Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

TETGen: A Transcriptome-Guided Transformer for Targeted Molecule Generation

Jiahui Gao
Anhui Medical University
Yanyun Cao
Anhui Medical University
Rui Zhou
Anhui Medical University
Yinzi Li
Anhui Medical University
Jiale Wang
Anhui Medical University
Wanyue Ding
Anhui Medical University
Mingzhi Ma
Anhui Medical University
Danchen Wu
Anhui Medical University
Yan Zhou
Anhui Medical University
Bin Yuan
Anhui Medical University
Qichao Luo
luoqichao@ahmu.edu.cn

Anhui Medical University

Research Article

Keywords: small molecule generation, transcriptomic data, protein-protein interaction, Transformer model, drug design

Posted Date: March 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9234106/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

TETGen: A Transcriptome-Guided Transformer for Targeted Molecule Generation

Jiahui Gao[a,1], Yanyun Cao[b,d,1], Rui Zhou[a,1], Yingzi Li[c], Jiale Wang[c], Wanyue Ding[a], Mingzhi Ma[b], Danchen Wu[a], Yan Zhou[b,d*], Bin Yuan[a,*], Qichao Luo[a,c,*]

[a] School of Pharmacy, Anhui Medical University, Hefei 230032, Anhui, China

[b]The First Affiliated Hospital of Anhui Medical University, Hefei 230032, Anhui, China.

[c]School of Biomedical Engineering, Anhui Medical University, Hefei 230032, Anhui, China

[d]Anhui Public Health Clinical Center, Anhui Medical University, Hefei 230032, Anhui, China

[1]These authors contributed equally to this work (co-first authors).

[*]Corresponding authors. E-mails: Qichao Luo (luoqichao@ahmu.edu.cn); Bin Yuan (yuanbin@ahmu.edu.cn); Yan Zhou (yfyb33719021@fy.ahmu.edu.cn).

Abstract

While artificial intelligence has advanced de novo molecule generation, effectively leveraging transcriptomic data to guide biologically meaningful molecular design remains underexplored. Here we present TETGen (Transcriptome-Enhanced Transformer Generator), a Transformer-based model that integrates transcriptomic profiles to generate molecules with targeted biological activities. TETGen incorporates biological context through gene vectors derived from protein-protein interaction network random walks, which serve as functional positional encodings. The model further supervises its cross-attention mechanism using transcriptomic profiles during training to ensure biological relevance. TETGen efficiently generates compounds with favorable drug-like properties and synthetic accessibility. Evaluation demonstrates that molecules with transcriptomic matching scores exceeding 0.6 exhibit significantly higher structural similarity to known bioactive compounds compared to those with lower scores. In a case study using NEK7-knockout transcriptomic data, TETGen generated candidate molecular glues whose stable binding modes were corroborated by docking and 100 ns molecular dynamics simulations, with the lead candidate maintaining persistent hydrogen bonding with NEK7 Glu37 and π-π stacking with CRBN Trp386. This work presents an interpretable

generative framework that effectively bridges chemical and biological spaces, offering a computational tool for phenotype-inspired drug design.

Keywords: small molecule generation, transcriptomic data, protein-protein interaction, Transformer model, drug design

Abbreviations

AI: Artificial intelligence ;BTK: Bruton's tyrosine kinase; CLL: Chronic lymphocytic leukemia; CMAP: Connectivity Map;CRBN: Cereblon ; Diffusion Models: Diffusion-based generative models; GPT: Generative pre-trained transformer; GROMACS: Groningen machine for chemical simulations; IntDiv: Internal diversity; LINCS: Library of Integrated Network-Based Cellular Signatures; L1000: Level 1000 assay (landmark gene expression profiling); MACCS: Molecular Access System; MD: Molecular dynamics; MOA: Mechanism of action; NEK7: NIMA-related kinase 7; NF-κB: Nuclear factor kappa-B ; NLRP3: NOD-like receptor family pyrin domain containing 3 ; PDB: Protein Data Bank; PME: Particle mesh Ewald ; PPI: Protein-protein interaction; QED: Quantitative Estimate of Drug-likeness; QSAR: Quantitative structure-activity relationship; RMSD: Root-mean-square deviation; RNN: Recurrent neural network ; SA: Synthetic accessibility; SI: Supplementary Information ; shRNA: Short hairpin RNA; SMILES: Simplified molecular-input line-entry system; TETGen: Transcriptome-Enhanced Transformer Generator; TransGEM: Transformer-based model from gene expression to molecules; Transformers: Transformer-based neural networks; VAE: Variational autoencoder; XlogP: Predicted partition coefficient

1 Introduction

Drug discovery is a time-consuming process with a low success rate, posing substantial economic and technical challenges. Various computational methods have been developed to accelerate the identification of early-stage lead compounds[1–5]. Traditional computer-aided drug design approaches are primarily divided into two major categories. Ligand-based methods, including QSAR modeling and pharmacophore mapping, operate on the principle of molecular similarity, presuming that similar structures exhibit similar functions. This approach, however, is fundamentally constrained by its reliance on already known bioactive chemical entities[6-7]. Conversely, structure-based techniques rely on the three-dimensional structure of the target protein. Methods such as molecular docking and virtual screening use this structural information to simulate ligand binding within a specific pocket. While powerful, this approach is

critically dependent on the availability and accuracy of the target protein structure with a well-defined druggable pocket[8-9]. Crucially, both approaches are inherently limited in complex therapeutic contexts. A key bottleneck is the rational design of multi-target interventions, where a single molecule must simultaneously engage multiple targets to produce a desired phenotype[10-12]. To bridge this gap, new computational frameworks are needed, ones that integrate holistic biological information like genomic and transcriptomic data to directly connect molecular design with complex biological outcomes.

In recent years, with the rapid advancement of deep learning technology, deep generative models have been widely introduced into the field of drug discovery, offering innovative solutions for the automated and efficient design of new molecules with specific properties[13-19]. These models are not confined to searching and optimizing within known chemical spaces. Instead, they can learn implicit chemical rules and structural features from vast molecular datasets, enabling the autonomous generation of novel structural compounds that meet specific requirements. Among various generative models, neural network architectures such as Recurrent Neural Networks[20-21], Variational Autoencoders[22-24], Generative Adversarial Networks[25-26], Transformers[27-28], and Diffusion Models[29-30] are particularly well-suited for molecular generation tasks. They capture the complex distribution of molecular structures through mechanisms like encoder-decoder frameworks, adversarial training, or probabilistic diffusion processes, achieving controllable mapping from continuous latent spaces or noise vectors to concrete molecular structures. The advantage of generative models lies in their ability to learn the intrinsic relationships between chemical and biological properties from real molecular training sets and automatically synthesize new entities with similar characteristics. This capability not only significantly enhances molecular design efficiency but also greatly expands the explorable bioactive chemical space.

Transcriptomic data provide a snapshot of cellular states by capturing molecular-level responses to disease, genetic perturbations, or drug treatments. This makes them a valuable resource for identifying compounds that can influence complex biological phenotypes, including those requiring multi-target interventions. To harness this potential, several deep generative models, such as Gex2SGen[31], BiCEV[32], and TransGEM[33], have been developed to connect transcriptomic signatures with molecular generation. BiCEV, for instance, enables transcriptomically guided molecular design through joint training of a variational autoencoder (VAE) for chemical structures and another VAE for the transcriptomic changes induced by molecular perturbations. Gex2SGen follows a different strategy, first training a gene expression VAE and a molecular VAE independently and then combining the pretrained encoder from the former with the decoder from the latter to form a unified generative model. TransGEM (Transformer-based model

from gene expression to molecules), on the other hand, integrates a gene expression encoder with a Transformer-based decoder tailored for molecular generation. Despite these advances, existing approaches often underutilize the functional relationships and biological context inherent in transcriptomic data. Moreover, they typically function as black boxes, lacking interpretable and quantifiable metrics to assess whether, and to what extent, the generated molecules correspond to the input transcriptomic profiles. This gap not only complicates reliable evaluation but also raises concerns about the trustworthiness of the models themselves.

To overcome the limitations discussed above, we introduce TETGen, a Transcriptome-Enhanced Transformer Generator specifically designed to incorporate transcriptomic information in a more biologically meaningful and interpretable manner. Rather than treating genes as independent features, TETGen first derives gene representations from protein-protein interaction networks via random walks. These functionally enriched vectors serve as positional encodings, enabling the model to capture underlying biological relationships from the outset. In parallel, we guide the cross-attention mechanism through a supervision strategy using transcriptomic profiles to ensure that generation remains grounded in the intended biological context. The resulting framework consistently produces molecules with favorable drug-like properties and synthetic accessibility. Notably, we find that a transcriptomic matching score above 0.6 correlates strongly with structural resemblance to known bioactive compounds, providing a quantitative indicator of biological relevance. To test its practical utility, we applied TETGen to transcriptomic data from NEK7-knockout experiments. The model successfully generated a series of candidate molecules, from which one potential NEK7 molecular glue was identified through docking-based screening. Subsequent molecular dynamics simulations confirmed that this compound adopts a stable binding mode with the NEK7–CRBN complex. By making the connection between transcriptional states and chemical structures both explicit and measurable, this work offers an interpretable framework that bridges chemical and biological spaces, representing a step toward more biologically aware, AI-driven drug discovery.

2 Methods

Gene Expression Data Set. LINCS provides publicly accessible, high-throughput gene expression profiles under various perturbations, including small molecules and shRNAs. Using the L1000 assay, expression can be measured for 978 landmark genes[34]. In this study, we employed the latest CMAP LINCS 2020 dataset and utilized Level 5 data, which contain real measured expression values for the landmark genes. The Python library cmapPy was used to extract

Level 5 signatures from GCTx files. To filter the perturbation profiles induced by compounds, we retained only those generated under the most common experimental conditions: a duration of 24 h and a dosage concentration of 10 μM. Additionally, moderated z-score summarization was applied to ensure that only one profile per compound per cell line was included. The four most frequently profiled cell lines (MCF7, PC3, HT29, A549) were selected. Profiles from the MCF7 cell line were used for model training, with an 8:2 split into training and validation sets. Profiles from the remaining cell lines served as an external validation set to assess the model's generalization capability. For shRNA-induced expression profiles, identical processing procedures were applied. We selected those profiles obtained in the MCF7 cell line after 96 h to evaluate the model's capability in generating molecules targeting NEK7 and other specified targets.

Molecule Data Sets. In this work, we obtained a dataset of 1.9 million drug-like molecules from the ZINC database[27], which had been curated with strict filtering criteria including molecular weight between 250 and 350 Da, fewer than seven rotatable bonds, and an XlogP below 3.5. The SMILES strings of these molecules were tokenized using a vocabulary of 38 tokens, with each string encoded as a fixed-length sequence of 150 tokens. Sequences shorter than 150 tokens were zero-padded to the required length, while those exceeding 150 tokens were truncated to retain only the first 150 characters. This tokenized dataset was then used to pre-train our chemical language model, a Generative Pre-trained Transformer (GPT) tailored for molecular generation. The dataset was randomly split into training and validation sets in an 8:2 ratio.

Model Configuration. We propose a Transformer-based generative model that de novo generates molecular SMILES strings conditioned on input differential expression profiles. The architecture follows the standard Transformer framework, consisting of an encoder for processing differential expression profiles and a decoder for generating SMILES strings, linked by cross-attention layers. Model building proceeds in two stages. First, we pre-train the decoder as a chemical language model—a Generative Pre-trained Transformer (GPT)—on the curated dataset of 1.9 million drug-like molecules to capture the syntactic and structural regularities of SMILES. The best pre-trained GPT weights are saved and later used to initialize the decoder in the full model. In the second stage, the pre-trained GPT decoder is connected to the encoder via cross-attention, and the entire model is fine-tuned to align generation with the input transcriptomic profiles. To ensure that the model attends to significantly altered genes during molecule generation, we directly supervise the cross-attention weights using the transcriptomic data. All models were implemented in PyTorch. Training was performed on a machine equipped with 24 vCPUs, 125 GB of RAM, and a

single NVIDIA A40 GPU (40 GB memory). Our model is illustrated in **Figure 1**, further details on the model are provided below.

GPT Layer. The SMILES string of each molecule was first tokenized into a sequence of 150 tokens. These tokenized integers were then passed through an embedding layer, which maps discrete token indices into continuous low-dimensional feature vectors of size 256, serving as the input representations for the model. To incorporate sequential information, positional embeddings were added to the token embeddings to retain the order of tokens within each sequence. The resulting embedded representations were fed into a stack of Transformer decoder blocks. Each decoder block consists of two core components, namely a masked multi-head self-attention layer and a fully connected feed-forward network. The masked self-attention mechanism ensures that during training, the prediction for a given token depends only on preceding tokens and not on future ones, mirroring the autoregressive generation process where only previously generated tokens are accessible. This masking prevents information leakage and enables the model to learn proper causal dependencies in molecular SMILES strings. Each self-attention layer employs eight attention heads, allowing the model to jointly attend to information from different representation subspaces. Following the self-attention layer, the output is passed through a feed-forward network comprising linear transformation with a GELU activation in between. The hidden layer dimension is 1024, and the output layer projects back to a vector of size 256, which serves as input to the subsequent decoder block. The GPT module is composed of eight such stacked decoder blocks. Hyperparameters, including the number of attention heads and the number of decoder blocks, were optimized through systematic experimentation, with detailed configurations provided in **Table S1**. The final output from the last decoder block is used to predict the next token in the SMILES sequence, enabling step-by-step molecule generation.

Gene Expression Profiles Encoder Layer. Following the embedding approach used in the TransGEM model, we first transformed each of the 978 gene expression values into continuous feature vectors of size 256, yielding a 978×256 matrix as the initial input representation. To preserve functional relationships among genes, we retrieved protein–protein interaction networks from the STRING database[35] and performed random walks to obtain functionally enriched embeddings for each protein. These embeddings were then incorporated as positional encodings for the corresponding genes, providing biological context that reflects their functional interdependencies. The resulting embedded representations were fed into a stack of Transformer encoder blocks. Each encoder block consists of three core components: a masked multi-head self-attention layer, a residual connection layer, and a fully connected feed-forward network. Following the self-attention layer, the output is passed through a residual connection that adds the

original input back to the layer output, helping to mitigate gradient vanishing and facilitate training. The output then goes through a feed-forward network comprising a linear transformation with a GELU activation in between. The hidden layer dimension is 1024, and the output layer projects back to a vector of size 256, which serves as input to the subsequent encoder block. The encoder module is composed of six such stacked encoder blocks.

Loss Function. The training objective of TETGen comprises two components: a reconstruction loss that ensures valid molecular generation, and an attention supervision loss that aligns the generation process with biologically relevant genes. The reconstruction loss is a standard cross-entropy loss computed over the generated SMILES sequences. At each decoding step, the model predicts the probability distribution over the vocabulary for the next token, and the loss minimizes the negative log-likelihood of the ground-truth token given the preceding context:

$$L_{recon} = -\sum_{t=1}^{T} \log p(y_t | y_{<t}, X)$$

where y_t represents the token at position t. $y_{<t}$ denotes all preceding tokens. And X is the input gene expression profile.

The attention supervision loss consists of two terms: an alignment loss and a sparsity loss. The alignment loss encourages the cross-attention weights to focus on genes exhibiting significant expression changes. To derive the target attention distribution, we first compute the absolute values of the input gene expression changes, then apply a softmax normalization with a temperature parameter to obtain probabilities p_i for each gene i. The model's attention distribution q_i is obtained by similarly normalizing the attention weights aggregated across heads and decoding steps. The alignment loss is then defined as a weighted cross-entropy between the target distribution and the model's attention distribution, where the weighting factor $1+p_i$ amplifies the supervision signal for genes with higher expression changes:

$$L_{align} = -\sum_{i=1}^{978} p_i \cdot (1 + p_i) \cdot \log q_i$$

This formulation ensures that genes with larger expression perturbations exert stronger influence on the attention learning process, guiding the model to focus on biologically meaningful signals.

The sparsity loss prevents the attention weights from being uniformly distributed across all genes, encouraging the model to concentrate on a compact set of relevant genes. This is implemented as an entropy penalty on the attention distribution:

$$L_{sparse} = -\sum_{i=1}^{978} q_i \log q_i$$

Minimizing this entropy term sharpens the attention distribution and helps mitigate overfitting by reducing reliance on noisy or irrelevant features.

The total loss is a weighted combination of these three terms:

$$L_{total} = L_{recon} + \lambda_1 L_{align} + \lambda_2 L_{sparse}$$

where λ_1 and λ_2 are hyperparameters controlling the contribution of each supervision term. In our experiments, we set $\lambda_1 = 0.15$ and $\lambda_2 = 0.015$

Evaluation Metrics. To comprehensively assess the quality of generated molecules, we employed four standard metrics. Validity measures the fraction of chemically valid molecules among all generated samples. A generated SMILES string is considered invalid if it contains syntactic or structural defects such as unbalanced parentheses or unreasonable valences. Novelty computes the proportion of generated molecules that are absent from the training set. Uniqueness quantifies the fraction of distinct and valid molecules within the generated set. Internal diversity (IntDiv) measures the structural heterogeneity of the generated molecules.

To evaluate the biological relevance of generated molecules with respect to the input transcriptomic profiles, we defined a gene mapping score. Let $\alpha \in R^{978}$ be the aggregated attention weights, obtained by averaging over decoding steps and attention heads, then taking the maximum over sequence positions for each gene. Let $e \in R^{978}$ be the input expression perturbation vector, with |e| representing the magnitude of change. Both vectors are converted to probability distributions using a temperature softmax (T=0.07):

$$p_q = softmax(\alpha|T), p_e = softmax(\alpha|s|\alpha)$$

The Spearman rank correlation ρ between p_a and p_e is computed and linearly mapped to [0,1]:

$$\text{Gene mapping score} = \frac{\rho + 1}{2}$$

A higher Gene mapping score indicates stronger concordance between the genes attended to during generation and those most perturbed in the input transcriptome, reflecting greater biological relevance of the generated molecule.

Molecular Docking and Dynamics Simulations. To assess whether TETGen can generate potential BTK inhibitors, we performed molecular docking using the online CB-Dock2 server[36]. Candidate molecules generated from the transcriptomic profiles of dasatinib-treated CLL patient lymphocytes were docked into the crystal structure

of the BTK/dasatinib complex (PDB: 3K54). The docking results were visualized with PyMOL to evaluate binding modes and identify promising candidates. For NEK7 molecular glue discovery, molecules generated from NEK7-knockout MCF7 transcriptomic data were docked into the known molecular glue binding pocket within the NEK7–CRBN ternary complex structure (PDB:9NFQ). To further validate binding stability, we conducted 100 ns molecular dynamics simulations using GROMACS. The simulation system was prepared with the CHARMM-GUI web server[37]. The CHARMM36m force field was applied, and the complex was solvated in a periodic rectangular box (140 Å × 140 Å × 140 Å) filled with TIP3P water molecules. The system was neutralized and supplemented with 0.15 M NaCl to mimic physiological conditions. Energy minimization was performed using the steepest descent algorithm to eliminate unfavorable contacts. The system then underwent a 35 ns equilibration phase under NPT conditions (300 K, 1.0 bar), with temperature controlled by the v-rescale thermostat and pressure maintained by the C-rescale barostat with isotropic coupling. During the subsequent 100 ns production run, bonds involving hydrogen atoms were constrained using the LINCS algorithm. Non-bonded interactions were handled with the Verlet cutoff scheme, updating the neighbor list every 20 steps. Van der Waals interactions employed a force-switch modifier (rvdw-switch =1.0 nm, rvdw = 1.2 nm), and Coulomb interactions were truncated at 1.2 nm with long-range electrostatics computed via the Particle Mesh Ewald (PME) method. Trajectory analysis focused on ligand RMSD, hydrogen bond occupancy, and contact frequency between NEK7 and CRBN to evaluate the compound's ability to stabilize the protein–protein interface.

3 Results

Model Performance Evaluation and Comparative Analysis

To systematically evaluate model performance, we selected three representative baseline models, namely BiCEV, Gex2SGen and TransGEM, for comparative analysis. BiCEV and Gex2SGen are based on variational autoencoder architectures, while TransGEM shares the Transformer architecture with our proposed TETGen. Evaluation results on a unified validation set demonstrate that all models generate valid molecular structures with high success rates **(Figure 2a)**, which can be attributed to pre-training on the large-scale ZINC compound library that enables learning of rich chemical space representations. Notably, TETGen outperforms the baseline models in terms of internal diversity of generated molecules, indicating enhanced capability to explore chemical space and generate structurally diverse compounds. Physicochemical property analysis reveals that molecules generated by all models exhibit molecular

weights concentrated around 500 Da and logP values predominantly distributed near 5 (**Figure 2c**), consistent with fundamental drug-like characteristics. However, on metrics designed to evaluate drug-likeness and synthetic feasibility, such as QED and SA scores, TETGen, while outperforming TransGEM, falls slightly short of BiCEV and Gex2SGen. These results suggest that although our model demonstrates advantages in structural diversity and chemical space coverage, further improvements are needed in precisely controlling optimization objectives such as drug-likeness and synthetic accessibility.

The principal improvements of TETGen over TransGEM are twofold. First, TETGen introduces protein feature vectors derived from protein-protein interaction network random walks into the encoding layer, constructing gene-specific, function-aware positional encodings. Ablation experiments show that removing these functional encoding vectors leads to significant degradation in the validity, uniqueness, QED and SA Score of generated molecules (**Figure 2a, 2c**), confirming that functional association information embedded in PPI networks is essential for guiding the model toward generating molecules with biological relevance. Second, to enhance the biological correlation between generated molecules and target transcriptomic profiles, TETGen directly supervises the cross-attention mechanism using transcriptomic data during training. As shown in **Figure 2b**, compared to the unsupervised version and TransGEM, this transcriptome-guided attention supervision substantially improves the alignment between attention weight distributions and ground-truth transcriptomic profiles, thereby strengthening consistency between generated outputs and the target biological context.

Evaluation of Model Generalization across Different Cell Lines

To assess the generalization capability of TETGen, we extended its application to three biologically distinct cell lines: PC3 (prostate cancer), A549 (lung cancer), and HT29 (colorectal cancer). For each cell line, 500 representative transcriptomic profiles were randomly selected to generate molecules under identical conditions. Evaluation results showed that all models maintained stable performance in overall molecular quality and key physicochemical properties across different cell lines. TETGen consistently outperformed baseline models in internal diversity across PC3, A549, and HT29 cell lines (**Figure 3a**), indicating that it generates more structurally diverse molecules when conditioned on different cellular contexts. When directly compared with TransGEM, TETGen exhibited higher quantitative estimate of drug-likeness and synthetic accessibility scores across all tested cell lines (**Figure 3c**).

Moreover, when the TETGen model trained exclusively on MCF7 cells was directly applied to transcriptomic data from PC3, A549, and HT29 cells without fine-tuning, its cross-attention mechanism still maintained alignment with biological context. As shown in **Figure 3b**, the cross-attention weight distributions correlated with ground-truth transcriptomic profiles across these unseen cell lines. This correlation indicates that TETGen's attention mechanism responds to gene expression patterns in the input, rather than relying on features memorized from the training set. The sustained alignment between attention weights and transcriptomic profiles across different cell lines suggests that the biological context representations learned by TETGen are generalizable, supporting consistent generation performance when applied to cell types not encountered during training.

Cross-Modal Alignment between Chemical Structure and Transcriptomic Profiles

To investigate whether TETGen achieves cross-modal alignment between chemical structures and transcriptomic profiles, we conducted systematic analyses guided by the similarity principle that molecules inducing similar gene expression patterns often share structural frameworks. **Figure 4a** shows that distributions of structural similarity between generated and known bioactive compounds remained highly consistent regardless of transcriptomic supervision, confirming that large-scale pre-training provides a foundation for cross-modal generation. Further analysis revealed that transcriptomic supervision substantially enhances alignment. As shown in **Figure 4b,** molecules with transcriptomic matching scores exceeding 0.6 exhibited significantly higher structural similarity to known bioactive compounds compared to those with scores below 0.4, indicating that higher transcriptomic alignment guides the model toward generating molecules structurally closer to known active compounds. Examination of individual generation cases corroborated this mechanism **(Figure 4c-h).** In four out of five cases where transcriptomic matching scores exceeded 0.6, TETGen generated molecules with high MACSS similarity to their corresponding reference compounds. These case-level observations further suggest that under conditions of high transcriptomic alignment, the model preferentially generates compounds structurally related to known active molecules, demonstrating effective cross-modal alignment between gene expression patterns and molecular structure features.

Generating Inhibitor-Like Molecules from Knockout Gene Expression Signatures

We next investigated whether TETGen could generate inhibitor-like molecules solely from gene expression signatures of knocked-out targets, without any prior structural information. This approach is grounded in the biological rationale that genetic ablation and pharmacological inhibition of a target often elicit convergent transcriptional responses. We obtained a benchmark set of ten targets from the literature, each with at least ten known inhibitors[38],

and used siRNA-mediated knockdown signatures in MCF-7 cells as conditional inputs for generation. As shown in **Figure 5a**, TETGen with transcriptome-supervised attention generated molecules with significantly higher structural similarity to known inhibitors compared to the unsupervised version, achieving a mean MACCS Tanimoto similarity of 0.75 ±0.09. We observed a clear positive correlation between transcriptomic matching scores and structural similarity to known actives (**Figure 5b**), indicating that molecules better aligned with the input knockout signature occupied chemical space closer to validated inhibitors. Representative examples (**Figure 5c**) show that generated molecules often reproduced key functional groups and scaffold arrangements of known inhibitors, despite no explicit exposure to inhibitor structures during training. These results demonstrate that TETGen can translate purely transcriptional perturbation signals into inhibitor-like molecular structures, advancing biology-driven drug design.

Generation and Validation of BTK Inhibitor Candidates from CLL Disease Transcriptomes

BTK is a tyrosine kinase that plays a central role in B-cell receptor signaling in chronic lymphocytic leukemia (CLL), and dasatinib, a second-generation tyrosine kinase inhibitor, has shown clinical potential in CLL treatment[39-40]. To evaluate whether TETGen can generate novel inhibitors based on disease-specific transcriptomic features, we conditioned the model on transcriptomic profiles obtained from peripheral blood lymphocytes of CLL patients following dasatinib treatment[41]. As shown in **Figure 6a**, the generated molecules included candidates exhibiting both high transcriptomic matching scores (>0.6) and favorable drug-likeness (QED > 0.7). Using the co-crystal structure of dasatinib bound to BTK (PDB: 3K54)[42], we performed molecular docking simulations. The results showed that several generated molecules interacted with key hydrophobic residues in the BTK catalytic pocket, with binding modes broadly similar to that of dasatinib (**Figure 6b–f**). This case demonstrates that TETGen can leverage disease-specific transcriptomic features to generate structurally novel candidates with plausible binding modes, providing a foundation for transcriptome-guided design of kinase inhibitors targeting B-cell receptor signaling.

Transcriptome-Driven Generation of NEK7 Molecular Glue Candidates with Computational Validation

NEK7 regulates NLRP3 inflammasome activation through kinase-independent promotion of NLRP3 oligomerization, driving IL-1β maturation and pyroptosis in inflammatory diseases including gout and atherosclerosis[43-44]. Recent studies show that molecular glue degraders can induce NEK7 ubiquitination by facilitating its interaction with CRBN E3 ligase, suppressing inflammasome hyperactivation. To investigate whether transcriptomic data alone can guide generation of potential NEK7 molecular glues, we conditioned TETGen on transcriptomic profiles from NEK7-knockout MCF7 cells. The model generated candidates with high transcriptomic

matching scores and drug-likeness, several exceeding 0.8 for both metrics (**Figure 7a**). Using the NEK7–CRBN–molecular glue crystal structure[45], we performed docking on high-confidence candidates. One molecule stably occupied the known binding pocket with characteristic interactions (**Figure 7b**). In 100 ns molecular dynamics simulations, a duration sufficient to assess initial binding stability, this candidate maintained heavy atom RMSD within 2–3 Å, formed persistent hydrogen bonding with NEK7 Glu37 and π–π stacking with CRBN Trp386, with average non-bonded contacts approximately 500 (**Figure 7c–f**). The NEK7–CRBN complex backbone RMSD remained around 3 Å **(Figure 7d),** with G-loop contacts near 250 **(Figure 7f)**. These results indicate that TETGen can generate molecules capable of stabilizing the NEK7–CRBN interface, providing a basis for further experimental investigation of these candidates as potential molecular glue degraders.

4 Discussions

In this study, we introduced TETGen, a transcriptome enhanced Transformer generative model that integrates PPI derived functional gene embeddings with a dual supervised cross attention mechanism for biologically aware molecular design. A key innovation lies in using PPI network random walks to construct functionally enriched positional encodings for genes, moving beyond approaches that treat genes as independent features such as one hot encoding or expression values alone. Ablation experiments confirmed that removing these functional encodings significantly degrades generation quality and physicochemical properties, underscoring the importance of capturing biological context at the input level. The transcriptomic supervision of cross attention weights further provided an interpretable bridge between gene expression changes and molecular generation, with the resulting gene mapping score emerging as a quantitative indicator of biological relevance.

Notably, TETGen demonstrated robust cross cell line generalization when applied to unseen cell lines without fine tuning, maintaining superiority in internal diversity and achieving consistently higher QED and synthetic accessibility scores compared to TransGEM across all tested cell lines. The model's ability to generate inhibitor like molecules solely from knockout signatures, with mean MACCS similarity of 0.75 ± 0.09 to known actives across ten targets, suggests that transcriptional perturbation signatures encode sufficient target specific information to guide molecular generation toward relevant regions of chemical space. Case studies further illustrated TETGen's translational potential. In the CLL setting, the model generated BTK inhibitor candidates that adopted binding modes broadly similar to

dasatinib in docking simulations. For NEK7 molecular glue discovery, TETGen produced a candidate that maintained stable binding throughout 100 ns molecular dynamics simulations, a duration sufficient to assess initial binding stability, forming key interactions with both NEK7 Glu37 and CRBN Trp386 while stabilizing the protein protein interface.

Despite these promising results, several limitations warrant acknowledgment. TETGen did not exceed BiCEV in QED and synthetic accessibility scores, indicating opportunities for incorporating multi objective optimization strategies to balance structural diversity with drug-like properties. Experimental validation, such as surface plasmon resonance or cellular assays, remains necessary to confirm actual biological activity, and the current reliance on L1000 landmark genes captures only a fraction of the transcriptome. Future work will explore attention perturbation methods to elucidate causal relationships between attended genes and specific structural features, while incorporating full transcriptomic profiles and multi objective optimization to enhance biological fidelity and generation quality.

Historically, many successful drugs have emerged from phenotypic screening, observing desired effects without prior knowledge of molecular targets[46-47]. TETGen offers a computational counterpart to this phenotype first approach by learning explicit and quantifiable associations between transcriptional states and molecular structures. Instead of screening millions of compounds in costly phenotypic assays, disease relevant gene expression signatures can now directly guide candidate generation. By making the connection between transcriptional phenotypes and chemical structures both explicit and quantifiable, frameworks like TETGen provide a computational bridge between chemical and biological spaces, offering a practical tool for phenotype inspired drug design.

5 Conclusion

This study presents TETGen, a transcriptome-enhanced Transformer model that integrates PPI-derived functional gene embeddings with supervised cross-attention for biologically aware molecular generation. The model consistently generates structurally diverse, drug-like molecules that align with input transcriptomic profiles, with a matching score above 0.6 serving as a reliable quantitative indicator of biological relevance. Case studies on NEK7 molecular glue and BTK inhibitor discovery further demonstrate the potential of translating disease-specific transcriptomic signatures into candidate therapeutic molecules, offering an interpretable framework for phenotype-inspired drug design.

Supplementary Information

Supplementary Figure S1: GPT Model Pre-training Loss results. Supplementary Figure S2. TETGen Model Loss results. Supplementary Table S1: Detailed hyperparameters of GPT model.

Statements and Declarations

Funding

This work was supported by the Youth Science Fund of the National Natural Science Foundation of China (No.82304382). Key Program of the Education Department of Anhui Province (No. 2025AHGXZK30758).

Competing Interests

The authors declare no competing financial interests.

Authors' Contributions

J.G., Y.C. and R.Z. contributed equally to this work. Conceptualization: Q.L., B.Y.; Methodology: J.G., R.Z.; Formal analysis and investigation: J.G., Y.C., Y.L., J.W., Y.D., M.M., D.W.; Writing - original draft preparation: J.G., Q.L.; Writing - review and editing: B.Y., Y.Z., Q.L.; Supervision: Q.L., B.Y., Y.Z.; Funding acquisition: Q.L., Y.Z.

Data Availability Statement

The complete code for this work is deposited and available for download at https://github.com/xiaogaohaha/TETGen/tree/master.

Ethics Approval

This study did not involve human participants, human data, or biological materials, so no ethical approval was required.

Author Information

Corresponding Author

Qichao Luo-School of Pharmacy, Anhui Medical University, Hefei 230032, Anhui, China; **Email**: luoqichao@ahmu.edu.cn

Bin Yuan- School of Pharmacy, Anhui Medical University, Hefei 230032, Anhui, China; **Email:** yuanbin@ahmu.edu.cn

Yan Zhou- The First Affiliated Hospital of Anhui Medical University, Hefei, Anhui, China; **Email:** yfyb33719021@fy.ahmu.edu.cn

Authors

Jiahui Gao-School of Pharmacy, Anhui Medical University, Hefei 230032, Anhui, China

Yanyun Cao - The First Affiliated Hospital of Anhui Medical University, Hefei 230032, Anhui, China

Rui Zhou -School of Pharmacy, Anhui Medical University, Hefei 230032, Anhui, China

Yingzi Li -School of Biomedical Engineering, Anhui Medical University, Hefei 230032, Anhui, China

Jiale Wang -School of Biomedical Engineering, Anhui Medical University, Hefei 230032, Anhui, China

Wanyue Ding-School of Pharmacy, Anhui Medical University, Hefei 230032, Anhui, China

Mingzhi Ma- The First Affiliated Hospital of Anhui Medical University, Hefei, 230032，Anhui, China.

Danchen Wu- School of Pharmacy, Anhui Medical University, Hefei 230032, Anhui, China

References

[1] Farghali H, Kutinová Canová N, Arora M (2021) The potential applications of artificial intelligence in drug discovery and development. *Physiol Res* **70**(Suppl4):S715–S722. https://doi.org/10.33549/physiolres.934765.

[2] Sabe VT, Ntombela T, Jhamba LA, Maguire GEM, Govender T, Naicker T, Kruger HG (2021) Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: a review. *Eur J Med Chem* **224**:113705. https://doi.org/10.1016/j.ejmech.2021.113705.

[3] Mak KK, Pichika MR (2019) Artificial intelligence in drug development: present status and future prospects. *Drug Discov Today* **24**(3):773–780. https://doi.org/10.1016/j.drudis.2018.11.014.

[4] Chan HCS, Shan H, Dahoun T, Vogel H, Yuan S (2019) Advancing drug discovery via artificial intelligence. *Trends Pharmacol Sci* **40**(8):592–604. https://doi.org/10.1016/j.tips.2019.06.004.

[5] Maia EHB, Assis LC, de Oliveira TA, da Silva AM, Taranto AG (2020) Structure-based virtual screening: from classical to artificial intelligence. *Front Chem* **8**:343. https://doi.org/10.3389/fchem.2020.00343.

[6] Delre P, Cerchia C, Falbo E, Lavecchia A (2026) Artificial intelligence in ligand-based drug design. In: Lavecchia A (ed) *Applied Artificial Intelligence for Drug Discovery: From Data-Driven Insights to Therapeutic Innovation*. Springer Nature Switzerland, Cham, pp 131–155. https://doi.org/10.1007/978-3-031-98022-0_5.

[7] Giordano D, Biancaniello C, Argenio MA, Facchiano A (2022) Drug design by pharmacophore and virtual screening approach. *Pharmaceuticals (Basel)* **15**(5):646. https://doi.org/10.3390/ph15050646.

[8] Ripphausen P, Stumpfe D, Bajorath J (2012) Analysis of structure-based virtual screening studies and characterization of identified active compounds. *Future Med Chem* **4**(5):603–613. https://doi.org/10.4155/fmc.12.18.

[9] Zhang X, Shen C, Zhang H, Kang Y, Hsieh CY, Hou T (2024) Advancing ligand docking through deep learning: challenges and prospects in virtual screening. *Acc Chem Res* **57**(10):1500–1509. https://doi.org/10.1021/acs.accounts.4c00093.

[10] Anighoro A, Bajorath J, Rastelli G (2014) Polypharmacology: challenges and opportunities in drug discovery. *J Med Chem* **57**(19):7874–7887. https://doi.org/10.1021/jm5006463.

[11] Zięba A, Stępnicki P, Matosiuk D, Kaczor AA (2022) What are the challenges with multi-targeted drug design for complex diseases? *Expert Opin Drug Discov* **17**(7):673–683. https://doi.org/10.1080/17460441.2022.2072827.

[12] Yuan Y, Yu L, Bi C, Huang L, Su B, Nie J, Dou Z, Yang S, Li Y (2025) A new paradigm for drug discovery in the treatment of complex diseases: drug discovery and optimization. *Chin Med* **20**(1):40. https://doi.org/10.1186/s13020-025-01075-4.

[13] De Cao N, Kipf T (2022) MolGAN: an implicit generative model for small molecular graphs. arXiv preprint arXiv:1805.11973 [stat.ML]. https://arxiv.org/abs/1805.11973.

[14] Elton DC, Boukouvalas Z, Fuge MD, Chung PW (2019) Deep learning for molecular design—a review of the state of the art. *Mol Syst Des Eng* **4**(4):828–849. https://doi.org/10.1039/C8ME00125G.

[15] Jin W, Barzilay R, Jaakkola T (2018) Junction tree variational autoencoder for molecular graph generation. In: Dy J, Krause A (eds) *Proceedings of the 35th International Conference on Machine Learning*. PMLR, Proceedings of Machine Learning Research **80**:2323–2332. https://proceedings.mlr.press/v80/jin18a.html.

[16] Liu P, Ren Y, Tao J, Ren Z (2024) GIT-Mol: a multi-modal large language model for molecular science with graph, image, and text. *Comput Biol Med* **171**:108073. https://doi.org/10.1016/j.compbiomed.2024.108073.

[17] Popova M, Isayev O, Tropsha A (2018) Deep reinforcement learning for de novo drug design. *Sci Adv* **4**(7):eaap7885. https://doi.org/10.1126/sciadv.aap7885.

[18] Xiong Y, Wang Y, Wang Y, Li C, Yusong P, Wu J, Wang Y, Gu L, Butch CJ (2023) Improving drug discovery with a hybrid deep generative model using reinforcement learning trained on a Bayesian docking approximation. *J Comput Aided Mol Des* **37**(11):507–517. https://doi.org/10.1007/s10822-023-00523-3.

[19] Zeng X, Wang F, Luo Y, Kang SG, Tang J, Lightstone FC, Fang EF, Cornell W, Nussinov R, Cheng F (2022) Deep generative molecular design reshapes drug discovery. *Cell Rep Med* **3**(12):100794. https://doi.org/10.1016/j.xcrm.2022.100794.

[20] Grisoni F, Moret M, Lingwood R, Schneider G (2020) Bidirectional molecule generation with recurrent neural networks. *J Chem Inf Model* **60**(3):1175–1183. https://doi.org/10.1021/acs.jcim.9b00943.

[21] Olivecrona M, Blaschke T, Engkvist O, Chen H (2017) Molecular de-novo design through deep reinforcement learning. *J Cheminform* **9**(1):48. https://doi.org/10.1186/s13321-017-0235-x.

[22] Lim J, Ryu S, Kim JW, Kim WY (2018) Molecular generative model based on conditional variational autoencoder for de novo molecular design. *J Cheminform* **10**(1):31. https://doi.org/10.1186/s13321-018-0286-7.

[23] Oliveira AF, Da Silva JLF, Quiles MG (2022) Molecular property prediction and molecular design using a supervised grammar variational autoencoder. *J Chem Inf Model* **62**(4):817–828. https://doi.org/10.1021/acs.jcim.1c01573.

[24] Kang S, Cho K (2019) Conditional molecular design with deep generative models. *J Chem Inf Model* **59**(1):43–52. https://doi.org/10.1021/acs.jcim.8b00263.

[25] Kadurin A, Nikolenko S, Khrabrov K, Aliper A, Zhavoronkov A (2017) druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Mol Pharm* **14**(9):3098–3104. https://doi.org/10.1021/acs.molpharmaceut.7b00346.

[26] Zou J, Yu J, Hu P, Zhao L, Shi S (2023) STAGAN: an approach for improve the stability of molecular graph generation based on generative adversarial networks. *Comput Biol Med* **167**:107691. https://doi.org/10.1016/j.compbiomed.2023.107691.

[27] Bagal V, Aggarwal R, Vinod PK, Priyakumar UD (2022) MolGPT: molecular generation using a transformer-decoder model. *J Chem Inf Model* **62**(9):2064–2076. https://doi.org/10.1021/acs.jcim.1c00600.

[28] Honda S, Shi S, Ueda HR (2019) Smiles transformer: pre-trained molecular fingerprint for low data drug discovery. arXiv preprint arXiv:1911.04738 [cs.LG]. https://arxiv.org/abs/1911.04738.

[29] Schneuing A, Harris C, Du Y, Didi K, Jamasb A, Igashov I, Du W, Gomes C, Blundell TL, Lio P, Welling M, Bronstein M, Correia B (2024) Structure-based drug design with equivariant diffusion models. *Nat Comput Sci* **4**(12):899–909. https://doi.org/10.1038/s43588-024-00737-x.

[30] Xie J, Chen S, Lei J, Yang Y (2024) DiffDec: structure-aware scaffold decoration with an end-to-end diffusion model. *J Chem Inf Model* **64**(7):2554–2564. https://doi.org/10.1021/acs.jcim.3c01466.

[31] Das D, Chakrabarty B, Srinivasan R, Roy A (2023) Gex2SGen: designing drug-like molecules from desired gene expression signatures. *J Chem Inf Model* **63**(7):1882–1893. https://doi.org/10.1021/acs.jcim.2c01301.

[32] Pravalphruekul N, Piriyajitakonkij M, Phunchongharn P, Piyayotai S (2023) De novo design of molecules with multiaction potential from differential gene expression using variational autoencoder. *J Chem Inf Model* **63**(13):3999–4011. https://doi.org/10.1021/acs.jcim.3c00457.

[33] Liu Y, Yu H, Duan X, Zhang X, Cheng T, Jiang F, Tang H, Ruan Y, Zhang M, Zhang H, Zhang Q (2024) TransGEM: a molecule generation model based on Transformer with gene expression data. *Bioinformatics* **40**(5):btae189. https://doi.org/10.1093/bioinformatics/btae189.

[34] Subramanian A, Narayan R, Corsello SM, et al (2017) A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**(6):1437–1452.e17. https://doi.org/10.1016/j.cell.2017.10.049.

[35] Szklarczyk D, Nastou K, Koutrouli M, Kirsch R, Mehryary F, Hachilif R, Hu D, Peluso ME, Huang Q, Fang T, Doncheva NT, Pyysalo S, Bork P, Jensen LJ, von Mering C (2025) The STRING database in 2025: protein networks with directionality of regulation. *Nucleic Acids Res* **53**(D1):D730–D737. https://doi.org/10.1093/nar/gkae1113.

[36] Liu Y, Yang X, Gan J, Chen S, Xiao Z-X, Cao Y (2022) CB-Dock2: improved protein-ligand blind docking by integrating cavity detection, docking and homologous template fitting. Nucleic Acids Res 50(W1):W159–W164. https://doi.org/10.1093/nar/gkac394.

[37] Guterres H, Im W (2023) CHARMM-GUI-Based induced fit docking workflow to generate reliable protein-ligand binding modes. J Chem Inf Model 63(15):4772–4779. https://doi.org/10.1021/acs.jcim.3c00416.

[38] Méndez-Lucio O, Baillif B, Clevert DA, Rouquié D, Wichard J (2020) De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nat Commun* **11**(1):10. https://doi.org/10.1038/s41467-019-13807-w.

[39] Feng Y, Duan W, Cu X, Liang C, Xin M (2019) Bruton's tyrosine kinase (BTK) inhibitors in treating cancer: a patent review (2010-2018). *Expert Opin Ther Pat* **29**(4):217–241. https://doi.org/10.1080/13543776.2019.1579962.

[40] Robak T, Witkowska M, Smolewski P (2022) The role of Bruton's kinase inhibitors in chronic lymphocytic leukemia: current status and future directions. *Cancers (Basel)* **14**(3):771. https://doi.org/10.3390/cancers14030771.

[41] Blätte TJ, Machnicki MM, Glodkowska-Mrowka E, Dolnik A, Karp M, Karczmarczyk A, Giannopoulos K, Bullinger L, Stoklosa T (2020) Gene expression profiling predicts sensitivity of chronic lymphocytic leukemia cells to dasatinib. *Hemasphere* **5**(1):e514. https://doi.org/10.1097/HS9.0000000000000514.

[42] Marcotte DJ, Liu YT, Arduini RM, et al (2010) Structures of human Bruton's tyrosine kinase in active and inactive conformations suggest a mechanism of activation for TEC family kinases. *Protein Sci* **19**(3):429–439. https://doi.org/10.1002/pro.321.

[43] Chen X, Liu G, Yuan Y, Wu G, Wang S, Yuan L (2019) NEK7 interacts with NLRP3 to modulate the pyroptosis in inflammatory bowel disease via NF-κB signaling. *Cell Death Dis* **10**(12):906. https://doi.org/10.1038/s41419-019-2157-1.

[44] Wang X, Zhao Y, Wang D, et al (2023) ALK-JNK signaling promotes NLRP3 inflammasome activation and pyroptosis via NEK7 during Streptococcus pneumoniae infection. *Mol Immunol* **157**:78–90. https://doi.org/10.1016/j.molimm.2023.03.016.

[45] Marcotte DJ, Liu YT, Arduini RM, et al (2010) Structures of human Bruton's tyrosine kinase in active and inactive conformations suggest a mechanism of activation for TEC family kinases. *Protein Sci* **19**(3):429–439. https://doi.org/10.1002/pro.321.

[46] Childers WE, Elokely KM, Abou-Gharbia M (2020) The resurrection of phenotypic drug discovery. *ACS Med Chem Lett* **11**(10):1820–1828. https://doi.org/10.1021/acsmedchemlett.0c00006.

[47] Vincent F, Nueda A, Lee J, Schenone M, Prunotto M, Mercola M (2022) Phenotypic drug discovery: recent successes, lessons learned and new directions. *Nat Rev Drug Discov* **21**(12):899–914. https://doi.org/10.1038/s41573-022-00472-w.

Figure legend

Figure 1. Architecture of the TETGen Model.

Figure 2. Evaluation of model performance. (a) Comparison of molecule generation performance across different models. (b) Gene mapping scores of TETGen compared with unsupervised-TETGen and TransGEM, demonstrating improved biological relevance through transcriptome-guided attention supervision. (c) Distributions of molecular weight, logP, quantitative estimate of drug-likeness (QED), and synthetic accessibility (SA) scores for molecules generated by each model.

Figure 3. Evaluation of TETGen generalization across different cell lines. (a) Comparison of molecule generation performance across different models on all tested cell lines. (b) Gene mapping scores of TETGen compared across MCF7, PC3, A549, and HT29 cell lines, demonstrating the transferability of transcriptome-guided attention supervision. (c) Distributions of molecular weight, logP, quantitative estimate of drug-likeness (QED), and synthetic accessibility (SA) scores for molecules generated by each model across all tested cell lines.

Figure 4. Cross-modal alignment between chemical structure and transcriptomic profiles. (a) Distributions of structural similarity (MACSS) between generated molecules and known bioactive compounds under different supervision conditions. (b) Comparison of structural similarity to known bioactive compounds between molecules with high (> 0.6) and low (< 0.4) Gene matching scores, Statistical significance was determined by two-sided t-test (*$p < 0.05$). (c–h) Representative generation cases illustrating the relationship between transcriptomic matching

score and structural similarity. For each case, the reference compound (input-derived active molecule), the generated molecule, and their MACSS similarity scores are shown.

Figure 5. Generation of inhibitor-like molecules from knockout gene expression signatures. (a) Comparison of structural similarity (MACCS Tanimoto) to known inhibitors between molecules generated by TETGen with transcriptome-supervised attention and the unsupervised version. TETGen achieved significantly higher similarity (mean 0.75 ± 0.09) across ten targets. (b) Comparison of structural similarity to known inhibitors between TETGen generated molecules with high (> 0.6) and low (< 0.4) gene mapping scores. (c) Representative generation examples for ten targets by TETGen.

Figure 6. Generation and validation of BTK inhibitor candidates from CLL disease transcriptomes. (a) Scatter plot showing generated molecules with high transcriptomic matching scores and favorable drug-likeness. (b) Representative generated molecules selected for docking analysis. (c) Superimposition of representative generated molecules (pink) with dasatinib (yellow) in the BTK binding pocket (PDB: 3K54). (d–f) Docking poses of representative generated molecules (pink) showing their binding modes with BTK.

Figure 7. Transcriptome-driven generation and computational validation of NEK7 molecular glue candidates. (a) Scatter plot showing generated molecules with high transcriptomic matching scores and favorable drug-likeness (QED). (b) Representative candidate molecule structure and its docked pose within the NEK7–CRBN–molecular glue ternary complex. (c) Binding mode of the representative candidate at the NEK7–CRBN interface. (d) Heavy atom

RMSD of the candidate molecule (green) and backbone RMSD of the NEK7–CRBN complex (blue) over 100 ns molecular dynamics simulations. (e) Key persistent interactions during the simulation, including hydrogen bonding with NEK7 Glu37 and π–π stacking with CRBN Trp386. (f) Time evolution of non-bonded contacts between the candidate molecule and the NEK7–CRBN complex (green), along with contacts between NEK7 and CRBN at the interface (blue).

Figure 1

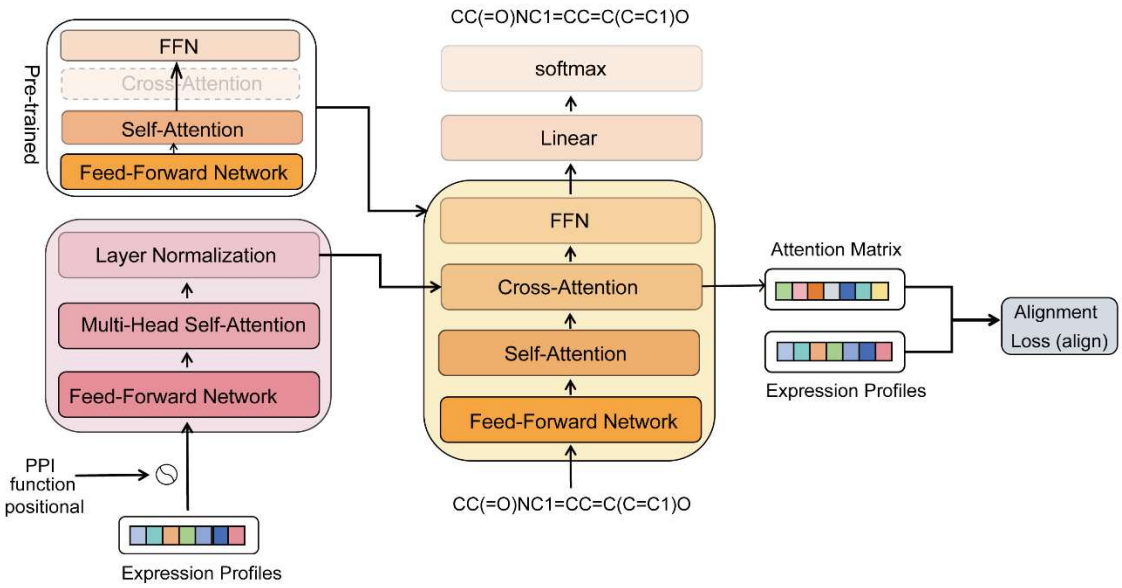

Pre-trained
FFN
Cross-Attention
Self-Attention
Feed-Forward Network
Layer Normalization
Multi-Head Self-Attention
Feed-Forward Network
PPI
function
positional
Expression Profiles
CC(=O)NC1=CC=C(C=C1)O
softmax
Linear
FFN
Cross-Attention
Self-Attention
Feed-Forward Network
CC(=O)NC1=CC=C(C=C1)O
Attention Matrix
Expression Profiles
Alignment
Loss (align)

Figure 2

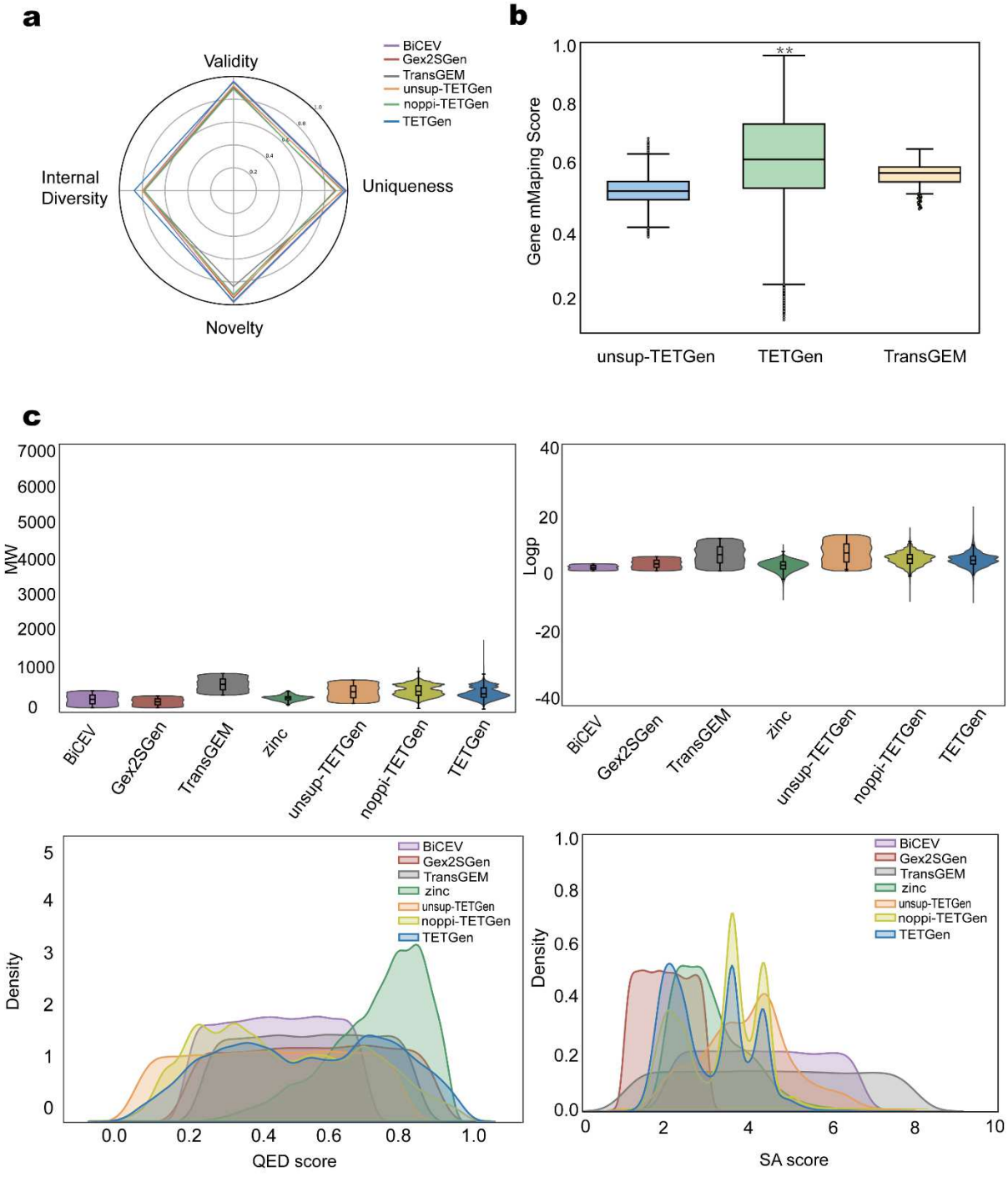

a
Validity
Uniqueness
Novelty
Internal Diversity
BiCEV
Gex2SGen
TransGEM
unsup-TETGen
noppi-TETGen
TETGen
b
Gene mMaping Score
1.0
0.8
0.6
0.4
0.2
**
unsup-TETGen
TETGen
TransGEM
c
MW
7000
6000
5000
4000
3000
2000
1000
0
Logp
40
20
0
-20
-40
BiCEV
Gex2SGen
TransGEM
zinc
unsup-TETGen
noppi-TETGen
TETGen
Density
QED score
SA score

Figure 3

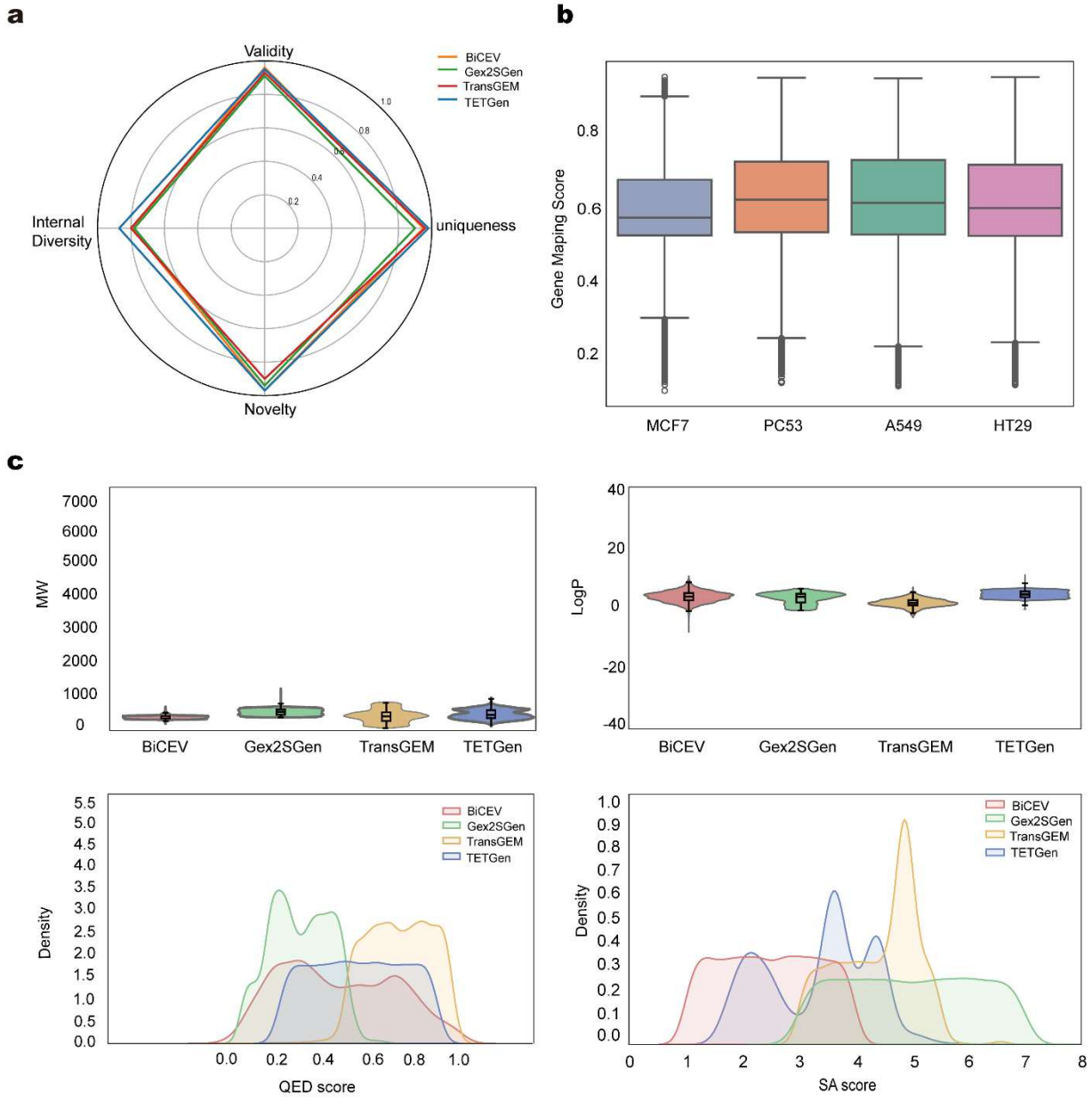

a
Validity
Internal Diversity
uniqueness
Novelty
BiCEV
Gex2SGen
TransGEM
TETGen
b
Gene Maping Score
MCF7
PC53
A549
HT29
c
MW
LogP
Density
QED score
SA score
BiCEV
Gex2SGen
TransGEM
TETGen

Figure 4

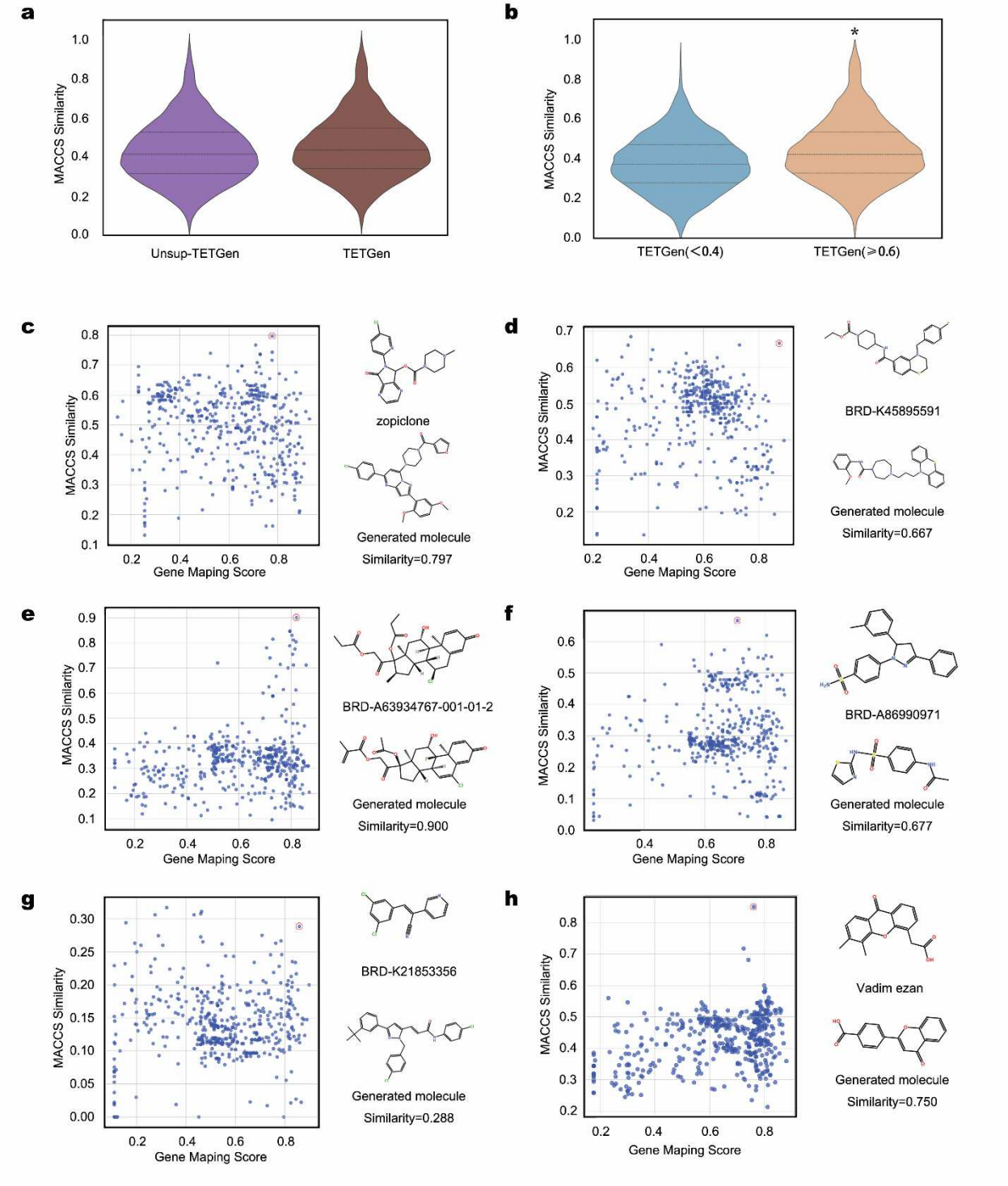

a
MACCS Similarity
Unsup-TETGen
TETGen
b
MACCS Similarity
*
TETGen(<0.4)
TETGen(≥0.6)
c
MACCS Similarity
Gene Maping Score
zopiclone
Generated molecule
Similarity=0.797
d
MACCS Similarity
Gene Maping Score
BRD-K45895591
Generated molecule
Similarity=0.667
e
MACCS Similarity
Gene Maping Score
BRD-A63934767-001-01-2
Generated molecule
Similarity=0.900
f
MACCS Similarity
Gene Maping Score
BRD-A86990971
Generated molecule
Similarity=0.677
g
MACCS Similarity
Gene Maping Score
BRD-K21853356
Generated molecule
Similarity=0.288
h
MACCS Similarity
Gene Maping Score
Vadim ezan
Generated molecule
Similarity=0.750

Figure 5

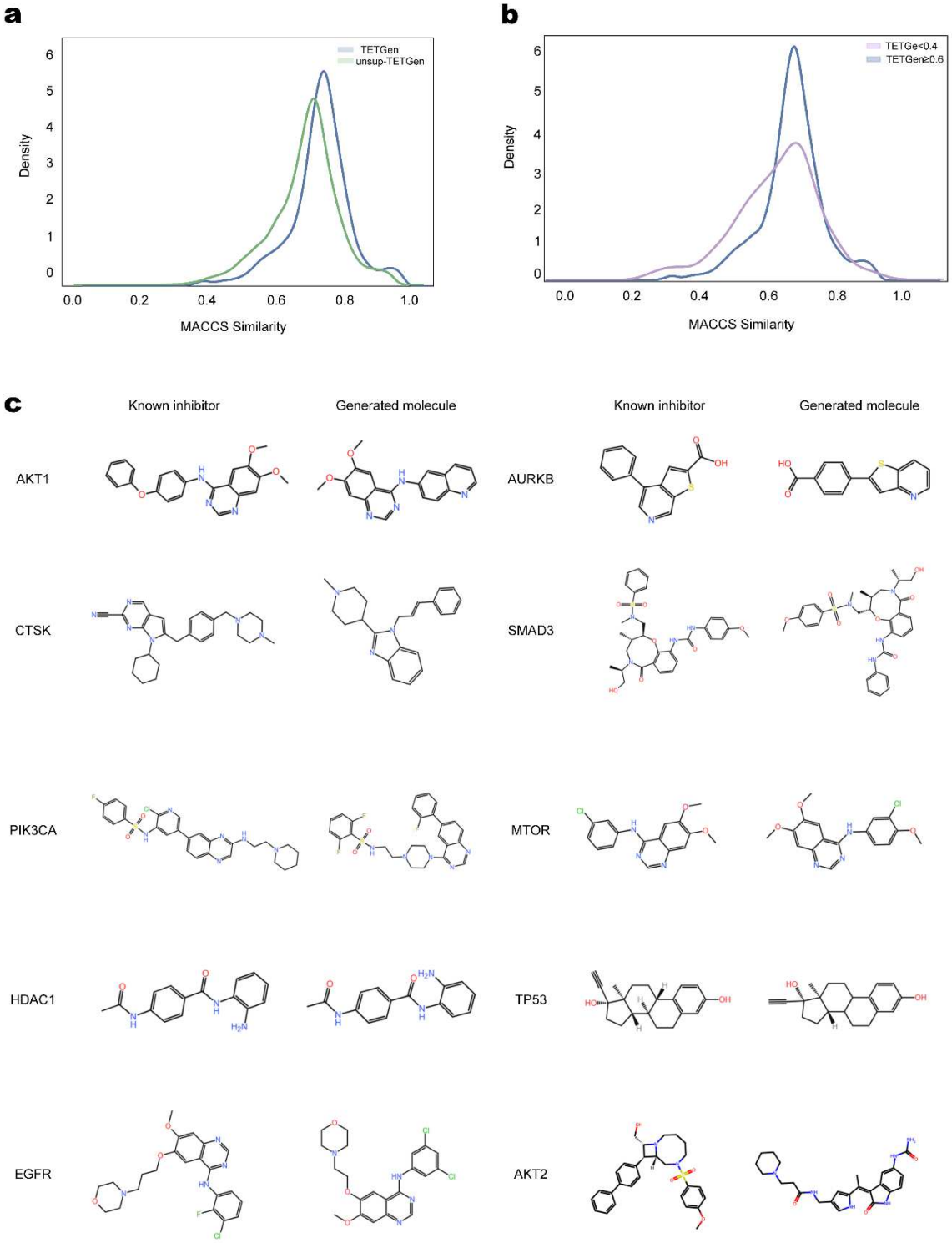

a
TETGen
unsup-TETGen
Density
MACCS Similarity
b
TETGe<0.4
TETGen≥0.6
Density
MACCS Similarity
c
Known inhibitor
Generated molecule
Known inhibitor
Generated molecule
AKT1
AURKB
CTSK
SMAD3
PIK3CA
MTOR
HDAC1
TP53
EGFR
AKT2

Figure 6

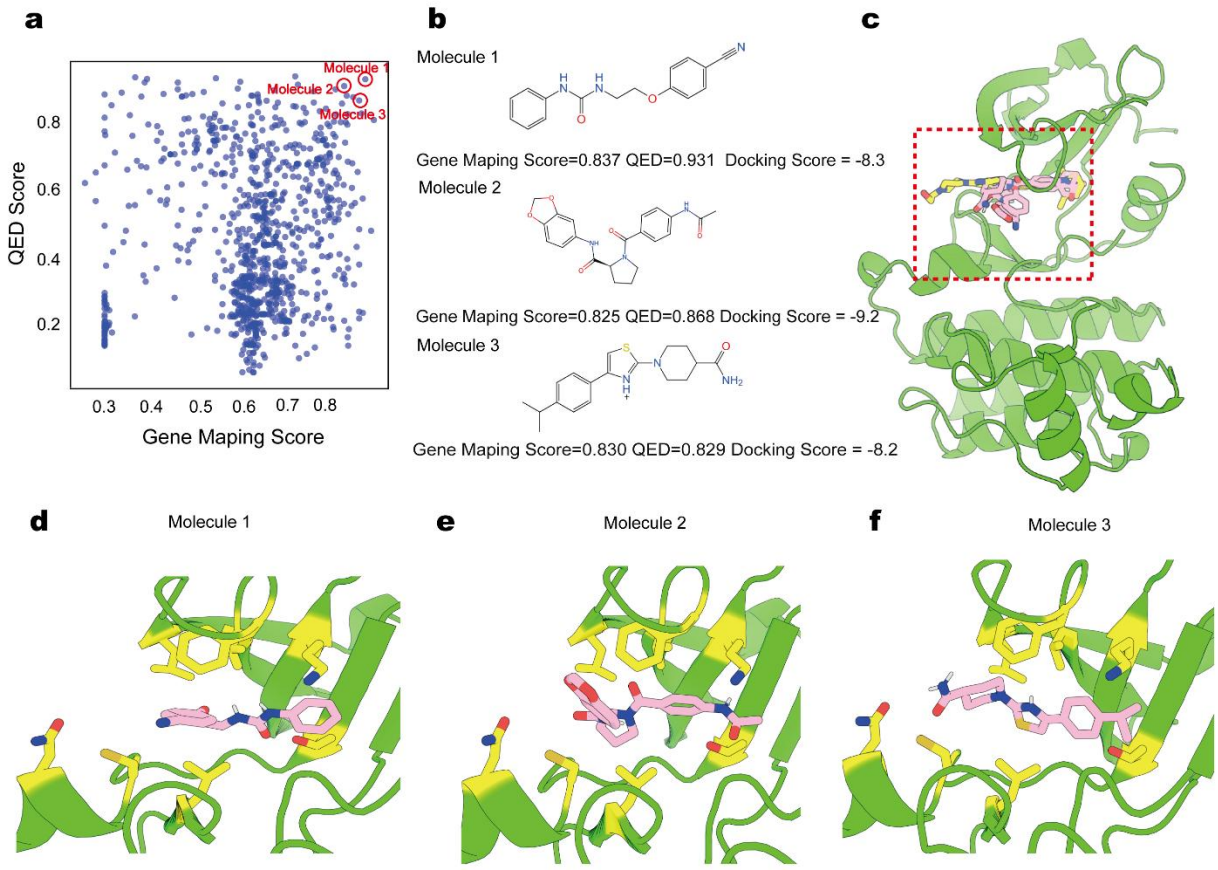
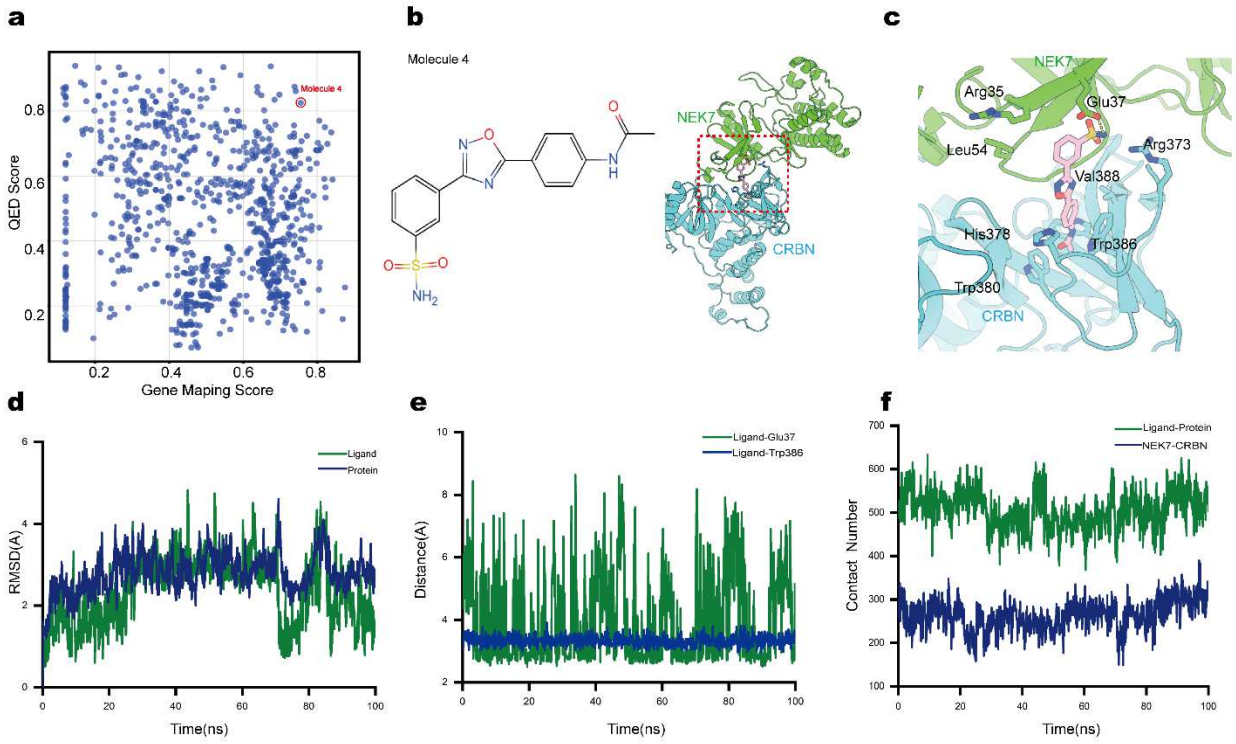

a
Molecule 1
Molecule 2
Molecule 3
QED Score
0.8
0.6
0.4
0.2
0.3 0.4 0.5 0.6 0.7 0.8
Gene Maping Score
b
Molecule 1
Gene Maping Score=0.837 QED=0.931 Docking Score = -8.3
Molecule 2
Gene Maping Score=0.825 QED=0.868 Docking Score = -9.2
Molecule 3
Gene Maping Score=0.830 QED=0.829 Docking Score = -8.2
c
d
Molecule 1
e
Molecule 2
f
Molecule 3

Figure 7


a
b
c
QED Score
Gene Maping Score
Molecule 4
Molecule 4
NEK7
CRBN
NEK7
Arg35
Glu37
Leu54
Arg373
Val388
His378
Trp386
Trp380
CRBN
d
e
f
RMSD(A)
Time(ns)
Ligand
Protein
Distance(A)
Time(ns)
Ligand-Glu37
Ligand-Trp386
Contact Number
Time(ns)
Ligand-Protein
NEK7-CRBN

Supplementary Files

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

- SI.docx