Supplementary Information for

An SE(3)-equivariant language model for pocket-aware 3D molecular generation enables discovery of potent HPK1 inhibitors

Bin Xi[1,2†], Han Wang[1,2†], Guanglong Sun[2†], Bowen Zhang[3†], Ruihan Mao[2], Yuyang Ge[2], Yang Wang[2], Jiangtao Zhang[3], Yiting Pan[2], Feng Zhou[2], Xiaojian Xu[4], Yuji Wang[5], Zhenming Liu[1,6]*, Daohua Jiang[3]*, Huting Wang[2]*, Wenbiao Zhou[2]* and Bo Huang[1,2,5]*

[1]State Key Laboratory of Natural and Biomimetic Drugs, School of Pharmaceutical Sciences, Peking University, Beijing, China.

[2]Beijing StoneWise Technology Co Ltd., Haidian Street #15, Beijing, 100080, China.

[3]Laboratory of Soft Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing, China.

[4]Beijing Neurosurgical Institute, Capital Medical University, Beijing, China.

[5]College of Pharmaceutical Sciences, Capital Medical University, Beijing 100069, P. R. China.

[6]Key Laboratory of Xinjiang Endemic Phytomedicine Resources Ministry of Education; School of Pharmacy, Shihezi University, Shihezi 832003, Xinjiang, China.

[†]These authors contributed equally to this work.

*Corresponding author(s). E-mail(s): bohuang_011@163.com; zhouwenbiao@stonewise.cn; wanghuting@stonewise.cn; jiangdh@iphy.ac.cn; zmliu@bjmu.edu.cn

Contents

1. Model Construction

1.1 Training and Inference Algorithms

Pseudocode for the algorithms utilized in both the training and inference phases of the SE3-BiLingoMol model is provided. Ligand representation was established using the FSMILES[1] methodology.

The specific procedures for training the self-refinement module are not included within the primary model training algorithm. This is because the self-refinement module was instantiated as a replicate of the pre-trained ligand decoder, inheriting its architectural layers and corresponding parameters. Consequently, the training of this module was readily accomplished by adhering to the established model training protocol, specifically by incorporating a bidirectional attention mask, while maintaining fixed parameters for all other modules. Notably, the self-refinement module converged efficiently, requiring a substantially reduced number of epochs (e.g., 30 epochs) compared to the comprehensive training of the entire SE3-BiLingoMol model.

Algorithm 1 Model Training

Input: PDB files of all binding proteins, each with a co-crystallized or docked tool ligand (reference ligand) in the pocket

Input: SDF files of all small-molecule ligands

Input: Identified non-covalent interactions(NCI) by ODDT for each complex

Input: Identified pocket features, e.g., hydrogen-bond donor/acceptor, hydrophobicity, aromaticity, formal charges, and residue types

// Prepare training dataset

Define pockets as all residues that are within 6 $\mathring{A}$ from its reference ligand

Define contact sites as all pocket atom sites that are within 4 $\mathring{A}$ from its reference ligand

Define pocket NCI sites as the pocket atom sites of identified NCIs

Define pocket feature sites as the pocket atom sites of identified pocket features

Obtain pocket atom tokens using element types, and additional "pad", "undefined" tokens

Obtain ligand tokens using FSMILES tokenizer

// Training

num_epochs=300

btach_size=30

for each epoch **do**

  for each batch **do**

    Generate pocket multivector embeddings R^{pocket}

    Generate pocket residue embeddings $E^{residue}$

    Generate pocket feature embeddings $E^{feature}$

    Generate pocket NCI embeddings E^{NCIs}

    Generate pocket contact embeddings $E^{contact}$

    $E^{pocket} = E^{residue} + Concat(E^{feature}, E^{NCIs} + E^{contact})$

    Generate ligand multivector and FSMILES embeddings R^{lig}, E^{lig}

    $Z^{pocket,MV}, Z^{pocket,S} = \text{Encoder}(R^{pocket}, E^{pocket})$

    $Z^{lig,MV}, Z^{lig,S} = \text{Decoder}(R^{lig}, E^{lig}, Z^{pocket,MV}, Z^{pocket,S})$

    $\text{ligand\_token\_logits} = \text{MLP}_{token}(\text{Z}^{lig,MV}, Z^{lig,S})$

    $R^{lig}_{preds} = \text{MLP}_{coords}(\text{Z}^{lig,MV}[:, : \ -1, ...], Z^{lig,S}[:, : \ -1, ...], E^{lig}[:, 1 \ : \ , ...])$

    Compute ligand token loss and ligand coordinates loss

    Compute auxiliary loss

    Backpropagate the total loss and use AdamW optimizer to update model parameters

  end for

end for

Algorithm 2 Molecule Inference

Input: PDB file of a binding pocket
Input: SDF file of a reference ligand only for pocket contact site featurization
Input: (*Optional*) SDF file of a molecular fragment
Input: (*Optional*) Non-covalent ineraction sites of pocket atom

// Prepare input pocket embeddings
Obtain pocket multivector embeddings R^{pocket}
Obtain pocket token embedding $E^{residue}$
Obtain pocket features embedding $E^{feature}$
Obtain pocket contact embedding $E^{contact}$
if input ineraction sites are provided **then**
    Obtain pocket interaction embedding E^{NCIs} from input
else
    Obatin E^{NCIs} by randomly selecting 4 pocket feature sites that are closed to the pocket center
end if
// Prepare input ligand embeddings
if input molecular fragment is provided **then**
    Obtain ligand multivector embeddings R^{lig} from fragment atomic positions
    Obtain ligand token embeddings E^{lig} by FSMILES tokenization of the fragment
else
    Set E^{lig} to the embedding of '[START]' token
    Obtain R^{lig} using randomly selected contact site position embeddings
end if
// Start auto-regressive generation (generate one molecule as an example)
max_length= 100
temperature= 1
$Z^{pocket,MV}, Z^{pocket,S} = \text{Encoder}(R^{pocket}, E^{pocket})$
while current ligand sequence length $l <$ max_length **do**
    $Z^{lig,MV}, Z^{lig,S} = \text{Decoder}(R^{lig}, E^{lig}, Z^{pocket,MV}, Z^{pocket,S})$
    ligand_token_logits = $\text{MLP}_{token}(\text{Z}_l^{lig,MV}, Z_l^{lig,S})$
    ligand_token_probs = Softmax(ligand_token_logits/temperature)
    Sample a ligand token by multinomial sampling
    Obtain the embedding E_l^{lig} of the sampled token
    $R_l^{lig} = \text{MLP}_{coords}(\text{Z}_l^{lig,MV}, Z_l^{lig,S}, E_l^{lig})$
    if current token is '[END]' token **then**
        Terminate generation before reaching max_length
    end if
end while
$R_{refined}^{lig} = \text{SelfRefinement}(E^{lig}, R^{lig})$
Obtain generated molecule by transforming all predicted tokens and coordinates

Output: generated molecule

Algorithm 3 Molecular representation using fragment-SMILES

Input: A molecule SMILES with its 3D positions
Input: Predefined SMILES vocabulary, vocab1
Input: Predefined fragment-SMILES vocabulary, vocab2

Fragment the molecule into different fragments based on each single bond that is (1) not in a ring, (2) attached to a ring and (3) not attached to hydrogen atoms
FSMILES_list = ['start']
for For each fragment **do**
    Recognize and tokenize atoms, bonds, branches, rings, stereochemistry, charges and isotopes from SMILES using vocab1
    // Incorporate ring information into SMILES tokens
    Recognize ring type for each atom and set ring type '0' for atoms that do not form a ring and non-element tokens
    for For each token **do**
        Combine the SMILES token with its ring type
        Push the new token into FSMILES_list
    end for
    // Use a 'sep' token to distinguish different fragments
    FSMILES_list += 'sep'
end for
FSMILES_list += 'end'
Transform FSMILES_list into FSMILES_code using the index of each fragment-SMILES token in vocab2

// Process token positions
Centralize all atomic positions
Find the parent token (defined as an element token) for each token
Assign non-element token position as the one of its parent token

Output: fragment-SMILES token list and position list

1.2 Optimization Techniques Used for SE3-Bilingomol Inference Speedup

SE3-BiLingoMol functions as an SE(3)-equivariant Transformer-based language model, facilitating molecular generation via autoregressive sampling. During the inference phase, three principal optimization techniques were employed to significantly reduce computational time.

Firstly, a Key-Value (KV) cache was implemented. This optimization, widely adopted in many large language models, stores the key and value matrices from previous decoding steps, enabling their reuse in the current decoding step. This approach substantially reduces computational consumption by eliminating redundant key and value computations.

Secondly, consistent with the official implementation of the Geometric Algebra Transformer[2], multi-query attention[3] was utilized instead of multi-head attention to mitigate memory consumption.

Finally, a specific optimization was applied to multi-query attention computation. In standard implementations, GPUs repeatedly load identical key and value data into shared memory for different queries, thereby increasing bandwidth demands. To address this, all queries were loaded into shared memory simultaneously during decoding steps, which consequently reduced I/O operations. Concurrently, by concatenating all queries into a single matrix, the self-attention computation was transformed from General Matrix-Vector multiplication (GEMV) to General Matrix-Matrix multiplication (GEMM). This enabled more efficient utilization of GPU tensor cores, resulting in accelerated computation.

1.3 Features representations using the projective geometric algebra

Table S1. Representations of scalar features, 3D points and symmetry operations using the projective geometric algebra $G(R^{3,0,1})$. The non-zero component coefficients of the 16-dimensional multivector x are listed, and a generic multivector can be expressed as in Eq. 1.

Geometric objects/ symmetry operations	Multivector	mapping
Scalar $s \in \mathrm{R}$	x_s	$= s$
Point $p \in \mathrm{R}^3$	$(x_{012}, x_{013}, x_{023}, x_{123})$	$= (p, 1)$
Translation $t \in \mathrm{R}^3$	$(x_s, x_{01}, x_{02}, x_{03})$	$= (1, t/2)$
Rotation expressed as quaternion $q \in \mathrm{R}^4$	$(x_s, x_{01}, x_{02}, x_{03})$	$= q$
Point reflection through $p \in \mathrm{R}^3$	$(x_{012}, x_{013}, x_{023}, x_{123})$	$= (p, 1)$
Reflection through plane w/ normal $n \in \mathrm{R}^3$, origin shift $d \in \mathrm{R}$	(x_0, x_1, x_2, x_3)	$= (d, n)$

1.4 Hyperparameters of SE3-BiLingoMol

Supplementary Information Table S2. Hyperparameters of SE3-BiLingoMol for model architecture, training and inference.

Hyperparameter	Value
Number of pocket encoder blocks	20
Number of ligand decoder blocks	20
Number of self-refinement blocks	20
Hidden size of multi-vector channels	64
Hidden size of scalar channels	128
Hidden size of scalar embeddings	512
Scalar channel dimensions of equivariant MLP	[128, 256, 128]
Multi-vector channel dimensions of equivariant MLP	[64, 128, 64]
Number of attention heads	8
Scalar channel dimensions of token head	[128, 256, 256]
Multi-vector channel dimension of token head	[64, 128, 64]
Scalar channel dimension of coordinate head	[640, 256, 128]
Multi-vector channel dimension of coordinate head	[64, 128, 1]
Scalar channel dimension of coordinate head of Self-refinement module	[640, 256, 128]
Multi-vector channel dimension of coordinate head of Self-refinement module	[64, 128, 1]
Training	
Learning rate	7e-5
Batch size	30
Weight decay	0.01
Ligand token loss weight	1
Ligand coordinates loss weight	10
Auxiliary loss of ligand bond length	5
Auxiliary loss of ligand bond angle	8
Auxiliary loss of ligand dihedral angle	3
Inference	
Max. sequence length of ligand tokens	100
Max. sequence length of pocket tokens	500
Sampling temperature (for *de novo* generation)	1.0 (*suggested*)
Sampling temperature (for substructure-guided generation)	1.35 (*suggested*)
Number of samples per epoch	200

2. Model Evaluation

2.1 Definition of Evaluation Metrics

(i) **# Generated Molecules.** The sum of total number of generated molecules across 102 targets, with an upper limit of 1000 per target.

(ii) **Mean MW.** The averaged molecular weight of all generated molecules.

(iii) **Mean QED.** The average quantitative estimate of drug-likeness of all generated molecules.

(iv) **Mean SAS.** The average synthetic accessibility score of all generated molecules.

(v) **% Druglike Molecules.** The ratio of molecules with QED > 0.3 and SAS < 5 over all generated molecules.

(vi) **% Non-aromatic rings.** The score estimates the average ratio of non-aromatic rings in the generated molecules. Specifically, for each molecule, the atomic masses of all atoms in non-aromatic rings (determined using RDKit) are summed, and the non-aromatic ring ratio is defined as the ratio of the atomic masses of atoms of non-aromatic rings to its total molecular weight.

(vii) **Diversity.** The score measures molecular diversity, where the similarity score of average pair-wise Tanimoto similarity of molecules generated for the same target was first computed, and the Diversity was defined as 1 - similarity score.

(viii) **% Conformation Validity.** The ratio of molecules passing high-energy atom detection (HEAD)[4] overall all generated molecules.

(ix) **Strain Energy 25%, 50%, 75%**. Strain energy (in unit kcal/mol) of the 25% quantile, medium, and 75% quantile of all predicted strain energies by PoseCheck[5] at a level of Universal Force Fields.

(x) **Mean Min-in-place Score.** The weighted average of min-in-place docking score using "mininplace" method of Shrödinger Glide module[6], where the number of druglike molecules for different targets are considered as the weights. During calculation, we only considered druglike molecules with a negative min-in-place score.

(xi) **Mean Redocking Score.** The weighted average of redocking score using "docking" method of Shrödinger Glide module, where the number of druglike molecules for different targets are considered as the weights. During calculation, we only considered druglike molecules with a negative redocking score.

(xii) **% Min-in-place < Re-docking Score.** The ratio of molecules whose min-in-place docking score is lower than redocking score over all generated druglike molecules. During calculation, we only considered druglike molecules with a negative min-in-place and redocking score.

(xiii) **% IMP[7].** The ratio of molecules whose min-in-place docking score is lower than the lowest docking score of active compounds (calculated using Shrödinger Glide module with

"docking" method) on that target, over all generated druglike molecules. During calculation, we only considered druglike molecules with a negative min-in-place score.

(xiv) **PLIF Recovery Rate[8].** This metric is a measurement of the fraction of non-covalent interactions in the crystal ligand pose that are successfully replicated in the generated molecule pose using ProLIF, defined as,

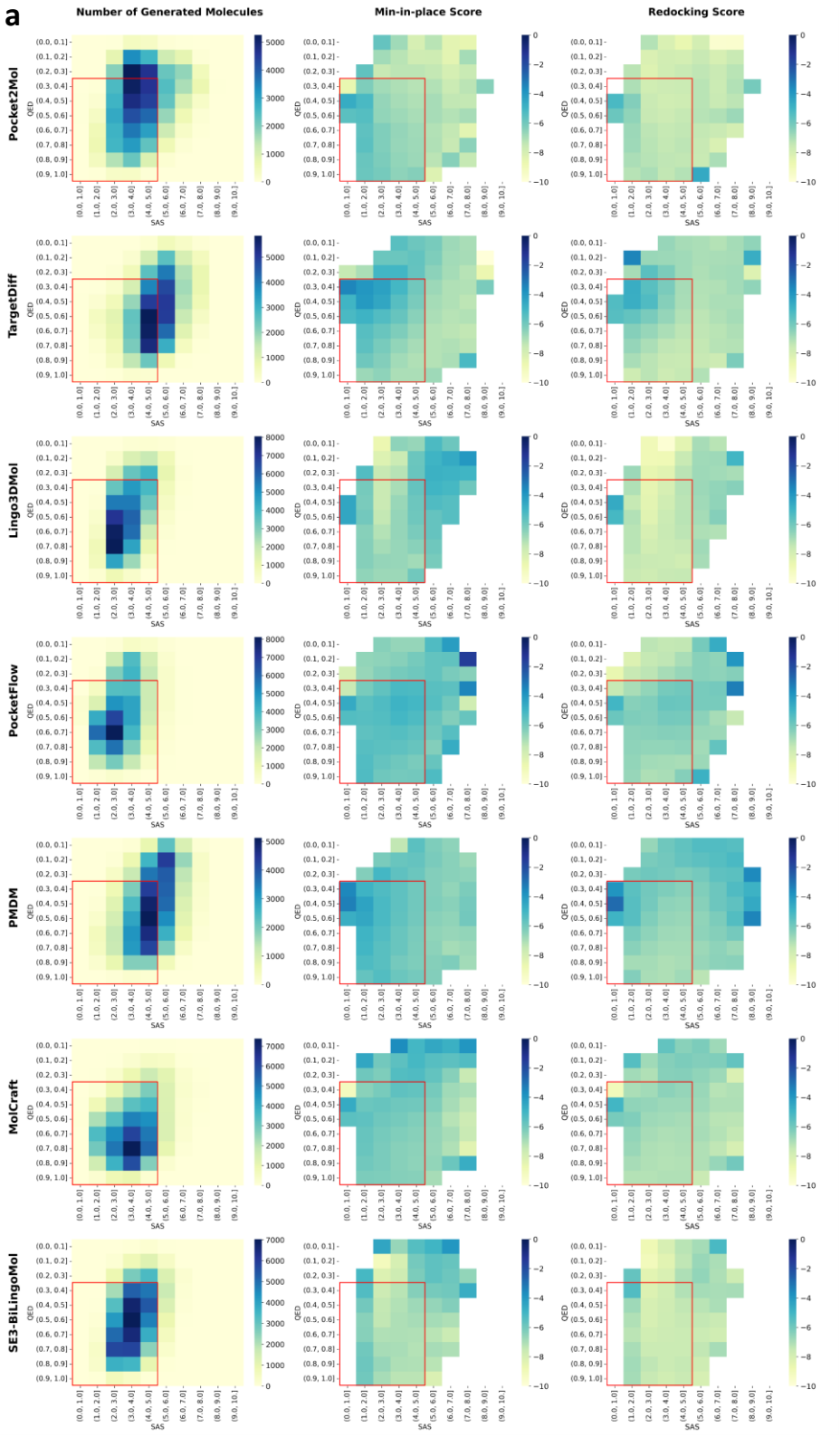
$$\text{PLIF Recovery} = \frac{\sum_{i,r} \min(C_{i,r}, P_{i,r})}{\sum_{i,r} C_{i,r}},$$

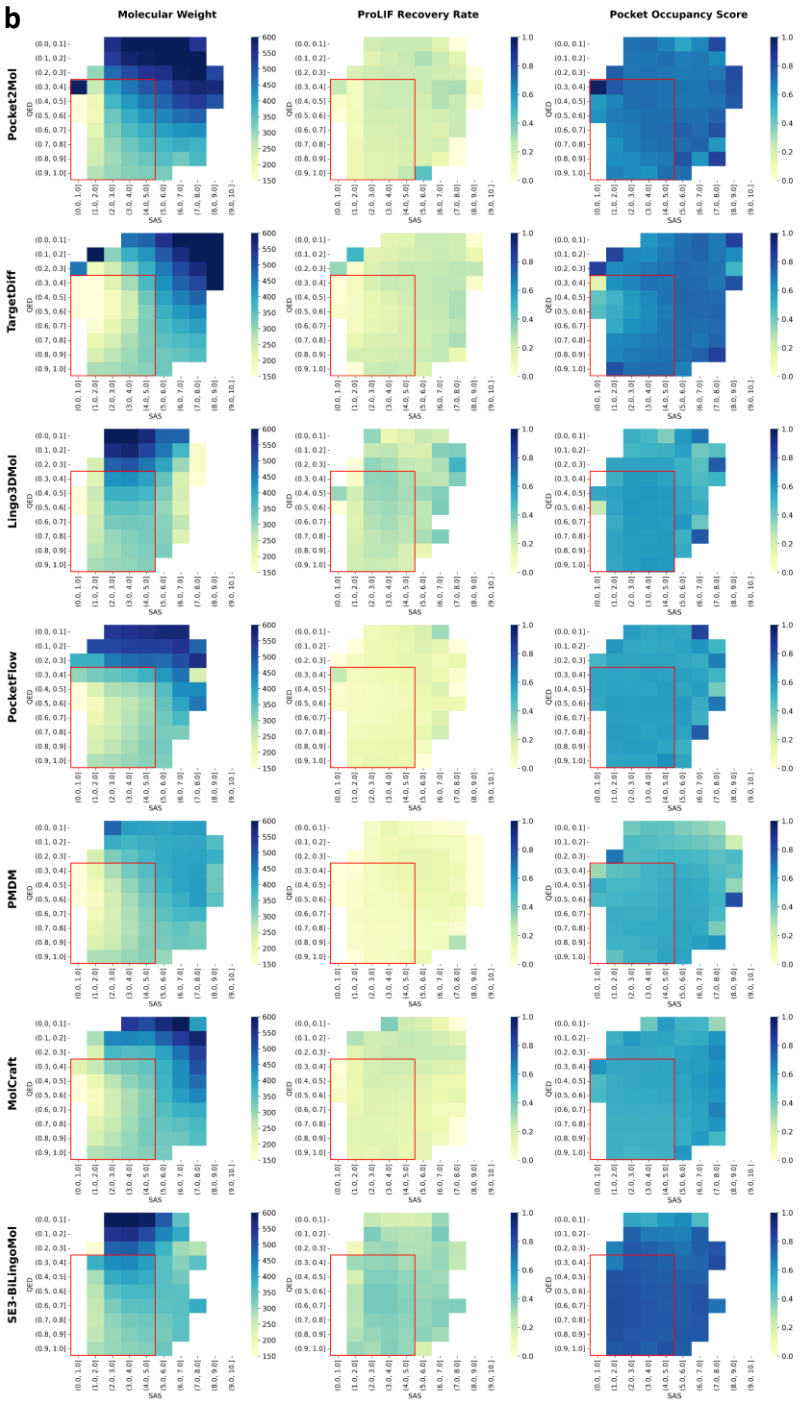
where $C_{i,r}$ is the count of type-i interaction that crystal ligand formed with residue r, and $P_{i,r}$ is the count of type-i interaction that a generated molecule formed with residue r. Here, we considered interactions including hydrogen bond acceptor/donor, halogen bond acceptor/donor, π-π stacking, cation-π, π-cation, anionic and cationic interactions were calculated by ProLIF.

(xv) **Mean Pocket occupancy.** The score represents the weighted average occupancy similarity between generated ligands and the co-crystal ligand within the binding pocket, evaluated across multiple targets. To compute this score, the solvent-accessible surface area (SASA) difference between the unbound pocket (defined as all residues within 6 Å from its cocrystal ligand) and the ligand-bound complex is calculated for both the co-crystal ligand and each generated ligand. The occupancy of a given ligand is then defined as the ratio of its SASA difference to that of the co-crystal ligand, with values exceeding 1 clipped to 1.

(xvi) **% ECFP_TS>0.5.** The ratio of number of targets, where at least one generated druglike molecule exhibits a Tanimoto similarity > 0.5 based on the ECFP4 fingerprints with that of any known active compound for that target, over 102 targets. For the active compounds, we only considered those with docking score < -5; for drug-like molecule candidates, those with QED > 0.3, SAS < 5 and min-in-place score < 0 are considered.

2.2 Property Distributions of Generated molecules across QED and SAS





Supplementary Information Figure S1. Property distributions of generated molecules across QED and SAS.This figure illustrates the distributions of various molecular properties for molecules generated across DUD-E 102 targets. A drug-like region, defined as QED > 0.3 and SAS < 5, is highlighted by a red box. The analysis encompasses around 1000 molecules generated per target across the 102 DUD-E targets for each model under evaluation.

(a) Distributions of number of generated molecules, min-in-place score, and redocking score.

(b) Distributions of molecular weight, ProLIF recovery rate, and pocket occupancy score.

3. HPK1 Novel Hit Scaffold Discovery

3.1 Screening Pipeline for Generated Molecules

To identify novel hit scaffolds, generated molecules underwent a multi-stage filtering pipeline followed by a clustering module. The filtering pipeline comprised several submodules, each with a pre-defined passing rate tailored to reflect specific screening objectives (as discussed subsequently). Filtering criteria were set to achieve these desired passing rates. The passing rate for each submodule, calculated as the proportion of all generated molecules (totaling approximately 1.2 million) satisfying its respective criterion, along with the associated criteria, are detailed below:

(i) **Binding Quality Measurement via Minimum-in-Place GlideSP Score.** The min-in-place docking score was calculated using the "mininplace" method within the Schrödinger Glide module. Molecules with a min-in-place score less than -8 were retained, resulting in a 10% passing rate.

(ii) **Evaluation of Ligand-Pocket Complementarity Using ExptGMS.** The electron density of a reference ligand-protein complex (PDB ID 7KAC) was utilized, and the complementarity score at 3.0 Å resolution (ed30_score) was predicted for each molecule. A threshold of ed30_score > 24,000 was applied to indicate favorable complementarity, leading to a 60% passing rate.

(iii) **Estimation of Occupancy within L23-G24-G25 Region.** For each generated ligand, the residue-level Solvent Accessible Surface Area (SASA) of L23, G24, and G25 in the bound protein state was calculated. These values were then compared to their unbound state to determine a SASA difference, indicative of each residue's contact with the ligand. Specifically, molecules exhibiting contact with L23 > 65 Å^2 and G24 > 8 Å^2 were considered to possess good sub-pocket occupancy, yielding a 40% passing rate.

(iv) **Conformational Validity Assessment via Torsional Energy Descriptor (TED).** The torsional energy of various dihedral angles was assessed for each generated ligand. A strict criterion was applied, retaining only molecules with a low torsional energy (<2 kcal/mol), which yielded a 40% passing rate.

(v) **Drug-Likeness Assessment (QED and SAS).** Molecules were filtered based on QED > 0.3 and SAS < 5, resulting in a 90% passing rate.

(vi) **Molecular Weight (MW) Compliance.** Molecules with molecular weights outside the range of 300 to 550 Da were excluded, leading to a 95% passing rate.

(vii) **Number of Rotatable Bonds < 6.** To promote higher molecular rigidity, the maximum allowed number of rotatable bonds was set to 5. This criterion resulted in a 60% passing rate.

The rationale behind the establishment of passing rates and the selection of specific thresholds for each filter was guided by the objective of novel scaffold discovery. At this initial stage, general drug-likeness metrics, including QED, SAS, and MW, were deliberately set to be less stringent, allowing over a 90% passing rate for each. This approach, aimed at maximizing the initial diversity of generated scaffolds, ensured that only a minor proportion of molecules were excluded based on these generalized criteria. In contrast, to identify novel scaffolds with a desired binding mode, more rigorous thresholds were applied for sub-pocket occupancy (L23-G24-G25) and pocket binding quality (minimum-in-place GlideSP score), achieving 40% and 10% passing rates, respectively. The torsional energy threshold was derived from prior work[9,10], yielding a 40% passing

rate. The number of rotatable bonds was optimized empirically: a setting exceeding 7 was observed to introduce undesirable aliphatic chains containing consecutive sp3 carbons. The ExptGMS threshold was set to exclude molecules with extreme overall shapes, thereby preventing an excessively high exclusion rate.

This filtering pipeline successfully identified 7,224 molecules, which were then clustered into 200 groups using a spectral clustering algorithm. Specifically, this algorithm operated on a similarity matrix derived from computing the maximum common substructure (MCS) of pairwise Bemis-Murcko (BM) scaffolds. Here, MCS of BM scaffolds were employed instead of Tanimoto similarity of entire molecules' ECFP4 fingerprints because the high diversity observed among molecules generated at this stage resulted in low similarity between most of the given pairs, which made the similarity matrix-based clustering inefficient for whole molecules. Finally, for each cluster, molecule with best (lowest) min-in-place GlideSP score was selected as the representative of this group.

3.2 Molecules Resulting from Screening Pipeline

The generated molecules selected from the screening pipeline are listed in the Supplementary Files within the Figshare repository (released_data → **RoundI_Scaffold_Discovery_Filtered_Representative_200_Molecules.sdf**)

4. HPK1 Lead Compound Optimization

4.1 Screening Pipeline for Generated Molecules

During our lead optimization stage, where drug-likeness and the mitigation of aromaticity were prioritized over novelty, the molecular filtering pipeline underwent several modifications. Specifically, an increased focus was placed on critical chemical properties, necessitating more stringent thresholds for QED, SAS, MW, and the number of rotatable bonds. Furthermore, the fraction of sp^3-hybridized carbons (Fsp3) and the number of aromatic rings were incorporated as additional filters to actively reduce the aromaticity of the compliant compounds. For binding mode filtering, the shared scaffold structure of regions A and C in the generated molecules rendered the general binding quality metrics, such as the min-in-place GlideSP score and ExptGMS, inefficient for evaluating subtle binding mode variations. Consequently, the thresholds for these two metrics were adjusted to reduce the exclusion rate and avoid discarding potentially valuable compounds. In contrast, because we particularly focused on the modifications in region B, the fulfillment of two critical NCIs observed in the Cmpd. 2-HPK1 crystal structure was rigorously monitored: a π-stacking interaction with Y28 and a hydrogen bond with K46.

The revised filtering pipeline comprised the following submodules. For each submodule, the passing rate was calculated as the proportion of all generated molecules (totaling approximately 0.2 million) satisfying its respective criterion:

(i) **Binding Quality Measurement via Minimum-in-Place GlideSP Score**: Molecules exhibiting a minimum-in-place GlideSP score of < -7 were considered, applying a relatively permissive criterion. This initial screening yielded a 90% passing rate.

(ii) **Evaluation of Ligand-Pocket Complementarity Using ExptGMS**: Instead of relying on PDB ID 7KAC, the electron density of the ligand-protein complex (PDB ID

9WD3) served as a reference for predicting the ed30_score for each molecule. An ed30_score > 18,450 filtering criteria was used to achieve a 95% passing rate.

(iii) **Fulfillment of Key NCIs (Y28 and K46)**: ProLIF was utilized to identify generated molecules forming either a π-stacking interaction with Y28 or a hydrogen bond with K46. This criterion yielded a 50% passing rate.

(iv) **Conformational Validity Assessment via TED**: Only molecules possessing valid torsional energy across all torsion dihedrals were considered, resulting in a 50% passing rate.

(v) **Mitigation of Aromaticity and Planarity**: To address potential cytotoxicity associated with excessive aromaticity and planarity observed in Cmpd. 4, metrics such as Fsp3 and the number of aromatic rings were integrated into the screening criteria. Molecules with Fsp3 > 0.35 were selected, referencing Fsp3 values of SW898 (0.19), Cmpd. 2 (0.27), and Cmpd. 4 (0.29). This filter resulted in an 80% passing rate.

(vi) **Number of Aromatic Rings < 4:** Molecules with fewer than 4 aromatic rings were considered, yielding a 90% passing rate.

(vii) **Number of Rotatable Bonds < 5**: To maintain relatively higher rigidity, the maximum allowed number of rotatable bonds was set to 4. This criterion resulted in a 70% passing rate.

(viii) **MW Compliance**: Molecules with molecular weights outside the range of 300 to 530 were excluded, with a 90% passing rate.

(ix) **Drug-Likeness Assessment (QED and SAS)**: More stringent criteria were applied for drug-likeness, specifically QED > 0.35 and SAS < 4.5, leading to a 50% passing rate.

Collectively, these filtering steps yielded a final set of 14,864 generated molecules. Subsequently, these molecules were grouped into 200 clusters using a spectral clustering algorithm. The input similarity matrix for clustering was computed based on the pairwise Tanimoto similarity of ECFP4 molecular fingerprints. For each cluster, the representative molecule was selected as the one exhibiting the highest Fsp3 value among its group members.

4.2 Molecules Resulting from Screening Pipeline

The generated molecules selected from the screening pipeline are listed in the Supplementary Files within the Figshare repository (released_data → **RoundII_Lead_Optimization_Filtered_Representative_200_Molecules.sdf**).

5. *In Vitro* and *In Vivo* Assays

5.1 HPK1 ADP-Glo Kinase Assay

HPK1 inhibitory activities were determined using the ADP-Glo kinase assay at 40 μM ATP. Initially, 8 μL of compounds were transferred to a 384-well LDV Echo plate. Subsequently, compounds were diluted and 50 nL aliquots were dispensed into a Greiner white assay plate using an Echo liquid handler, yielding 10 concentration points in duplicate (starting from a top concentration of 49.5 μM with 3-fold serial dilutions). The enzyme was prepared in assay buffer to a final concentration of 5 nM. A 2.5 μL aliquot of the enzyme mixture was added to each well of the assay plate. The plate was then spun down at 1000 rpm, centrifuged for 30 seconds, and pre-incubated for 15 minutes at 25°C.

A substrate mixture was prepared, and 2.5 μL was added to a Corning 4512 assay plate to initiate the reaction. The plate was then briefly centrifuged at 1000 rpm for 30 seconds to ensure all reagents were at the bottom of the wells and properly mixed, and then incubated at 25°C for 90 minutes. Following the 90-minute incubation, 5 μL of ADP-Glo™ Reagent was added. The plate was incubated at room temperature for 60 minutes, after which 10 μL of Kinase Detection Reagent was added. A final incubation at room temperature was performed for another 60 minutes. Finally, luminescence was recorded using an Envision plate reader. The 50% inhibitory concentration (IC_{50}) values were subsequently calculated using the log (inhibitor) vs. normalized response–variable slope model provided by GraphPad Prism.

5.2 HPK1 HTRF Kinase Assay

HPK1 inhibitory activities were determined using the HTRF Assay at 1 mM ATP. Initially, 10 μL of compounds were transferred to a 384-well LDV Echo plate. Subsequently, compounds were diluted, and 100 nL aliquots were dispensed into a Greiner white assay plate using an Echo liquid handler, yielding 10 concentration points in duplicate. The enzyme was prepared in assay buffer to a final concentration of 5 nM. A 5 μL aliquot of the enzyme mixture was added to each well of the assay plate. The plate was then spun down at 1000 rpm, centrifuged for 30 seconds, and pre-incubated for 15 minutes at 25°C. A substrate mixture was prepared, and 5 μL was added to a Corning 784075 assay plate to initiate the reaction. The plate was then briefly centrifuged at 1000 rpm for 30 seconds to ensure all reagents were at the bottom of the wells and properly mixed, and then incubated at 25°C for 16 hours. Following the 16-hour incubation, 10 μL of the XL665 and antibody detection reagent mixture was added. The plate was incubated for 120 minutes at room temperature. Finally, the TR-FRET signal (665/612 nm) was recorded using an Envision plate reader.

5.3 Kinase Selectivity Assays

Kinase activity inhibition was assessed using the KinaseProfiler (Eurofins Discovery, France) at 10 μM ATP. This platform utilizes a radiometric assay to measure the inhibition of kinase catalytic activity.

5.4 Jurkat pSLP76 Cellular Assay

Jurkat E6-1 cells were cultured overnight in RPMI 1640 medium supplemented with 0.5% Fetal Bovine Serum (FBS). Following overnight culture, 30 μL of cells (0.3×10^6 cells) were added per well to the assay plate. The test compounds were prepared via a three-fold serial dilution. Then, 7.5 μL of each diluted compound was transferred to the assay plate, resulting in 10 concentration points in duplicate. The plate was subsequently incubated for 4 hours in a 37°C incubator with 5% CO_2. After this initial incubation, 7.5 μL of 6× anti-human CD3 antibody was added to each well, and the plate was incubated for an additional 20 minutes in a 37 °C incubator with 5% CO_2. Finally, phosphorylated SLP-76 (S376) levels were determined using a Cisbio HTRF assay kit. Based on these results, the IC_{50} values of the compounds were calculated.

5.5 Human T-cell IL-2 Induction Assay

Human primary T cells were isolated from human peripheral blood mononuclear cells (PBMCs) using a human T cell isolation kit (Miltenyi Biotec, CAT: 130-096-535). Isolated T cells were plated in 96-well plates and pre-treated with varying concentrations of test compounds for 30 minutes. Subsequently, the cells were stimulated with 0.5 μg/mL anti-human CD3 antibody and 0.5 μg/mL anti-human CD28 antibody for 48 hours. Following the stimulation, cell culture supernatants were harvested and analyzed for IL-2 release using an ELISA kit.

5.6 Cell Viability Assay

HepG2 cells were cultured and assayed in EMEM medium supplemented with 10% FBS. After seeding into 96-well plates and incubating overnight in a 37 °C incubator with 5% CO_2, cells were treated with the indicated concentrations of test compounds for 72 hours. Cell viability was subsequently measured using the CellTiter-Glo® Luminescent (CTG) assay.

To assess the direct antiproliferative or cytotoxic effect of the compounds on CT26 tumor cell lines, a CTG assay was performed according to the manufacturer's instructions (Promega #G7573). CT26 cells were cultured in RPMI 1640 medium supplemented with 10% FBS, seeded into 384-well plates, and incubated overnight in a 37 °C incubator with 5% CO_2. Cells were subsequently treated with the indicated concentrations of test compounds for 72 hours, and cell viability was measured using the CTG assay.

5.7 CT26 Murine Syngeneic Model Efficacy Studies

All animal studies were conducted at ProOnco Therapeutics Co., Ltd, and were approved by the Institutional Animal Care and Use Committee (IACUC, permit: IACUC-ALM-01). CT26 tumor cells were maintained *in vitro* in RPMI 1640 medium supplemented with 10% FBS at 37°C in a 5% CO_2 atmosphere. For in vivo studies, BALB/c mice were subcutaneously inoculated in their right flank regions with 3×10^5 CT26 tumor cells suspended in 0.1 mL of phosphate-buffered saline (PBS). Drug treatments were initiated when tumor volumes reached approximately 50–60 mm^3. All enrolled tumor-bearing mice were randomized based on tumor size and body weight and then allocated to study groups. Test compounds were administered orally as a homogeneous suspension. Anti-mouse PD-1 (clone RMP1-14, Bio X Cell) was administered intraperitoneally (i.p.) as a solution. Study animals were monitored daily for signs of morbidity and mortality. Body weight and tumor volumes were measured every 2–3 days. Tumor volumes were determined in two dimensions using a caliper and expressed in mm^3 according to the formula: $V = (L \times W \times W) / 2$, where V represents tumor volume, L is the longest tumor dimension, and W is the longest tumor dimension perpendicular to L. The experiment was terminated with all individual animals euthanized when the mean tumor volume of the control group exceeded 2,000 mm^3.

5.8 Solubility Assay

Kinetic solubility was measured in various standard pH buffers (pH 6.5, pH 7.4).

Initially, 2 µL of a DMSO stock solution containing the test compound was added to 198 µL of the respective buffer, followed by thorough mixing. The samples were then shaken for 4 hours in an Eppendorf ThermoMixer at 350 rpm and 25°C. Subsequently, samples were filtered through a filter plate. The resulting filtrate was diluted with DMSO according to the established dilution scheme (details to be provided in supplementary methods if complex). Finally, 5 µL of the diluted sample solutions and calibration standard solutions were spiked into 300 µL of quenching solution, mixed thoroughly, and prepared for subsequent LC/MS analysis.

5.9 Caco-2 Permeability

The bi-directional apparent permeability of test compound was determined using a Caco-2 monolayer cell model. Caco-2 cells (50 µL of cell suspension at a density of 6.86 × 10^5 cells/mL) were seeded onto Transwell inserts. These cells were cultured for 14–18 days to form a confluent cell monolayer. The test compound was subsequently added to either the apical or basolateral side of the cell monolayer. The system was then incubated at 37°C for up to 2 hours to facilitate bi-directional transport. Samples were processed and analyzed by LC-MS/MS. Finally, the bi-directional apparent permeability, efflux ratio, and system recovery were calculated.

5.10 Hepatocyte Stability Assay

The metabolic stability of test compound was assessed using hepatocyte suspensions from mouse, rat, dog, and human. Hepatocyte suspensions from the four species, each containing 1 µM of test compound, were incubated in an orbital shaker at 37°C for up to 120 minutes. Cell aliquots of 25 µL were removed at time points of 0, 15, 30, 60, 90, and 120 minutes. To terminate the metabolic reaction, these aliquots were immediately mixed with 6 volumes (150 µL) of acetonitrile containing internal standards (i.e. 100 nM alprazolam, 200 nM labetalol, 200 nM caffeine, and 2 µM ketoprofen). The plate was then centrifuged for 20 minutes at 3,220 × g. A 100 µL aliquot of the resulting supernatant was mixed with 100 µL of ultra-pure H_2O and subsequently used for LC-MS/MS analysis.

5.11 Plasma Protein Binding

The plasma protein binding of test compound in plasma from mice, rats, dogs, and humans was determined by equilibrium dialysis. Plasma samples from each species, containing 1 µM of test compound and the control compound ketoconazole, were added to a HTD 96-well dialysis device. The device was then incubated at 37°C with 5% CO_2 at 100 rpm for 6 hours. Following incubation, the samples were processed, and the concentrations of test compound were determined by LC-MS/MS. This data was then used to calculate the bound fraction, fraction unbound, and system recovery.

5.12 PXR Activation in DPX2 Cells

DPX2 cells were seeded into 384-well plates at a density of 1 × 10^4 cells/well and incubated for 24 hours at 37°C with 5% CO_2. Following incubation, the plates were removed from the incubator. Subsequently, 25 nL of the negative control, inducers, test compound (at concentrations of 1 µM and 10 µM), and the positive control rifampicin (10

μM) were added. The plates were then returned to the incubator for an additional 48 hours. Cell viability was determined using the CellTiter-Fluor Assay reagent, with fluorescent excitation and emission wavelengths of 400 nm and 505 nm, respectively. Subsequently, PXR activation was investigated by measuring luciferase activity with ONE-Glo reagent (Promega). PXR activation was expressed as the ratio of the average relative light units (RLU) from triplicate determinations at each dose of test compound to the average RLU from triplicate determinations obtained with the corresponding vehicle control.

5.13 CYP Inhibition Assay

The direct inhibition of test compound on five isoforms of cytochrome P450 (CYP1A2, CYP2C9, CYP2C19, CYP2D6, and CYP3A) was evaluated using human liver microsomes. Specific probe substrates for each CYP isoform (CYP1A2: 0.5 μM α-naphthoflavone; CYP2C9: 10 μM sulfaphenazole; CYP2C19: 1 μM (+)-N-3-Benzylnirvanol; CYP2D6: 0.5 μM quinidine; and CYP3A: 0.5 μM ketoconazole) were mixed with 0.2 mg/mL human liver microsomes, containing 10 μM of test compound, for each respective isoform inhibition assay. Following a pre-incubation in a water bath for 8 minutes at 37°C, NADPH solution was added to all incubation wells, and incubation continued at 37°C. The reaction was terminated by the addition of 400 μL of methanol containing internal standards (i.e. 100 nM alprazolam, 500 nM labetalol, and 2 μM ketoprofen) at the designated time point (Panel 1: 5 minutes). Samples were then centrifuged at 3,220 × g for 40 minutes to precipitate proteins. Subsequently, 100 μL of the supernatant was transferred to a new 96-well plate and mixed with 100 μL of water for LC-MS/MS analysis.

5.14 Time-dependent Inhibition (TDI) Assay

The assay was performed in potassium phosphate buffer. Verapamil, a known CYP3A4/5 inhibitor, was utilized as a positive control to validate the incubation system. Test compound was diluted to achieve six concentration points, starting from a top concentration of 50 μM with 4-fold serial dilutions. These dilutions were then incubated with midazolam (a substrate for CYP3A4/5) in human liver microsomes. The reaction was initiated by the addition of the coenzyme NADPH and incubated at 37°C for 30 minutes. To terminate the reaction, ice-cold acetonitrile containing an internal standard was added to the incubation system. The mixture was then vortexed for 1 minute and centrifuged at 4,000 rpm for 15 minutes at 4°C. A 200 μL aliquot of the resulting supernatant was then transferred and thoroughly mixed with 50 μL of water. These samples were subsequently centrifuged again at 4,000 rpm for 5 minutes at 4°C prior to LC-MS/MS analysis.

5.15 hERG Inhibition

HEK 293-hERG cells (Invitrogen, Cat. K1236) were cultured in DMEM-based medium (85% DMEM, 10% dialyzed FBS, 0.1 mM NEAA, 25 mM HEPES, 100 U/mL Pen-Strep, 5 μg/mL Blasticidin, 400 μg/mL Geneticin) at 37°C, 5% CO_2. hERG expression was induced with 1 μg/mL doxycycline for 48 hours. For assay, cells (approx. 5×10^5 cells/3.5 cm dish) were plated on coverslips in Blasticidin/Geneticin-free medium. Whole-cell patch-clamp recordings utilized a PatchMaster system (HEKA Elektronik). A giga-ohm (GΩ) seal and whole-cell configuration were established via suction, with capacity currents compensated (Cfast/Cslow). Membrane potential was -60 mV. Holding potential:

-90 mV for 500 ms (20 kHz recording, 10 kHz filter). Leak current: -80 mV for 500 ms. hERG current was elicited by +30 mV for 4.8 s, then -50 mV for 5.2 s; tail current amplitude was measured. Current stability was confirmed for 120 s. After 5 min vehicle control baseline, test compound was applied. Steady-state hERG current was recorded for ~5 min (5 sweeps). Five cumulative concentrations of compound were tested for dose-response. Dofetilide (150 nM) served as an internal low control post-highest test compound. System validation included a full dofetilide dose-response (5 concentrations) on an identical cell batch.

5.16 Animal Pharmacokinetic (PK) Studies

For *in vivo* pharmacokinetic experiments, the test compound was administered to male CD1 mice and male Sprague Dawley rats either intravenously or via oral gavage. For intravenous (IV) dosing, the test compound was typically administered at a dose of 1 mg/kg as an IV bolus to both mice and rats. For oral (PO) dosing, the test compound was typically administered at a dose of 3.0 mg/kg. Blood samples were collected at pre-dose and at various specified time points up to 24 hours post-dose. All blood samples were collected using EDTA as the anticoagulant, and subsequently centrifuged to obtain plasma samples. The plasma concentrations of the test compound were determined using established LC-MS/MS methods. The measured plasma concentrations were then used to calculate pharmacokinetic (PK) parameters by applying standard noncompartmental analysis methods, utilizing the Phoenix® WinNonlin software program (version 8.0, Pharsight Corporation).

Reference

1 Feng, W. *et al.* Generation of 3D molecules in pockets via a language model. **6**, 62-73 (2024).

2 Brehmer, J., De Haan, P., Behrends, S. & Cohen, T. S. J. A. i. N. I. P. S. Geometric algebra transformer. **36**, 35472-35496 (2023).

3 Shazeer, N. J. a. p. a. Fast transformer decoding: One write-head is all you need. (2019).

4 Fan, F. *et al.* Assessing Conformation Validity and Rationality of Deep Learning-Generated 3D Molecules. 2024.2011. 2010.622844 (2024).

5 Harris, C. *et al.* Posecheck: Generative models for 3d structure-based drug design produce unrealistic poses. (2023).

6 Friesner, R. A. *et al.* Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. **47**, 1739-1749 (2004).

7 Lin, H. *et al.* CBGBench: fill in the blank of protein-molecule complex binding graph. (2024).

8 Errington, D., Schneider, C., Bouysset, C. & Dreyer, F. A. J. J. o. C. Assessing interaction recovery of predicted protein-ligand poses: D. Errington et al. **17**, 76 (2025).

9 Fan, F. *et al.* Assessing Conformation Validity and Rationality of Deep Learning-Generated 3D Molecules. 2024.2011.2010.622844, doi:10.1101/2024.11.10.622844 %J bioRxiv (2024).

10 Rai, B. K. *et al.* TorsionNet: A Deep Neural Network to Rapidly Predict Small-Molecule Torsional Energy Profiles with the Accuracy of Quantum Mechanics. *Journal of Chemical Information and Modeling* **62**, 785-800, doi:10.1021/acs.jcim.1c01346 (2022).