input/output scaffolds (right) for (A) canonical or (B) augmented and aligned SMILES. (C) Analysis results are reported for 100 randomly selected skeletons

of the tripeptide in its bioactive conformation with a sum of squared deviations (SSD) value of 0.944 Å^2 (Fig. 7). The structure of skeleton **8** was not found in compounds used for model training. With five-fold augmentation of the skeleton **8** input SMILES, 100 unique output molecules retaining the input structure were obtained, 98 of which contained heteroatoms and double bonds. For the 100 structures, the mean SA score was 6.17, indicating reasonable synthetic accessibility.

Helix mimetics

The peroxisome proliferator-activated receptor-γ (PPAR-γ) is a transcription factor that interacts with steroid receptor co-activating factor-1 (SRC-1) via the LxxLL motif presented by an α-helical peptide structure [23]. This LxxLL motif is widely observed at protein-protein interfaces [24]. If peptidomimetics were designed to precisely mimic varying conformations of this motif at different interfaces, selective PPI inhibitors might be obtained. Following the same calculation route as described above, DeepCubist identified

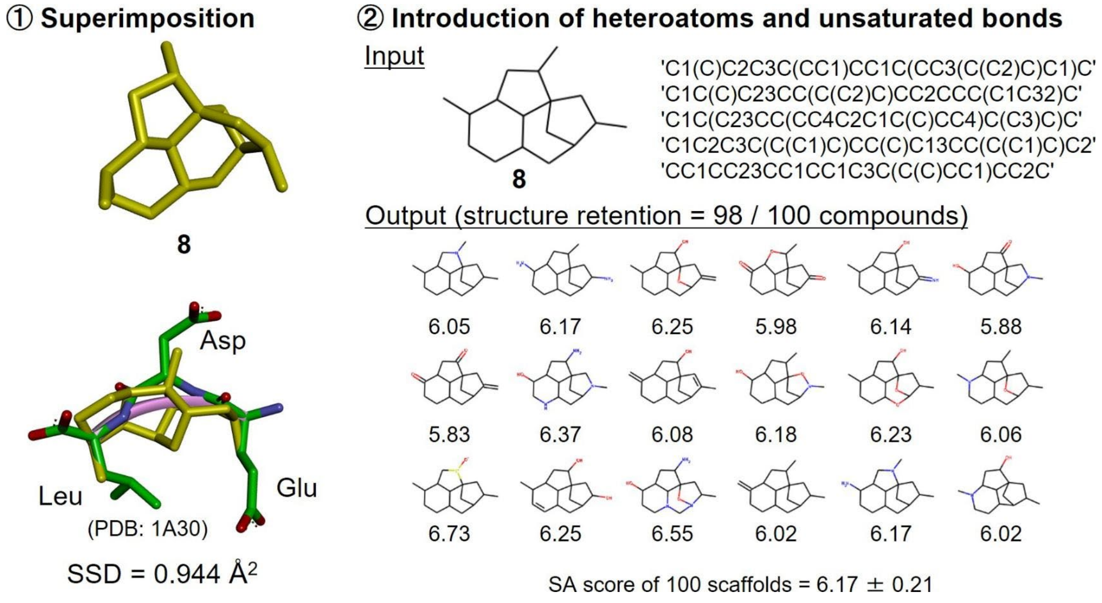
① Superimposition
8
Asp
Leu
Glu
(PDB: 1A30)
SSD = 0.944 Å²
② Introduction of heteroatoms and unsaturated bonds
Input
8
'C1(C)C2C3C(CC1)CC1C(CC3(C(C2)C)C1)C'
'C1C(C)C23CC(C(C2)C)CC2CCC(C1C32)C'
'C1C(C23CC(CC4C2C1C(C)CC4)C(C3)C)C'
'C1C2C3C(C(C1)C)CC(C)C13CC(C(C1)C)C2'
'CC1CC23CC1CC1C3C(C(C)CC1)CC2C'
Output (structure retention = 98 / 100 compounds)
6.05
6.17
6.25
5.98
6.14
5.88
5.83
6.37
6.08
6.18
6.23
6.06
6.73
6.25
6.55
6.02
6.17
6.02
SA score of 100 scaffolds = 6.17 ± 0.21

Fig. 7 Design of turn mimetics. Shown is the skeleton that best reproduced the side chain orientations of the tripeptide turn in its bioactive conformation with the corresponding sum of squared deviations (SSD) value obtained after rigid-body superimposition of the corresponding atom pairs (including Cα and Cβ atoms of the peptide residues). In addition, exemplary output structures are shown. Figures 8 and 9 are represented accordingly

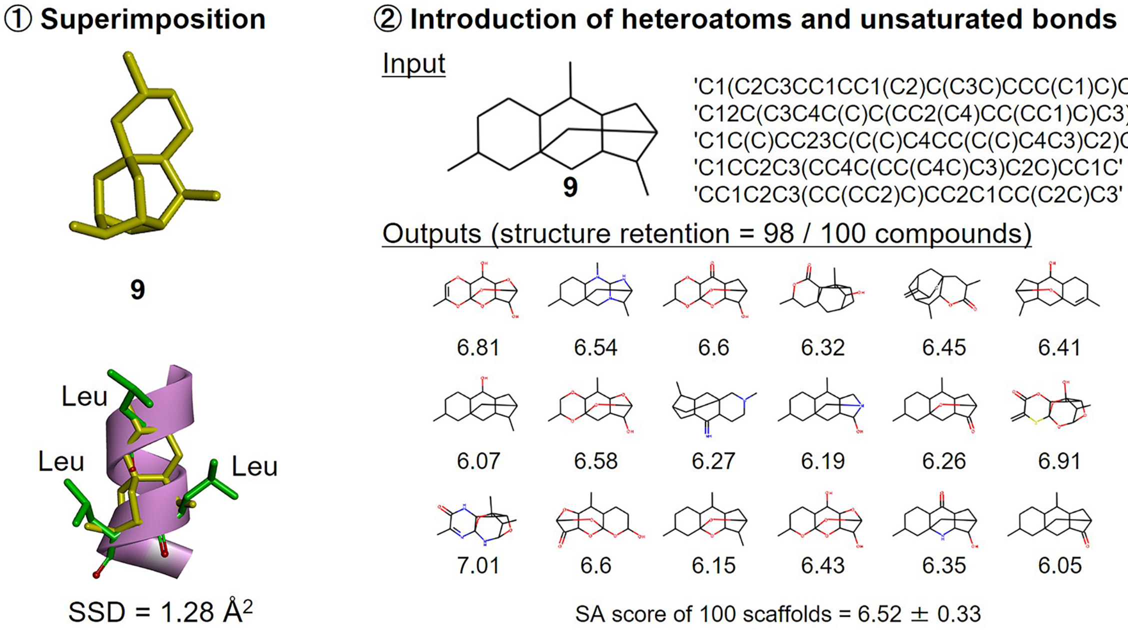
① Superimposition
9
Leu
Leu
Leu
SSD = 1.28 Å²
② Introduction of heteroatoms and unsaturated bonds
Input
9
'C1(C2C3CC1CC1(C2)C(C3C)CCC(C1)C)C'
'C12C(C3C4C(C)C(CC2(C4)CC(CC1)C)C3)C'
'C1C(C)CC23C(C(C)C4CC(C(C)C4C3)C2)C1'
'C1CC2C3(CC4C(CC(C4C)C3)C2C)CC1C'
'CC1C2C3(CC(CC2)C)CC2C1CC(C2C)C3'
Outputs (structure retention = 98 / 100 compounds)
6.81
6.54
6.6
6.32
6.45
6.41
6.07
6.58
6.27
6.19
6.26
6.91
7.01
6.6
6.15
6.43
6.35
6.05
SA score of 100 scaffolds = 6.52 ± 0.33

Fig. 8 Design of helix mimetics

tetracyclic skeleton **9** as the best available template for developing peptidomimetics, with an SSD value of 1.28 Å^2 (Fig. 8). Of 100 newly generated 3D scaffolds, 98 retained the input structure. In this case, the mean SA score was 6.52.

Loop mimetics

Modulation of nuclear factor erythroid 2-related factor(Nrf2) and Kelch-like-ECH-associated protein 1 (KEAP1) has been identified as an attractive therapeutic strategy for interfering with oxidative stress-related diseases such as cancer, neurodegenerative, cardiovascular, metabolic, or inflammatory

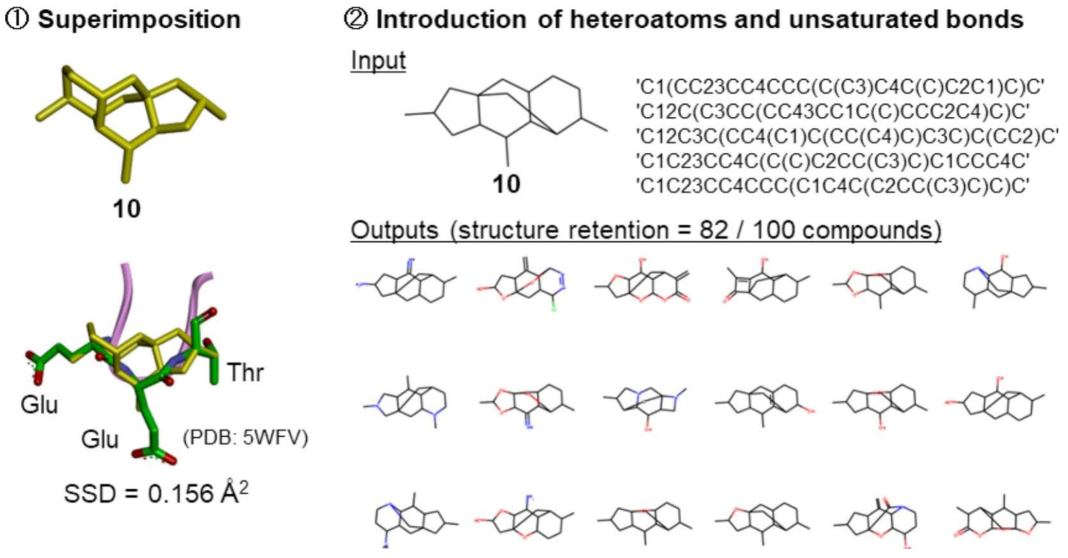
① Superimposition
10
Thr
Glu
Glu
(PDB: 5WFV)
SSD = 0.156 Å2
② Introduction of heteroatoms and unsaturated bonds
Input
10
'C1(CC23CC4CCC(C(C3)C4C(C)C2C1)C)C'
'C12C(C3CC(CC43CC1C(C)CCC2C4)C)C'
'C12C3C(CC4(C1)C(CC(C4)C)C3C)C(CC2)C'
'C1C23CC4C(C(C)C2CC(C3)C)C1CCC4C'
'C1C23CC4CCC(C1C4C(C2CC(C3)C)C)C'
Outputs (structure retention = 82 / 100 compounds)

Fig. 9 Design of loop mimetics

diseases [25]. Nrf2 was identified to form a peptide loop structure (different from well-defined turns) at a hot spot of Nrf2-KEAP1 interaction, indicating the potential of loop mimetics as inhibitors of this interaction [26]. DeepCubist identified unique tetracyclic skeleton **10** hat –to our knowledge– has thus far not been considered in drug discovery and design (Fig. 9). The SSD value for the superimposition was 0.156 Å^2. In this case, 85 of 100 unique output 3D scaffolds retained the input skeleton. The mean SA score for generated 100 scaffolds was 6.60.

Conclusion

In this study, we have introduced DeepCubist, a molecular generator for the design of peptidomimetics. DeepCubist includes a specialized database of complex sp3-rich skeletons as templates for design and a transformer model trained to convert preferred skeletons into viable 3D scaffolds including heteroatoms and unsaturated bonds. Minimizing the edit distance of input and output SMILES was found to be a simple and effective way to control overfitting and tune the transformer for the construction of chemically meaningful 3D scaffolds retaining input structures. To establish proof-of-concept for DeepCubist's design capacity, we have reported peptidomimetic designs for different peptide secondary structure motifs in high-profile therapeutic targets. While mostly favorable synthetic accessibility scores are obtained so far for newly generated 3D scaffolds, sp^3-rich compounds have often more limited synthetic accessibility thancombinations of popular aromatic ring systems. Therefore, future work will primarily concentrate on ensuring a high degree of synthetic feasibility of newly generated scaffolds, for which different methodological avenues such as template- or reaction-based design can be considered.

Author contributions All authors contributed to planning the study, conducting it, reviewing the results, and preparing the original and further revised draft.

Funding declaration The work was supported by the JSPS Overseas Program for Young Researchers (Fellowship to K.U.) and the China Scholarship Council (Fellowship to H.C.). No other funders are available.
Open Access funding enabled and organized by Projekt DEAL.

Data Availability The analysis was exclusively carried out using publicly available data, as specified in the text. Code and the data sets used are made freely available.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Barker A, Kettle JG, Nowak T, Pease JE (2013) Expanding medicinal chemistry space. Drug Discov Today 18:298–304. https://doi.org/10.1016/j.drudis.2012.10.008
2. Sauer WHB, Schwarz MK (2003) Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity. J Chem Inf Comput Sci 43:987–1003. https://doi.org/10.1021/ci025599w
3. Pelay-Gimeno M, Glas A, Koch O, Grossmann TN (2015) Structure-based design of inhibitors of protein-protein interactions: mimicking peptide binding epitopes. Angew Chem Int Ed 54:8896–8927. https://doi.org/10.1002/anie.201412070
4. Meier K, Arús-Pous J, Reymond J-L (2021) A potent and selective Janus kinase inhibitor with a chiral 3D-Shaped Triquinazine Ring System from Chemical Space. Angew Chem Int Ed 60:2074–2077. https://doi.org/10.1002/ange.202012049
5. Dassault Systèmes BIOVIA, Studio BIOVIAD (2020) San Diego: Dassault Systèmes, 2020
6. Bemis GW, Murcko MA (1996) The properties of known drugs. 1. Molecular frameworks. J Med Chem 39:2887–2893. https://doi.org/10.1021/jm9602928
7. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28:31–36. https://doi.org/10.1021/ci00057a005
8. Merk D, Friedrich L, Grisoni F, Schneider G (2018) De Novo Design of Bioactive Small Molecules by Artificial Intelligence. Mol Inf 37:1700153. https://doi.org/10.1002/minf.201700153
9. Zheng S, Yan X, Gu Q, Yang Y, Du Y, Lu Y, Xu J (2019) QBMG: quasi-biogenic molecule generator with deep recurrent neural network. J Cheminform 11:1–12. https://doi.org/10.1186/s13321-019-0328-9
10. Arús-Pous J, Johansson SV, Prykhodko O, Bjerrum EJ, Tyrchan C, Reymond J-L, Chen H, Engkvist O (2019) Randomized SMILES strings improve the quality of molecular generative models. J Cheminform 11:71. https://doi.org/10.1186/s13321-019-0393-0
11. Dollar O, Joshi N, Beck DAC, Pfaendtner J (2021) Attention-based generative models for de novo molecular design. Chem Sci 12:8362–8372. https://doi.org/10.1039/D1SC01050F
12. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magariños MP, Mosquera JF, Mutowo P, Nowotka M, Gordillo-Marañón M, Hunter F, Junco L, Mugumbate G, Rodriguez-Lopez M, Atkinson F, Bosc N, Radoux CJ, Segura-Cabrera A, Hersey A, Leach AR (2019) ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 47:D930–D940. https://doi.org/10.1093/nar/gky1075
13. Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C (2021) COCONUT online: Collection of Open Natural Products database. J Cheminform 13:1–13. https://doi.org/10.1186/s13321-020-00478-9
14. Arús-Pous J, Patronov A, Bjerrum EJ, Tyrchan C, Reymond J-L, Chen H, Engkvist O (2020) SMILES-based deep generative scaffold decorator for de-novo drug design. J Cheminform 12:1–18. https://doi.org/10.1186/s13321-020-00441-8
15. Kaitoh K, Yamanishi Y (2022) Scaffold-Retained structure generator to exhaustively create molecules in an arbitrary Chemical Space. J Chem Inf Model 62:2212–2225. https://doi.org/10.1021/acs.jcim.1c01130
16. Langevin M, Minoux H, Levesque M, Bianciotto M (2020) Scaffold-Constrained Molecular Generation. J Chem Inf Model 60:5637–5646. https://doi.org/10.1021/acs.jcim.0c01015
17. Zhong Z, Song J, Feng Z, Liu T, Jia L, Yao S, Wu M, Hou T, Song M (2022) Root-aligned SMILES: a tight representation for chemical reaction prediction. Chem Sci 13:9023–9034. https://doi.org/10.1039/D2SC02763A
18. Landrum G, Tosco P, Kelley B, Ric gedeck, Vianello R, NadineSchneider, Kawashima E, Cosgrove D, Dalke A, Dan N, Jones G, Cole B, Swain M, Turk S, AlexanderSavelyev, Vaucher A, Wójcikowski M, Take I, Probst D, Ujihara K, Scalfani VF, Godin G, Pahl A, Berenger F (2022) JLVarjo, strets, JP, DoliathGavid rdkit/rdkit: 2022_03_5 (Q1 2022) Release
19. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–1423. https://doi.org/10.1093/bioinformatics/btp163
20. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv:1912.01703v1. https://doi.org/10.48550/arXiv.1912.01703
21. Ertl P, Schuffenhauer A (2009) Estimation of synthetic accessibility score of Drug-Like Molecules based on Molecular Complexity and Fragment Contributions. J Cheminform 1:8. https://doi.org/10.1186/1758-2946-1-8
22. Louis JM, Dyda F, Nashed NT, Kimmel AR, Davies DR (1998) Hydrophilic peptides derived from the transframe region of Gag-Pol inhibit the HIV-1 protease. Biochemistry 37:2105–2110. https://doi.org/10.1021/bi972059x
23. Nolte RT, Wisely GB, Westin S, Cobb JE, Lambert MH, Kurokawa R, Rosenfeld MG, Willson TM, Glass CK, Milburn MV (1998) Ligand binding and co-activator assembly of the peroxisome proliferator-activated receptor-gamma. Nature 395:137–143. https://doi.org/10.1038/25931
24. Plevin MJ, Mills MM, Ikura M (2005) The LxxLL motif: a multifunctional binding sequence in transcriptional regulation. Trends Biochem Sci 30:66–69. https://doi.org/10.1016/j.tibs.2004.12.001
25. Magesh S, Chen Y, Hu L (2012) Small molecule modulators of Keap1-Nrf2-ARE pathway as potential preventive and therapeutic agents. Med Res Rev 32:687–726. https://doi.org/10.1002/med.21257
26. Zhong M, Lynch A, Muellers SN, Jehle S, Luo L, Hall DR, Iwase R, Carolan JP, Egbert M, Wakefield A, Streu K, Harvey CM, Ortet PC, Kozakov D, Vajda S, Allen KN, Whitty A (2020) Interaction Energetics and Druggability of the protein-protein Interaction between Kelch-like ECH-Associated protein 1 (KEAP1) and nuclear factor erythroid 2 like 2 (Nrf2). Biochemistry 59:563–581. https://doi.org/10.1021/acs.biochem.9b00943