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.

Inverse design of chemical perturbations for desired cellular state transitions with GEMGen

Sida Shao

shaosida@westlake.edu.cn

Westlake University

Qun Jiang

Tsinghua University

Xing Ye

Westlake University

Zekun Guo

Beijing Zhongguancun Academy

Yingce Xia

Beijing Zhongguancun Academy

Zequn Liu

Beijing Zhongguancun Academy

Jiaying Xu

Beijing Zhongguancun Academy

Peiran Jin

Beijing Zhongguancun Academy

Fusong Ju

Beijing Zhongguancun Academy

Huanhuan Xia

Beijing Zhongguancun Academy

Shangya Feng

Westlake University

Haiguang Liu

Beijing Zhongguancun Academy

Tao Qin

Beijing Zhongguancun Academy

Pan Deng

Beijing Zhongguancun Academy

Article

Keywords:

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11151710/v1

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

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

Inverse design of chemical perturbations for desired cellular state transitions with GEMGen

Qun Jiang[1,2,7], Xing Ye[3,4,7], Zekun Guo[5,7], Yingce Xia[1,7], Zequn Liu[1], Jiaying Xu[5,6], Peiran Jin[1], Fusong Ju[1], Huanhuan Xia[5], Shangya Feng[3], Rui Jiang[2], Haiguang Liu[1], Tao Qin[1], Pan Deng[1,8,*], Sida Shao[3,4,8,*]

1. Zhongguancun Academy, Beijing, China.
2. Ministry of Education Key Laboratory of Bioinformatics, Bioinformatics Division at the Beijing National Research Center for Information Science and Technology, Center for Synthetic and Systems Biology, Department of Automation, Tsinghua University, Beijing, China
3. Zhejiang Key Laboratory of Precise Synthesis of Functional Molecules, Department of Chemistry, School of Science and Research Center for Industries of the Future, Westlake University, Hangzhou, Zhejiang 310030, P. R. China.
4. School of Life Sciences, Westlake University, Hangzhou, China
5. Zhongguancun Institute of Artificial Intelligence, Beijing, China
6. School of Global College, Shanghai Jiao Tong University, Shanghai, China
7. These authors contributed equally.
8. These authors contributed equally.
*: dengpan@bza.edu.cn; shaosida@westlake.edu.cn

Abstract

Perturbation biology links interventions to resulting cellular changes, providing a basis for understanding and manipulating biological processes. Recent advances in computational modelling have enabled prediction of cellular responses to diverse perturbations. However, many biomedical applications require going beyond response prediction to identify perturbations capable of inducing desired responses, an inverse design problem that remains largely underexplored compared to the forward prediction task. Here we introduce GEMGen, a large language model-based framework for the inverse design of chemical perturbations conditioned on desired cellular state transitions. GEMGen represents transitions as text-based, rank-ordered differential-gene instructions, generates molecular structures autoregressively, and uses a compound-response scorer to prioritize candidates for experimental testing. Across multiple benchmarks, GEMGen recovered small molecules structurally or mechanistically relevant to ground-truth compounds, and also generalized across perturbation modalities by generating chemical interventions from genetic perturbation signatures, outperforming existing perturbation design approaches. Beyond *in silico* evaluation, when conditioned on transcriptional signatures related to tissue fibrosis, GEMGen yielded eight novel antifibrotic structures from 17 *de novo* synthesized candidates. Notably, one lead candidate, GEMGen-010, exhibited strong antifibrotic activity in animal models of liver fibrosis without any structural optimization. Together, these results establish GEMGen as an effective approach extending perturbation modeling toward intervention design, broadening the methodology and application scope of perturbation biology.

Introduction

Perturbation biology seeks to understand and control biological systems with genetic, chemical and environmental interventions(1–5). Perturbation experiments, such as those in genetics and chemical genetics, have driven major biological discoveries(6–10) Continued advances in pooled genetic screen and high-throughput chemical transcriptomics have enabled perturbation responses to be profiled at far greater scale (11–16) and the resulting compendia span thousands of perturbations across many cellular contexts. Combined with concurrent advances in computational modeling, these resources now enable prediction of responses to perturbations that have not been experimentally tested(17–22). More recently, these capabilities have begun to converge with broader AI virtual cell models that seek to simulate cellular states and responses *in silico*(22–26). These developments have substantially expanded the perturbational landscape that can be modeled.

However, many important questions in perturbation biology run in the opposite direction(27,28). We refer to this as the inverse design problem of perturbation biology: whereas conventional perturbation-response modeling asks what cellular state will result from a given perturbation, the inverse design problem asks what perturbation can produce a desired change in cellular state. This distinction becomes particularly important when the objective is to actively control a biological system, such as redirecting a dysfunctional cellular state toward a native state(29,30), inducing a specific differentiation trajectory(31,32), or altering a defined molecular program(33,34). A related objective underlies many drug discovery campaigns, where compounds are sought for their ability to restore a specified function in cellular and more complex biological systems(35,36).

Existing perturbation-response models do not solve this problem. To identify an intervention for a desired cellular response, these models must instead evaluate a predefined set of candidate perturbations and rank them according to their predicted effects(37–39). This approach is effective when the relevant interventions are represented in the candidate library, but cannot identify interventions outside the search space, and the highest-ranked candidate may still produce a response far from the desired transition. Such limitation is particularly consequential for chemical perturbations. Although make-on-demand libraries now contain billions of accessible molecules, they still represent only a minute fraction of the chemical space that could in principle be explored, which has been estimated to contain on the order of 10^{60} structures(40–42). Several generative approaches have begun to tackle the inverse design problem by directly generating molecular structures conditioned on gene expression(43–48). However, their validation has remained largely retrospective or *in silico*, potentially due to their limited ability to generalize to unseen chemical and biological contexts. Whether previously unseen chemical perturbations can be prospectively generated to produce intended biological state transitions therefore remains unresolved.

To this end, we introduce GEMGen (Gene Expression-guided Molecular Generator), a generative framework for the inverse design of chemical perturbations for desired

transcriptomic state transitions. GEMGen converts transitions between initial and desired states into text-based, rank-ordered lists of differentially expressed genes and uses these representations to condition autoregressive molecular generation(49,50). A separately trained compound-response scorer then ranks generated structures by their predicted compatibility with the target transcriptional response, producing focused candidate lists for experimental testing. Together, these components translate a specified cellular state transition directly into a focused set of candidate chemical perturbations, without restricting the search to predefined compound libraries.

We evaluated GEMGen across progressively more challenging perturbation settings. First, using held-out compounds and experimental conditions from multiple large-scale chemical perturbation datasets, we showed that GEMGen can recover molecules structurally or mechanistically related to those that produced the target transcriptional responses, outperforming existing generative approaches. Second, we demonstrated that GEMGen can address the inverse design problem across perturbation modalities. Using CRISPR interference signatures as inputs, GEMGen recovered known inhibitors of targeted genes and designed structurally distinct molecules beyond the training database, one of which, flavokawain C, was experimentally validated as a previously unreported NRF2 activator. Finally, we applied GEMGen to a disease-relevant state transition by generating novel anti-fibrotic molecules. In addition to recovering known antifibrotic compounds and identifying existing molecules with previously unreported antifibrotic activities, GEMGen generated novel chemical structures with a high experimental success rate (47%): eight of 17 de novo synthesized candidates attenuated fibrotic phenotypes in cellular assays. One candidate, GEMGen-010, was further evaluated in a CCl_4-induced mouse model of liver fibrosis, where it reduced collagen deposition, lowered serum markers of hepatocellular injury and suppressed fibrosis-associated gene expression in the liver. Note this in vivo activity was achieved using GEMGen-generated compounds before any structural optimization.

Together, these results establish GEMGen as a general approach to the inverse design problem of perturbation biology. By directly generating chemical interventions from desired transcriptomic state transitions, GEMGen extends perturbation modeling from predicting the consequences of interventions to designing interventions capable of producing specified biological outcomes. More broadly, our results suggest that cellular state transitions can serve as a common design interface across perturbation modalities, enabling genetic, chemical and disease-associated states to be translated into experimentally actionable interventions.

Overview of the GEMGen framework

GEMGen is a generative LLM framework for the inverse design of chemical perturbations from desired cellular state transitions (**Figure 1A**). Given transcriptomic profiles representing the initial and target cellular state, GEMGen proposes a ranked list of candidate molecules predicted to induce the corresponding transitions through sequential

modules performing preprocess, generation, filtering, and scoring (**Figure 1B**). By generating molecular structures directly rather than searching over predefined compound collections, GEMGen expands perturbation design beyond the compounds already represented in existing libraries (**Figure 1C**).

At its core, GEMGen learns the correspondence between transcriptomic state transitions and chemical perturbations through a chemical perturbation generation task (**Figure 1D-E**). GEMGen leverages a science-focused foundation language model trained on a curated corpus spanning natural language, scientific literature, and structured biological databases (**Methods**) (51). This model operates in an instruction-response format that generates contextually relevant outputs conditioned on textual prompts. Before perturbation-generation training, the foundation model was further tuned to generate descriptions of genes and cell types curated from the NCBI database and the ATCC cell catalog (bioentity representation generation task; **Figure S1A, and Methods**). Joint training on gene- and cell-type-level samples provided a shared semantic representation of biological entities for subsequent perturbation modeling.

To learn broad correspondences between chemical structures and their transcriptional effects, we curated a high-quality subset from the large-scale LINCS database for model training (**Methods**) (52). Perturbation-induced transcriptional changes were transformed into text-based, rank-ordered lists of differentially expressed genes (**Figure 1D**) (49,50). Compared to quantitative differential expression values, we found that this representation improves robustness to technical variation. Specifically, numerical fold-change tended to cluster by experimental plate, whereas text-based embeddings grouped samples by compound identity across plates (**Figure S1B**). This effect is particularly evident in cases where fragmented clusters observed in the numerical space were pulled into a single, unified cluster through text-based abstraction (**Figure S1C, Methods**). These results suggest that text-based abstraction reduces sensitivity to dataset-specific expression magnitudes while preserving the direction of transcriptional responses, providing more consistent and stable representations of perturbation-induced state transitions.

A central motivation for inverse perturbation generation is that the chemical perturbation space extends far beyond the molecules that can be explicitly enumerated and searched. Across sampled outputs, 60.99% of generated compounds were absent from both the training dataset and the ZINC 11M reference library (53) (**Figure S1D**). These molecules covered a broad range of structural similarity to known compounds, with Tanimoto similarities spanning 0.2 to 0.9 based on Morgan fingerprints (**Figure S1E**). Projection of generated and reference molecules further showed that GEMGen accessed regions of chemical space that were underrepresented in both the training data and commonly used screening library (**Figure S1F**). Representative generated molecules remained structurally distinct from their closest database matches while retaining favorable drug-likeness and synthetic accessibility (**Figure S1F**). Together, these analyses demonstrate the ability of GEMGen to expand the candidate space beyond the chemical structures represented in

the training data and existing collections.

To translate the broad generative space into experimentally tractable candidates, GEMGen further incorporates a compound-response scoring module that prioritizes generated molecules according to their predicted ability to reproduce the desired transcriptional response (**Figure 1E and Methods**). Combined with filters for molecular properties (**Figure 1B**), GEMGen is able to produce focused and experimentally executable candidate lists.

GEMGen recovered ground-truth compounds and identified mechanistically related analogs across diverse *in silico* evaluation settings (**Figure 1F**). Prospective validation further demonstrated that GEMGen-generated candidates could produce the desired biological responses, including previously unseen *de novo* structures with cellular and *in vivo* activity (**Figure 1F**). Together, these results establish GEMGen as a practical framework for inverse perturbation design, enabling directly generation of experimentally actionable chemical interventions for desired cellular state transitions while extending exploration beyond predefined chemical collections.

GEMGen recovers chemical perturbations across unseen compounds and conditions

To benchmark GEMGen's inverse-design capability, we assembled two complementary evaluation datasets that test generalization across chemical structures and experimental conditions. First, we held out 404 LINCS samples as a "novel compound" test set, consisting of compounds structurally distinct from the training data (**Figure 2A, Methods**). This setting evaluates whether GEMGen can propose compounds beyond previously observed chemical structures. Second, to assess generalization across experimental conditions, we curated 640 high-quality compound–response pairs from the Tahoe-100M dataset (**Figure 2A**)(16), a large-scale single-cell perturbation atlas profiling transcriptional responses to small molecules across 50 cancer cell lines (**Methods**). Unlike LINCS data, which captures population-averaged expression changes, Tahoe-100M measures heterogeneous responses at single-cell resolution. This "novel condition" setting therefore evaluates whether a model trained exclusively on bulk transcriptomic profiles can generalize across measurement modality and experimental regime.

For each target transcriptional response, GEMGen generated 1,000 candidate compounds in SMILES format predicted to reproduce the corresponding gene expression changes. Generated molecules were filtered for syntactic validity and de-duplicated prior to analysis. We trained three GEMGen variants with increasing numbers of training epochs (Ckpt1-3). As expected, longer training improved syntactic validity while reducing molecular diversity (**Figure S2A-B**), while all variants maintained comparable drug-likeness, assessed via Lipinski's Rule of Five (Ro5) (54), quantitative estimate of drug-

likeness (QED) (55), and synthetic accessibility score (SAS) (56) (**Figure S2C-E**). We therefore combined the outputs of these variants into an ensemble pool for all downstream analyses.

Each perturbation-response condition is associated with a reference compound used to induce the transcriptional change, defined here as the “benchmark compound”. We first evaluated model performance by measuring the fraction of samples for which generated molecules recover the benchmark compound under varying structural similarity thresholds. In the novel compound setting, GEMGen successfully recovered benchmark compounds for 203 out of 404 samples at a Tanimoto similarity threshold of 0.6 (**Figure 2B**), while under the novel condition setting, GEMGen identified similar compounds for 242 out of 640 samples (**Figure 2C**).

Recognizing that mechanistically converging but structurally distinct compounds can induce similar transcriptional responses, we introduced “mechanistic analogs”: compounds acting on the same molecular targets or pathways as the benchmark compounds (**Methods, Table S1**). Incorporating mechanistic analogs substantially increased recovery rates across both settings, reaching 344 out of 404 samples in the novel compound test set and 466 out of 640 samples in the novel condition test set at a 0.6 similarity threshold (**Figure 2B-C**). This recovery of functionally related yet structurally diverse molecules suggests that GEMGen learned response-structure associations extend beyond direct analog retrieval, enabling generation of candidates based on their ability to produce related biological responses through convergent mechanisms. Collectively, 42.21% of hit molecules were absent from both the training dataset and the ZINC reference library, indicating the generation of previously unseen chemical entities (**Figure S3A**).

We then compared GEMGen’s performance against four alternative approaches **(Figure S2F-J)**. First, as variational autoencoder (VAE)-based architectures are widely used in molecular design for their smooth latent spaces and interpolation capabilities (43,57,58), we implemented a VAE-based transcriptomic-conditioned molecular generator, following the design principles of Gex2SGen (46) and trained on the same data as GEMGen (**Methods**). We also evaluated GexMolGen (47), which builds on the single-cell foundation model scGPT (26), and directly encodes gene-level expression profiles for molecule generation. In addition, we tested a general-domain large language model (GPT-5) prompted with expression-derived instructions but without domain-specific fine-tuning, to assess whether generic LLMs can infer biologically meaningful molecules from transcriptional patterns alone. Finally, we included random sampling from ChemDiv as a reference baseline representing unguided exploration of chemical diversity. Across both settings, the baseline approaches recovered substantially fewer compounds than GEMGen, highlighting GEMGen’s effectiveness at identifying response-relevant chemical structures (**Figure 2B-C**).

We further quantified the discovery yield of GEMGen by analyzing the number of hits per perturbation-response condition across diverse cell types. In the novel compound set, GEMGen successfully proposed benchmark structures and their analogs for samples across 33 cell types. A subset of samples was enriched for multiple hit molecules, implying that higher model confidence may lead to clustered generation of structurally related candidates around the ground truth (**Figure S3B**). While the yield was slightly lower in the novel condition set (across 48 cell types), the model maintained high hit density for many conditions (**Figure S3C**). Expanding the reference set to include mechanistic analogs increased overall hit rates in both settings (**Figure S3B-C**). For instance, when prompted with the SKBR3 cell line signature, GEMGen recovered the benchmark BCR-ABL inhibitor nilotinib and identified several mechanistic analogs, including imatinib and dasatinib, alongside novel saracatinib analogs with distinct scaffolds (**Figure 2D**) (59,60). Similarly, in the novel condition set using MIA PaCa-2 signatures, GEMGen recovered the CDK inhibitor dinaciclib and proposed functionally equivalent analogs such as alvocidib and PHA-793887 (**Figure 2E**) (61). These results demonstrate that response-conditioned generation can recover reference perturbations and structurally distinct molecules with annotated convergent mechanisms across held-out datasets and experimental regimes.

Response-based prioritization of GEMGen-generated chemical perturbations

To prioritize generated chemical perturbations for experimental testing, we introduced a compound-response scoring model that estimates the likelihood that a candidate molecule reproduces a target transcriptional response. By enriching for response-relevant candidates before experimental validation, this prioritization module serves as a practical filter for the generation pipeline.

We first evaluated the scorer's ability to discriminate true compound-response pairs from randomly shuffled negatives (**Methods**). For this assessment, positive pairs were restricted to benchmark compounds to provide a well-defined reference. Across the combined novel compound and novel condition test sets, the scorer achieved an area under the receiver operating characteristic curve (AUROC) of 0.945 and an area under the precision-recall curve (AUPRC) of 0.939 (**Figure 3A-B**). The score distributions exhibited strong separation between negative (median = 0.085) and positive (median = 0.849) pairs (**Figure 3C**). At a score threshold of 0.8, the precision reached 0.950, demonstrating that the scorer provides a strong signal for compound–response matching (**Figure 3D**).

Applying the scorer to GEMGen-generated compounds revealed a clear correspondence between structural proximity and predicted response relevance. Compounds with a Tanimoto similarity greater than 0.6 to benchmark molecules ("hits") received significantly higher scores than those with lower similarity ("non-hits"; **Figure 3E**). Critically, as benchmark compounds represent only a subset of potential solutions, we further examined the non-hit candidates. When compared with randomly paired controls

(**Methods**), non-hits exhibited significantly higher match scores than the random background (**Figure 3F**). This indicates that GEMGen generates chemically novel molecules that are nonetheless predicted to induce the desired transcriptional responses.

We next assessed whether score-based filtering improves the efficiency of compound discovery. For both test sets, increasing the score threshold progressively reduced the total number of candidate molecules while monotonically increasing hit rates (**Figure 3G and S4A**). To further evaluate ranking performance, we performed a top-K analysis on the score-ranked pool (~2,000 molecules per response condition). As K decreased, the proportion of true hits increased steadily, confirming that higher-scoring compounds are preferentially enriched for benchmark matches (**Figure S4B-C**). Together, these results establish response-based scoring as an effective prioritization step for GEMGen-generated chemical perturbations.

GEMGen designs chemical perturbations from genetic perturbation states

An important question for inverse perturbation design is whether transcriptomic changes can serve as a measurable interface for translating desired cellular responses across perturbation modalities. We therefore asked whether GEMGen, despite being trained on chemical perturbation-response data, could design chemical perturbations directly from transcriptional states induced by genetic perturbations (**Figure 4A-B**). Specifically, we tested whether CRISPR interference signatures could be used as inverse-design objectives to generate molecules that phenocopy the corresponding loss-of-function states.

To this end, we leveraged a CRISPR interference Perturb-seq dataset from Replogle et al.(13), which provides single-cell transcriptomic profiles for genome-scale genetic perturbations. From this dataset, we manually selected 21 genes whose knockdown elicited robust and distinctive transcriptional responses and for which sufficient cell coverage was available. In parallel, we assembled a ground-truth compound set by curating known inhibitors of the selected targets from the ChEMBL database (**Figure 4A, Table S2, Methods**).

For 11 of the 21 genetic perturbations, GEMGen generated hit compounds with Tanimoto similarity greater than 0.6 to the corresponding ground-truth inhibitors (**Figure 4C**), with a variable number of hits across perturbations (**Figure 4D**). By contrast, existing generative baselines and random sampling from the ChemDiv library showed limited ability to recover ground-truth-like compounds, producing only a single sporadic hit across all perturbation inputs (**Figure 4C**).

Interestingly, for genetic perturbation samples, the compound-response scoring module shows reduced discriminative power compared to previous test sets. Increasing the score threshold does not consistently improve hit rates for several genetic perturbations (**Table S3**), suggesting that high-scoring molecules do not always resemble known inhibitors.

This reduced separation between hits and non-hits may reflects the intrinsic difference between genetic and small-molecule perturbations. Unlike single-gene knockdown, small-molecule inhibitors may induce multifaceted transcriptomic responses due to polypharmacological and off-target effects (62). Consequently, high-scoring compounds in this setting do not necessarily resemble known reference inhibitors, raising the possibility that, for specific genetic perturbations, response matching may arise from chemically distinct compounds operating through alternative mechanisms. We therefore asked whether GEMGen can identify such functionally relevant yet structurally distinct compounds in this setting (**Figure 4B**).

Inverse design of NRF2 pathway activators from *KEAP1* genetic perturbation

To explore this possibility in a concrete setting, we focused on *KEAP1*, a genetic perturbation for which score-based prioritization did not result in higher hit rate across increased score threshold (**Figure S4D and Table S3**). KEAP1 is an E3 ubiquitin ligase that targets NFE2L2 (also known as NRF2) for proteasomal degradation and serves as a central regulator of the cellular oxidative stress response (**Figure 4E**)(63). Pharmacological activation of NRF2 through KEAP1 inhibition is a validated therapeutic strategy, including for the treatment of Friedreich's ataxia (64).

We examined compounds that were structurally dissimilar to known KEAP1 inhibitors (Tanimoto similarity < 0.6) but received high scores from the GEMGen scorer (score > 0.6). For those present in the LINCS dataset, we directly retrieved their experimentally measured transcriptomic profiles and ranked the compounds based on the magnitude of induction of NRF2 target genes (**Figure 4B**). As expected, high-scoring molecules ranked above random LINCS compounds (**Figure 4F, Methods**), demonstrating GEMGen's ability to recover and prioritize compounds that recapitulate NRF2-associated transcriptional programs.

For compounds absent from the LINCS library, to determine whether these molecules exhibit *bona fide* NRF2 activating activity, we prioritized commercially available compounds for experimental validation (**Figure 4B**). Among the 2,010 GEMGen-generated compounds absent from LINCS, 87 were associated with registered CAS numbers. After further filtering based on scorer ranking (score > 0.6) and immediate availability, 16 compounds were selected for experimental testing. Using the expression of *NQO1*, a canonical NRF2 target gene, as a functional readout, we identified a chalcone derivative, flavokawain C, that dose-dependently induced *NQO1* expression in K562 cells (EC_{50}=6.3 μM; **Figure 4G-H**). RNA sequencing followed by gene set enrichment analysis confirmed that flavokawain C upregulated a broad set of NRF2 target genes to a degree comparable to the known NRF2 activator bardoxolone (**Figure 4I**). Gene ontology analysis further supported activation of an antioxidative transcriptional program (**Figure 4K-L**). Additionally, flavokawain C treatment in HEK293T cells increased NRF2 protein abundance, suggestive of stabilization of NRF2 through reduced KEAP1-mediated

ubiquitination and degradation (**Figure 4J**).

Interestingly, flavokawain C possesses a chalcone scaffold containing a Michael acceptor motif known to ligand KEAP1 (65,66). The identification and experimental validation of flavokawain C therefore suggest that GEMGen can generalize on the basis of mechanistic features across perturbation modalities.

GEMGen generates chemical perturbations reversing fibrotic state transition

We next asked whether inverse chemical perturbation design could be applied to a disease-associated cellular state transition. Fibrosis is characterized by persistent activation of fibroblasts into myofibroblasts and a coordinated transcriptional program involving extracellular-matrix deposition and tissue remodeling (67). We therefore used the transition from the fibrotic state toward the quiescent state as an inverse-design objective for GEMGen (**Figure 5A and Figure S7C**), which is represented by downregulation of fibrotic transcriptional signatures including *ACTA2, FN1, COL1A1 and TGFB1.*

We categorized GEMGen-generated candidates into three classes: compounds with previously reported antifibrotic activity, known structures lacking established antifibrotic effects, and novel structures without close analogs in existing collections that would require *de novo* synthesis for testing. Among the first class, GEMGen recovered established antifibrotic agents such as SB-525334 and sorafenib (68), supporting the biological relevance of the fibrosis-conditioned outputs (**Figure 5A**). We next focused on the commercially available molecules without documented antifibrotic activities. From this group, we selected 19 molecules for testing in TGFβ-stimulated LX-2 cells, a human hepatic stellate cell line widely used in fibrosis research. Among these molecules, we identified HG-9-91-01, a salt-inducible kinase inhibitor, and indirubin-3′-monoxime, a dual GSK3β and CDK1/5 inhibitor (**Figure 5A**), which dose-dependently suppressed the expression of *FN1* in TGFβ-stimulated LX-2 cells (**Figure 5B**). Consistent with these transcriptional effects, treatment with either compound markedly reduced TGFβ-induced fibronectin deposition, indicating attenuation of fibroblast activation at the phenotypic level (**Figure 5C**). As antifibrotic activities of HG-9-91-01 and indirubin-3′-monoxime had not been reported previously, this result demonstrates the ability of GEMGen to repurpose existing compounds capable of producing the desired response from a transcriptomic state-transition objective.

GEMGen generates novel chemical perturbations with antifibrotic activity

Encouraged by these results, we asked whether GEMGen could generate previously unenumerated structures with antifibrotic activity. We retained compounds that were absent from the LINCS collection and ZINC20 collection of 1.9 billion molecules (69). These candidates were then ranked by GEMGen scores and synthetic accessibility, prioritizing structures predicted to be active and practically synthesizable. From this subset,

we selected 17 molecules for *de novo* synthesis. These compounds are dissimilar to molecules in both ZINC and LINCS (**Figure S5**), indicating that GEMGen can propose chemotypes that extend beyond commonly sampled regions of immense chemical space. Among all the *de novo* synthesized molecules, eight exhibited measurable antifibrotic effects, as demonstrated by dose-dependent reductions in *FN1* and *Acta2* transcript levels in TGFβ-stimulated LX-2 and NIH-3T3 fibroblasts, respectively (**Figure 5D and Figure S6A-F**). Bemis-Murcko scaffold analysis against the ChemDiv BMS diversity set showed that most active molecules adopt novel core scaffolds and topologies, with only one residing on a known scaffold, supporting their structural novelty relative to conventional commercial collections (**Figure 5D**) (70).

We next focused on GEMGen-010, a quinazoline-2,4-diamine derivative, and GEMGen-012, a methyleneoxindole derivative, which showed strong antifibrotic activities (**Figure 5D**). Scaffold analysis confirmed that both compounds are based on cores that diverge substantially from their nearest known scaffolds (**Figure 5E**). In TGFβ-stimulated LX-2 and NIH-3T3 cells, they reduced *FN1* and *Acta2* transcript levels with sub-micromolar potency, respectively (**Figure 5F-G**), and attenuated FN1 protein levels (**Figure 5H**). To determine whether these effects reflected suppression of isolated markers or broader modulation of the fibrotic program, we performed RNA sequencing of TGFβ-stimulated LX-2 cells treated with GEMGen-010 or GEMGen-012. Transcriptomic profiling revealed coordinated downregulation of multiple fibrosis-associated gene modules, including extracellular matrix core and remodeling genes, and TGFβ/SMAD–EMT signaling components, with global expression patterns shifting toward the quiescent baseline (**Figure S7A-B and Table S4**). Additionally, both molecules reduced the expression of the genes used as prompts in the fibrosis-associated transcriptional signature (**Figure S7C-D**), further supporting the fidelity of GEMGen's inverse-design objective. These results indicate that GEMGen-010 and GEMGen-012 do not merely reduce individual marker expression but instead broadly dampen the underlying fibrotic transcriptional program.

Upon structural analysis, GEMGen-010 was identified as a quinazoline-2,4-diamine derivative whose close analogs, to the best of our knowledge, have not been associated with antifibrotic activity. Although 4-aminoquinazolines constitute a privileged scaffold in kinase inhibitor design, as exemplified by gefitinib and erlotinib (71), substitution with an extended moiety at the 2-position has been minimally explored. We hypothesized that the elongated 2-substituent of GEMGen-010 allows productive engagement with the allosteric back pocket of its kinase target via a type II, DFG-out binding mode (72,73), thereby conferring high target specificity in line with its sub-micromolar cellular potency.

GEMGen-010 alleviates liver fibrosis *in vivo*

On the basis of its *in vitro* activity and structural novelty, we sought to further evaluate the *in vivo* efficacy of GEMGen-010 in a tissue fibrosis model. We first characterized its pharmacokinetic profile to determine the appropriate route and dose of administration.

Whereas oral dosing at 20 mg/kg achieved poor systemic exposure ($AUC_{0\text{-}24h}$ of 83 ± 10 ng·h/mL), intraperitoneal administration provided sufficient exposure ($AUC_{0\text{-}24h}$ of 4088 ± 1119 ng·h/mL), comparable to the concentrations required for cellular activity (**Figure 5E and 6A-C**). We therefore selected the intraperitoneal route for subsequent efficacy studies.

We employed a carbon tetrachloride (CCl_4)-induced mouse model of liver fibrosis, a well-established model of chronic hepatic injury and fibrosis. Mice received CCl_4 twice weekly for four weeks, and GEMGen-010 (50 mg/kg/day) was administered once daily from week 3 for two weeks (**Figure 6D**). Histological analysis by Sirius Red staining revealed extensive collagen deposition in CCl_4-treated livers, which was markedly attenuated by GEMGen-010 (**Figure 6E-F**). Consistently, GEMGen-010–treated livers displayed a smoother surface with fewer fibrotic nodules than vehicle controls on gross examination (**Figure 6G**). In addition, GEMGen-010 treatment effectively lowered the elevated serum activities of alanine aminotransferase (**Figure 6H**) and aspartate aminotransferase (**Figure 6I**), indicating attenuated hepatocellular injury. At the transcriptional level, GEMGen-010 suppressed the CCl_4-driven upregulation of the fibrosis-associated genes *Acta2* and *Col1a1* in the liver (**Figure 6J-K**), recapitulating the antifibrotic activity observed *in vitro*.

Together, these results demonstrate that chemical perturbations generated from a desired disease-associated state transition can produce the intended response across molecular, cellular and tissue-level readouts. The *in vivo* efficacy of GEMGen-010 further provides prospective evidence that inverse perturbation design can yield experimentally actionable molecules beyond predefined chemical collections, providing a direct route from perturbation design to therapeutic discovery in disease-relevant settings.

Discussion

In this study, we introduce GEMGen, a generative framework for inverse perturbation design conditioned on desired transcriptomic state transitions. Across retrospective benchmarks and prospective studies of genetic perturbation-guided NRF2 activation and fibrosis reversal, GEMGen recovered known perturbations, identified mechanistically related alternatives and generated previously unenumerated structures with experimentally validated activities. The activity of *de novo*-designed molecules in cellular and *in vivo* settings demonstrates that inverse perturbation design can yield experimentally validated interventions rather than merely reconstructing compounds represented in existing datasets.

The inverse-design formulation complements rapid progress in forward perturbation modeling. Large-scale perturbational datasets and increasingly capable computational models, including emerging AI virtual cell frameworks, are improving our ability to predict how genetic and chemical interventions reshape cellular states across contexts. Such forward models answer how a biological system is expected to respond to a specified intervention. In many settings, however, the desired response is known while the

intervention is not. GEMGen addresses this complementary endpoint by treating the desired state transition itself as the design objective.

A central methodological challenge is how to represent cellular state transitions in a form that supports inverse molecular generation across heterogeneous perturbation datasets. Whereas many task-specific generators operate on continuous expression vectors or separately aligned transcriptomic and molecular embeddings, GEMGen converts differential gene expression into ranked gene lists and represents them as text-based instructions for an autoregressive scientific language model. This abstraction reduces dependence on dataset-specific expression magnitudes while preserving the direction of transcriptional change, consistent with the improved cross-batch consistency observed in our analyses. It also places biological entities, perturbation responses and molecular structures within a shared language-model framework. Together, these properties provide a practical interface between high-dimensional transcriptomic responses and autoregressive molecular generation, and may contribute to GEMGen's generalization across bulk and single-cell datasets as well as to genetic perturbation inputs.

A second challenge for inverse perturbation design is whether a response-structure relationship learned from observed perturbations can generalize to previously unenumerated molecules. GEMGen formulates this problem as conditional generation between two structured domains: rank-ordered transcriptional changes that specify the desired response and molecular strings that encode chemical structure. Training across diverse perturbation-response pairs allows the generator to learn statistical regularities linking transcriptional programs to chemical structures, rather than simply retrieving compounds associated with similar responses. Autoregressive generation can then use these learned relationships to propose previously unseen structures conditioned on the target transition, while the response-matching scorer provides an additional measure of response compatibility. We tested this generalization prospectively in the fibrosis setting, where eight of 17 synthesized candidates showed antifibrotic activity, including a *de novo* synthesized molecule that demonstrated efficacy *in vivo*. These results provide direct evidence that the learned response-structure mapping can yield biologically active molecules beyond the compounds observed during training.

The experimental results also highlight that successful inverse design does not require a unique molecular solution. In the benchmark datasets, GEMGen recovered both benchmark compounds and mechanistically related alternatives. For genetic perturbation inputs, high response-matching scores were not always associated with structural similarity to known inhibitors, consistent with the possibility that distinct chemical mechanisms can converge on related cellular states. The identification of flavokawain C as an NRF2 pathway activator provides an experimental example of such functional convergence. Similarly, the fibrosis application yielded both existing compounds with previously unreported antifibrotic activities and de novo structures that shifted the broader

fibrosis-associated transcriptional program toward the desired state. These observations support evaluating inverse-design models by biological response rather than by structural recovery alone.

The ability to generate chemical perturbations from CRISPR-induced transcriptional states further suggests that transcriptomic changes can provide a measurable, shared interface across perturbation modalities, because both genetic and chemical interventions can be represented by the cellular state transitions they induce. Although the present implementation generates chemical perturbations, the inverse-design formulation is not inherently restricted to a single output modality: with appropriate perturbation-response data and modality-specific output representations, it could in principle be reversed to generate genetic perturbations predicted to produce a desired state, or generalized to translate between other pairs of perturbation modalities. Foundation models trained across molecular and genetic entities may make such cross-modal generation particularly tractable by supplying shared representations and forward priors across intervention types.

Despite the advantages of the GEMGen framework, several limitations of the present study should be acknowledged. First, GEMGen relies on transcriptomic state transitions as its design objective. Transcriptomic similarity provides a high-dimensional representation of perturbation response but does not fully capture downstream functional, physiological or safety-related outcomes, and transcriptional changes associated with a desired state are not necessarily causal. Incorporating multimodal molecular measurements, functional phenotypes and explicit causal information should therefore improve the specification of design objectives. Second, perturbation responses are strongly dependent on cellular context, dose and time. Although GEMGen generalizes across several experimental settings, more explicit modeling of these variables and of response dynamics will be important for broader intervention design.

Third, the accessible chemical space remains shaped by the diversity of perturbation-response pairs available for training. Continued expansion of high-throughput perturbation atlases, together with iterative experimental feedback, should improve coverage and out-of-distribution generation. More broadly, extending inverse design beyond transcriptomic response representations and across intervention modalities may ultimately shift biological modeling from predicting state transitions to identifying routes toward desired biological states.

References

1. Dimitrov D, Schrod S, Rohbeck M, Stegle O. Interpretation, extrapolation and perturbation of single cells. Nat Rev Genet. 2026 May;27(5):349–70. doi:10.1038/s41576-025-00920-4

2. Trapnell C. Revealing gene function with statistical inference at single-cell resolution. Nat Rev Genet. 2024 Sep;25(9):623–38. doi:10.1038/s41576-024-00750-w

3. Stockwell BR. Chemical genetics: ligand-based discovery of gene function. Nat Rev Genet. 2000 Nov;1(2):116–25. doi:10.1038/35038557

4. Jones LH, Bunnage ME. Applications of chemogenomic library screening in drug discovery. Nat Rev Drug Discov. 2017 Apr;16(4):285–96. doi:10.1038/nrd.2016.244

5. Bock C, Datlinger P, Chardon F, Coelho MA, Dong MB, Lawson KA, et al. High-content CRISPR screening. Nat Rev Methods Primer. 2022 Feb 10;2(1):8. doi:10.1038/s43586-021-00093-4

6. Hartwell LH, Culotti J, Reid B. Genetic Control of the Cell-Division Cycle in Yeast, I. Detection of Mutants. Proc Natl Acad Sci. 1970 Jun;66(2):352–9. doi:10.1073/pnas.66.2.352

7. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature. 1998 Feb;391(6669):806–11. doi:10.1038/35888

8. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, et al. Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. Cell. 2007 Nov 30;131(5):861–72. doi:10.1016/j.cell.2007.11.019

9. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors: Cell

10. Hou P, Li Y, Zhang X, Liu C, Guan J, Li H, et al. Pluripotent Stem Cells Induced from Mouse Somatic Cells by Small-Molecule Compounds. Science. 2013 Aug 9;341(6146):651–4. doi:10.1126/science.1239278

11. Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell. 2016 Dec 15;167(7):1853-1866.e17. doi:10.1016/j.cell.2016.11.038

12. Adamson B, Norman TM, Jost M, Cho MY, Nuñez JK, Chen Y, et al. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. Cell. 2016 Dec 15;167(7):1867-1882.e21. doi:10.1016/j.cell.2016.11.048

13. Replogle JM, Saunders RA, Pogson AN, Hussmann JA, Lenail A, Guna A, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. Cell. 2022 Jul 7;185(14):2559-2575.e28. doi:10.1016/j.cell.2022.05.013

14. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. Science. 2006 Sep 29;313(5795):1929–35. doi:10.1126/science.1132939

15. Ye C, Ho DJ, Neri M, Yang C, Kulkarni T, Randhawa R, et al. DRUG-seq for miniaturized high-throughput transcriptome profiling in drug discovery. Nat Commun. 2018 Oct 17;9(1):4307. doi:10.1038/s41467-018-06500-x

16. Zhang J, Ubas AA, Borja R de, Svensson V, Thomas N, Thakar N, et al. Tahoe-100M: A Giga-Scale Single-Cell Perturbation Atlas for Context-Dependent Gene Function and Cellular Modeling [Internet]. bioRxiv; 2025 [cited 2026 Aug 31]. p. 2025.02.20.639398. Available from: https://www.biorxiv.org/content/10.1101/2025.02.20.639398v1 doi:10.1101/2025.02.20.639398

17. Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. Nat Methods. 2019 Aug;16(8):715–21. doi:10.1038/s41592-019-0494-8

18. Roohani Y, Huang K, Leskovec J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. Nat Biotechnol. 2024 Jun;42(6):927–35. doi:10.1038/s41587-023-01905-6

19. Tong X, Qu N, Kong X, Ni S, Zhou J, Wang K, et al. Deep representation learning of chemical-induced transcriptional profile for phenotype-based drug discovery. Nat Commun. 2024 Jun 25;15(1):5378. doi:10.1038/s41467-024-49620-3

20. Wenkel F, Tu W, Masschelein C, Shirzad H, Hodgson L, Bendidi I, et al. TxPert: using multiple knowledge graphs for prediction of transcriptomic perturbation effects. Nat Biotechnol. 2026 May 1;1–8. doi:10.1038/s41587-026-03113-4

21. Zhu J, Wang J, Wang X, Gao M, Guo B, Gao M, et al. Prediction of drug efficacy from transcriptional profiles with deep learning. Nat Biotechnol. 2021 Nov;39(11):1444–52. doi:10.1038/s41587-021-00946-z

22. Li Y, Zeng M, Zhu J, Liu L, Wang F, Huang L, et al. UniPert-G2CP bridges genetic and chemical screens from molecular representation to phenotype modeling. Cell. 2026 Aug 6;189(16):4946-4963.e8. doi:10.1016/j.cell.2026.06.005

23. Bunne C, Roohani Y, Rosen Y, Gupta A, Zhang X, Roed M, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. Cell. 2024 Dec 12;187(25):7045–63. doi:10.1016/j.cell.2024.11.015

24. Predicting cellular responses to perturbation across diverse contexts with State | bioRxiv [Internet]. [cited 2026 Aug 31]. Available from: https://www.biorxiv.org/content/10.1101/2025.06.26.661135v1

25. Hao M, Gong J, Zeng X, Liu C, Guo Y, Cheng X, et al. Large-scale foundation model on single-cell transcriptomics. Nat Methods. 2024 Aug;21(8):1481–91. doi:10.1038/s41592-024-02305-7

26. Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods. 2024 Aug;21(8):1470–80. doi:10.1038/s41592-024-02201-0

27. Rood JE, Hupalowska A, Regev A. Toward a foundation model of causal cell and tissue biology with a Perturbation Cell and Tissue Atlas. Cell. 2024 Aug 22;187(17):4520–45. doi:10.1016/j.cell.2024.07.035

28. Kholodenko BN, Kolch W, Rukhlenko OS. Reversing pathological cell states: the road less travelled can extend the therapeutic horizon. Trends Cell Biol. 2023 Nov;33(11):913–23. doi:10.1016/j.tcb.2023.04.004

29. Huang ME, Ye YC, Chen SR, Chai JR, Lu JX, Zhoa L, et al. Use of all-trans retinoic acid in the treatment of acute promyelocytic leukemia. Blood. 1988 Aug;72(2):567–72.

30. Kisseleva T, Cong M, Paik Y, Scholten D, Jiang C, Benner C, et al. Myofibroblasts revert to an inactive phenotype during regression of liver fibrosis. Proc Natl Acad Sci U S A. 2012 Jun 12;109(24):9448–53. doi:10.1073/pnas.1201840109

31. Lian X, Hsiao C, Wilson G, Zhu K, Hazeltine LB, Azarin SM, et al. Robust cardiomyocyte differentiation from human pluripotent stem cells via temporal modulation of canonical Wnt signaling. Proc Natl Acad Sci U S A. 2012 Jul 3;109(27):E1848-1857. doi:10.1073/pnas.1200250109

32. Pagliuca FW, Millman JR, Gürtler M, Segel M, Van Dervort A, Ryu JH, et al. Generation of functional human pancreatic β cells in vitro. Cell. 2014 Oct 9;159(2):428–39. doi:10.1016/j.cell.2014.09.040

33. Black JB, Adler AF, Wang HG, D'Ippolito AM, Hutchinson HA, Reddy TE, et al. Targeted Epigenetic Remodeling of Endogenous Loci by CRISPR/Cas9-Based Transcriptional Activators Directly Converts Fibroblasts to Neuronal Cells. Cell Stem Cell. 2016 Sep 1;19(3):406–14. doi:10.1016/j.stem.2016.07.001

34. Wang G, Chow RD, Bai Z, Zhu L, Errami Y, Dai X, et al. Multiplexed activation of endogenous genes by CRISPRa elicits potent antitumor immunity. Nat Immunol. 2019 Nov;20(11):1494–505. doi:10.1038/s41590-019-0500-4

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

36. Moffat JG, Vincent F, Lee JA, Eder J, Prunotto M. Opportunities and challenges in phenotypic drug discovery: an industry perspective. Nat Rev Drug Discov. 2017 Aug;16(8):531–43. doi:10.1038/nrd.2017.111

37. Pham TH, Qiu Y, Zeng J, Xie L, Zhang P. A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing. Nat Mach Intell. 2021 Mar;3(3):247–57. doi:10.1038/s42256-020-00285-9

38. DeMeo B, Nesbitt C, Miller SA, Burkhardt DB, Lipchina I, Fu D, et al. Active learning framework leveraging transcriptomics identifies modulators of disease phenotypes. Science. 2025 Oct 23;390(6776):eadi8577. doi:10.1126/science.adi8577

39. Xing J, Tan M, Leshchiner D, Sun M, Abdelgied M, Huang L, et al. Deep-learning-based de novo discovery and design of therapeutics that reverse disease-associated transcriptional phenotypes. Cell. 2026 Apr 30;189(9):2556-2572.e19. doi:10.1016/j.cell.2026.02.016 PubMed PMID: 41850287.

40. Reymond JL. The chemical space project. Acc Chem Res. 2015 Mar 17;48(3):722–30. doi:10.1021/ar500432k

41. Bohacek RS, McMartin C, Guida WC. The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev. 1996 Jan;16(1):3–50. doi:10.1002/(SICI)1098-1128(199601)16:1

42. Kirkpatrick P, Ellis C. Chemical space. Nature. 2004 Dec 1;432(7019):823–823. doi:10.1038/432823a

43. Kaitoh K, Yamanishi Y. TRIOMPHE: Transcriptome-Based Inference and Generation of Molecules with Desired Phenotypes by Machine Learning. J Chem Inf Model. 2021 Sep 27;61(9):4303–20. doi:10.1021/acs.jcim.1c00967

44. Liu Y, Yu H, Duan X, Zhang X, Cheng T, Jiang F, et al. TransGEM: a molecule generation model based on Transformer with gene expression data. Bioinformatics. 2024 May 2;40(5):btae189. doi:10.1093/bioinformatics/btae189

45. Méndez-Lucio O, Baillif B, Clevert DA, Rouquié D, Wichard J. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. Nat Commun. 2020 Jan 3;11(1):10. doi:10.1038/s41467-019-13807-w

46. Das D, Chakrabarty B, Srinivasan R, Roy A. Gex2SGen: Designing Drug-like Molecules

from Desired Gene Expression Signatures. J Chem Inf Model. 2023 Apr 10;63(7):1882–93. doi:10.1021/acs.jcim.2c01301

47. Cheng J, Pan X, Fang Y, Yang K, Xue Y, Yan Q, et al. GexMolGen: cross-modal generation of hit-like molecules via large language model encoding of gene expression signatures. Brief Bioinform. 2024 Sep 23;25(6):bbae525. doi:10.1093/bib/bbae525

48. Born J, Manica M, Oskooei A, Cadow J, Markert G, Rodríguez Martínez M. PaccMannRL: De novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning. iScience. 2021 Apr 23;24(4):102269. doi:10.1016/j.isci.2021.102269

49. Chen Y, Zou J. Simple and effective embedding model for single-cell biology built from ChatGPT. Nat Biomed Eng. 2025 Apr;9(4):483–93. doi:10.1038/s41551-024-01284-6

50. Rizvi SA, Levine D, Patel A, Zhang S, Wang E, Perry CJ, et al. Scaling Large Language Models for Next-Generation Single-Cell Analysis. bioRxiv: The Preprint Server for Biology; 2026. doi:10.1101/2025.04.14.648850

51. Xia Y, Jin P, Xie S, He L, Cao C, Luo R, et al. Nature Language Model: Deciphering the Language of Nature for Scientific Discovery [Internet]. arXiv; 2025 [cited 2026 Aug 31]. Available from: http://arxiv.org/abs/2502.07527 doi:10.48550/arXiv.2502.07527

52. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017 Nov 30;171(6):1437-1452.e17. doi:10.1016/j.cell.2017.10.049

53. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. ZINC: A Free Tool to Discover Chemistry for Biology. J Chem Inf Model. 2012 May 15;52(7):1757–68. doi:10.1021/ci3001277

54. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 2001 Mar 1;46(1–3):3–26. doi:10.1016/s0169-409x(00)00129-0

55. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs. Nat Chem. 2012 Jan 24;4(2):90–8. doi:10.1038/nchem.1243

56. Ertl P, Schuffenhauer A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. J Cheminformatics. 2009 Jun 10;1(1):8. doi:10.1186/1758-2946-1-8

57. Jin W, Barzilay R, Jaakkola T. Junction Tree Variational Autoencoder for Molecular Graph Generation [Internet]. arXiv; 2019 [cited 2026 Aug 31]. Available from:

http://arxiv.org/abs/1802.04364 doi:10.48550/arXiv.1802.04364

58. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci. 2018 Feb 28;4(2):268–76. doi:10.1021/acscentsci.7b00572

59. Rossari F, Minutolo F, Orciuolo E. Past, present, and future of Bcr-Abl inhibitors: from chemical development to clinical efficacy. J Hematol OncolJ Hematol Oncol. 2018 Jun 20;11(1):84. doi:10.1186/s13045-018-0624-2

60. Hennequin LF, Allen J, Breed J, Curwen J, Fennell M, Green TP, et al. N-(5-chloro-1,3-benzodioxol-4-yl)-7-[2-(4-methylpiperazin-1-yl)ethoxy]-5- (tetrahydro-2H-pyran-4-yloxy)quinazolin-4-amine, a novel, highly selective, orally available, dual-specific c-Src/Abl kinase inhibitor. J Med Chem. 2006 Nov 2;49(22):6465–88. doi:10.1021/jm060434q

61. Brasca MG, Albanese C, Alzani R, Amici R, Avanzi N, Ballinari D, et al. Optimization of 6,6-dimethyl pyrrolo[3,4-c]pyrazoles: Identification of PHA-793887, a potent CDK inhibitor suitable for intravenous dosing. Bioorg Med Chem. 2010 Mar 1;18(5):1844–53. doi:10.1016/j.bmc.2010.01.042 PubMed PMID: 20153204.

62. Jalencas X, Mestres J. On the origins of drug polypharmacology. MedChemComm. 2013 Jan 1;4(1):80–7. doi:10.1039/c2md20242e

63. Cuadrado A, Rojo AI, Wells G, Hayes JD, Cousin SP, Rumsey WL, et al. Therapeutic targeting of the NRF2 and KEAP1 partnership in chronic diseases. Nat Rev Drug Discov. 2019 Apr;18(4):295–317. doi:10.1038/s41573-018-0008-x

64. Pilotto F, Chellapandi DM, Puccio H. Omaveloxolone: a groundbreaking milestone as the first FDA-approved drug for Friedreich ataxia. Trends Mol Med. 2024 Feb;30(2):117–25. doi:10.1016/j.molmed.2023.12.002

65. Kumar V, Kumar S, Hassan M, Wu H, Thimmulappa RK, Kumar A, et al. Novel chalcone derivatives as potent Nrf2 activators in mice and human lung epithelial cells. J Med Chem. 2011 Jun 23;54(12):4147–59. doi:10.1021/jm2002348

66. Bollong MJ, Lee G, Coukos JS, Yun H, Zambaldo C, Chang JW, et al. A metabolite-derived protein modification integrates glycolysis with KEAP1-NRF2 signalling. Nature. 2018 Oct;562(7728):600–4. doi:10.1038/s41586-018-0622-0

67. Rockey DC, Bell PD, Hill JA. Fibrosis--a common pathway to organ injury and failure. N Engl J Med. 2015 Mar 19;372(12):1138–49. doi:10.1056/NEJMra1300575

68. Yuan S, Wei C, Liu G, Zhang L, Li J, Li L, et al. Sorafenib attenuates liver fibrosis by

triggering hepatic stellate cell ferroptosis via HIF-1α/SLC7A11 pathway. Cell Prolif. 2022 Jan;55(1):e13158. doi:10.1111/cpr.13158

69. Irwin JJ, Tang KG, Young J, Dandarchuluun C, Wong BR, Khurelbaatar M, et al. ZINC20-A Free Ultralarge-Scale Chemical Database for Ligand Discovery. J Chem Inf Model. 2020 Dec 28;60(12):6065–73. doi:10.1021/acs.jcim.0c00675 P

70. Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. J Med Chem. 1996 Jul 19;39(15):2887–93. doi:10.1021/jm9602928

71. Recent advancements of 4-aminoquinazoline derivatives as kinase inhibitors and their applications in medicinal chemistry - PubMed [Internet]. [cited 2026 Aug 31]. Available from: https://pubmed.ncbi.nlm.nih.gov/30878832/

72. Exploration of type II binding mode: A privileged approach for kinase inhibitor focused drug discovery? - PubMed [Internet]. [cited 2026 Aug 31]. Available from: https://pubmed.ncbi.nlm.nih.gov/24730530/

73. Vijayan RSK, He P, Modi V, Duong-Ly KC, Ma H, Peterson JR, et al. Conformational analysis of the DFG-out kinase motif and biochemical profiling of structurally validated type II inhibitors. J Med Chem. 2015 Jan 8;58(1):466–79. doi:10.1021/jm501603h

74. Zhang W, Guo Z, Xia Y, Jin P, Xie S, Qin T, et al. MolChord: Structure-Sequence Alignment for Protein-Guided Drug Design [Internet]. arXiv; 2025 [cited 2026 Aug 31]. Available from: http://arxiv.org/abs/2510.27671 doi:10.48550/arXiv.2510.27671

75. Duan Q, Reid SP, Clark NR, Wang Z, Fernandez NF, Rouillard AD, et al. L1000CDS2: LINCS L1000 characteristic direction signatures search engine. Npj Syst Biol Appl. 2016 Aug 4;2(1):16015. doi:10.1038/npjsba.2016.15

76. Clark NR, Hu KS, Feldmann AS, Kou Y, Chen EY, Duan Q, et al. The characteristic direction: a geometrical approach to identify differentially expressed genes. BMC Bioinformatics. 2014 Mar 21;15:79. doi:10.1186/1471-2105-15-79

77. Peidli S, Green TD, Shen C, Gross T, Min J, Garda S, et al. scPerturb: harmonized single-cell perturbation data. Nat Methods. 2024 Mar;21(3):531–40. doi:10.1038/s41592-023-02144-y

78. Wu K, Xia Y, Deng P, Liu R, Zhang Y, Guo H, et al. TamGen: drug design with target-aware molecule generation through a chemical language model. Nat Commun. 2024 Oct 29;15(1):9360. doi:10.1038/s41467-024-53632-4

Acknowledgements

Funding: We acknowledge the support from Zhongguancun Academy (Grant No. C20250607 to P.D.), Zhejiang Provincial Key Laboratory Construction Project (Grant No. 2025E10060 to S.S.), Key R&D Program of Zhejiang (QKWL25B0701 to S.S.), Westlake University startup (to S.S.), and the Research Center for Industries of the Future at Westlake University (to S.S.). We thank Dr. Jianyang Zeng (Westlake University) for his critical comments and insightful suggestions on this study.

Author Contributions: S.S., P.D., and Y.X. conceptualization; Q.J., X.Y., P.D., S.S., and Z.L. data curation; Y.X., P.D., P.J., F.J., and H.X. methodology; Q.J., Z.G., and Z.L. model implementation; X.Y., Q.J., P.D., S.S., and J.X. result interpretation; X.Y., S.S., and S.F. cellular experiments; P.D. and S.S. Writing-original draft; S.S., P.D., Q.J., X.Y., Z.G., and Y.X. Writing-revision; R.J., H.L., and T.Q. supervision. All authors have read and approved the manuscript.

Declaration of interests: All authors declare no competing interests. A provisional patent application related to the methods described in this paper has been filed.

Code availability: GEMGen is available on GitHub (https://github.com/DLS5-Omics/GEMGen).

Supplementary Materials

Methods

GEMGen Training

GEMGen is fine-tuned from NatureLM (51) for chemical perturbation generation and response-based scoring. The training was implemented through three complementary instruction-response tasks.

NatureLM foundation model. NatureLM (51) is a unified scientific language model trained on heterogeneous data modalities, including biological sequences (e.g., protein FASTA and molecular SMILES), natural language text, and other cross-modal scientific data. It adopts a decoder-only Transformer architecture, analogous to GPT, and is trained autoregressively by representing all modalities as token sequences within a shared vocabulary.

NatureLM is released in multiple model sizes. Given the limited scale of cell–compound perturbation data available for our downstream tasks, and to reduce the risk of overfitting to compound-specific patterns, we employed the 1B-parameter variant reproduced by Zhang et al. (74). This model consists of 16 Transformer layers with a hidden size of 2,048, a feed-forward network (FFN) dimension of 5,504, and 32 attention heads, including 8 key–value (KV) heads. The vocabulary comprises 129,687 tokens spanning diverse scientific entities and English text, including 1,401 tokens corresponding to unique SMILES symbols. As a general-purpose scientific foundation model, NatureLM provides a strong and flexible initialization for text-guided chemical perturbation generation and compound–response modeling tasks.

Bioentity representation task. This task was implemented through instruction fine-tuning of NatureLM on a diverse mixture of biological data. Descriptive annotations were curated for a wide range of human and mouse cell types and protein-coding genes (detailed data preparation procedures are provided in subsequent sections). Each entry was reformatted into an instruction–response structure with the assistance of GPT-4o. This process yielded 255 unique instruction–response pairs for cell type descriptions and 75,203 for gene descriptions. During training, the two data types were combined and randomly shuffled while maintaining their original distribution ratios. The model was trained using a cross-entropy loss function, consistent with the optimization setup used in NatureLM.

Chemical perturbation generation task. The LINCS chemical perturbation transcriptomic dataset (14), containing paired compound and perturbed-transcriptome information, was used to train the model to generate chemical

structures from perturbation-induced transcriptional responses. Transcriptomic signatures were transformed into text-based ranked lists of differentially expressed genes as input. Training was performed in two phases to balance data coverage and model performance. Phase I was conducted on compound perturbations curated from the original LINCS dataset. The model was trained for two epochs across 16 GPUs using a cross-entropy loss function, a learning rate of 1e-5, and a batch size of 256. Phase II involved continued instruction fine-tuning with the L1000CDS^2 dataset (75), which provides higher-quality measurements and broader gene coverage. This stage was designed to improve generalization across diverse transcriptomic profiling platforms and protocols. Training in Phase II was carried out for 2,475 steps on 8 GPUs with a cross-entropy loss function, a reduced learning rate of 8e-6 and a batch size of 64. Details of data preprocessing and train/test splitting are provided in subsequent sections.

During evaluation, the model was configured to generate 1,000 candidate molecules per target transcriptional response unless otherwise specified. This number is a user-defined sampling parameter and can be adjusted according to practical requirements.

Compound–response matching task. A separate model was fine-tuned to specialize in scoring the correspondence between chemical perturbations and transcriptomic changes, represented as text-based ranked lists of differentially expressed genes. Similar to the compound generation task, training was conducted in two sequential phases using the same data corpus. To adapt the framework for scoring, several task-specific modifications were introduced.

The model was trained to produce a binary response (“yes” or “no”) for each compound–response pair, indicating whether the compound was predicted to reproduce the given transcriptomic change. The training objective combined two complementary losses: a cross-entropy loss for autoregressive language modeling and a binary cross-entropy loss applied to a classification head. Specifically, the final hidden state corresponding to the response token (“yes” or “no”) was passed through a two-layer multilayer perceptron (MLP; hidden dimension = 2,048) for explicit binary classification. During inference, we use the output from the classification head as the final binary prediction.

For scorer training, Phase I was conducted for two epochs on 8 GPUs with a learning rate of 2e-5 and a batch size of 512. Phase II continued for 3,008 steps on 8 GPUs using a reduced learning rate of 8e-6 and a batch size of 128.

Datasets

LINCS L1000 dataset. We retrieved the Level 3 transcriptomic profiles and associated metadata from the LINCS dataset of the CMap LINCS (Library of Integrated Network-Based Cellular Signatures) resource (https://clue.io/data/CMap2020#LINCS2020, December 2020 release). In our study, we focused exclusively on compound perturbation profiles. The Level 3 data contain quantile-normalized expression values for 978 landmark genes (experimentally measured) and an additional 11,350 genes computationally inferred. For our task, we used only the measured expression values of the 978 landmark genes.

To compute differential gene expression (DGE) signatures, we applied the Characteristic Direction (CD) method (76). CD is a multivariate geometrical approach that quantifies expression change levels and computes statistical significance for each gene between two groups of transcriptomic profiles. Specifically, we grouped samples based on perturbation metadata-including cell type, compound, time, dose, and plate ID. Samples sharing identical perturbation conditions on the same plate were considered replicates. We computed DGE signatures for each such group against the matched vehicle-treated controls on the same plate to minimize confounding from plate-specific technical variability.

For each cell type–compound pair (hereafter referred to as a combination), multiple transcriptomic experiments may be available, conducted under different time points, doses, or plates. We refer to all such experiments within a given combination as replicates, regardless of the specific experimental settings. To ensure data quality and prevent overrepresentation of certain combinations, we retained two DGE signatures per combination, selected based on internal consistency measured by cosine similarity:

- For combinations with 2 to 20 replicates (54.6% of all combinations), we computed pairwise cosine similarities among all DGE signatures and selected the two with the highest average similarity to the rest.
- For combinations with more than 20 replicates (1.2%), we first applied Leiden clustering to group similar DGE signatures, then calculated intra-cluster cosine similarities. The two most consistent DGE signatures were selected from the cluster with the highest mean internal similarity.
- For combinations with only one replicate (44.2%), we generated five synthetic control sets by repeatedly sampling control cells (matched in size to the

perturbed group) from the same plate. We computed five CD-based DGE signatures and selected the two with the highest pairwise similarity.

This procedure yielded a curated dataset of 183,440 unique cell type–compound combinations, each represented by two high-consistency DGE signatures, spanning 27,770 compounds and 213 cell types.

L1000CDS² dataset. L1000CDS² is a publicly available database of precomputed CD-derived DGE signatures derived from the LINCS dataset (75). We downloaded 119,155 DGE signatures from the official L1000CDS² database (https://maayanlab.cloud/public/L1000CDS_download/), and filtered entries with compound perturbations and adjusted *p*-values less than 0.1, resulting in 32,809 high-confidence records. These records span 63 human cell lines and 3,872 small molecules.

Compared to the CD-derived DGE signatures we computed from the LINCS dataset, the L1000CDS² signatures include both landmark and inferred genes, thereby offering broader transcriptomic coverage. This expanded gene space is particularly valuable for model training in scenarios where the downstream applications (e.g., single-cell or other non-plate-based sequencing data) are not restricted to landmark genes alone.

Train/test split. To evaluate the model's ability to generalize to previously unseen chemical space, we constructed a novel compound test set based on the LINCS and L1000CDS² datasets. The key criterion was that no compound in this test set exhibits a Tanimoto similarity greater than 0.8 to any compound used during training.

To achieve this, we first pooled all small molecules from the filtered LINCS set and the L1000CDS² dataset into a unified compound collection and computed the pairwise Tanimoto similarity matrix across all molecules. For each compound in L1000CDS², we identified its maximum similarity to any other compound in the combined set. Compounds whose maximum similarity was below 0.8 and that originated from L1000CDS² were retained as candidates for the novel compound test set. The resulting novel compound test set comprises 404 samples.

Tahoe-100M dataset. The recently released Tahoe-100M dataset (16) captures the transcriptional responses of 50 cancer cell lines to 335 compound perturbations using the Mosaic platform. We downloaded the full dataset and associated metadata from the official Tahoe-100M repository hosted on Hugging Face [https://huggingface.co/datasets/tahoebio/Tahoe-100M].

For data preprocessing, we applied the following strategy: First, cells were filtered using quality-control criteria, including a mitochondrial gene fraction below 20%, more than 50 expressed genes, and restriction to the G1 phase of the cell cycle. Second, only cells treated with a 5 μM drug dose were retained to ensure consistency across perturbation conditions. Third, for each cell type–compound combination, we randomly sampled 100 cells before and 100 cells after perturbation. Fourth, we computed the E-distance between perturbed and control cells (77), which quantifies the effect size of perturbation in high-dimensional expression space. We retained combinations with E-distance > 500, resulting in 640 high-confidence cell type–compound combinations with strong perturbation effects. Finally, we calculated CD-derived DGE signatures for each selected combination for model evaluation.

Mechanistic analogs. To expand the list of ground truth molecules for evaluating the ability of GEMGen to recover mechanistically related but structurally distinct molecules, we selected 30 and 39 ground truth molecules from LINCS and Tahoe-100M test sets, respectively. We then manually curated a list of mechanistic analogs for each molecule based on potency and selectivity. A full list of mechanistic analogs is available in **Table S1**.

K562 CRISPR perturbation dataset. The CRISPR perturbation dataset for the K562 cell line used in this study was derived from the experiments conducted by Replogle et al. (13). We downloaded the scRNA-seq dataset of the K562 cell line before and after different gene edits from the website https://gwps.wi.mit.edu/. Following the analysis described in the original paper, we selected 21 genetic perturbations that induce significant transcriptomic changes and identified the corresponding sets of upregulated and downregulated genes. Under the assumption that chemical inhibition of a protein phenocopies the effects of genetic knockdown or knockout, we manually curated a list of small molecule inhibitors of each target from the ChEMBL database, restricted to those characterized by quantitative, protein-based assays. To preclude low-activity inhibitors, only high-potency inhibitors with IC50 values below 500 nM were retained as ground-truth molecules for analysis. A full list of small molecules for each genetic perturbation is available in **Table S2**.

Cell type descriptive annotations. We curated descriptive annotations for all cell types used in this study by combining two sources: descriptions retrieved from the ATCC cell catalog, and independently generated summaries by GPT-4o, which were manually reviewed to ensure biological plausibility. In total, we compiled 255 cell type descriptions, including 42 primary cells and 213 immortalized cell lines.

Gene descriptive annotations. We retrieved human and mouse protein-coding gene list from the NCBI database (https://ftp.ncbi.nlm.nih.gov/gene/DATA/; version dated

April 2025). For each gene, we extracted descriptive summaries from the RefSeq database using the official API. In total, we collected annotations for 20,786 human genes and 18,756 mouse genes.

Instruction-response pair construction

Perturbation-response pair construction. To support both training and inference for the compound generator and the scoring module, we constructed instruction–response pairs derived from compound–response pairs.

From each DGE signature, we derive upregulated and downregulated gene list based on the CD vector, which encodes the magnitude and direction of gene differential expressions. Specifically, we selected the top 50/40/30 genes with the largest absolute CD values and then assigned each gene to the upregulated or downregulated category based on the sign of its value (positive or negative). This yielded two directional gene sets per DGE signature.

Given that many small molecules appear across multiple cell type–compound combinations, we introduced molecular representation diversity by generating alternative SMILES strings for each compound using the SMILES augmentation method implemented in TamGen (78) [https://github.com/microsoft/TamGen]. This augmentation enriches the chemical space explored during generative modeling and helps prevent overfitting to SMILES forms in the training dataset.

For both the generator and the scorer, the input and output take the form of natural language instructions and responses, allowing interface with natural language prompts in instruction-based generation and evaluation workflows.

Cell type and genes. To construct instruction–response pairs for cell types, we first manually designed a base instruction template (for example, *"Describe the cell type [cell type]"*). To increase linguistic diversity and reduce structural uniformity across instructions, we then used GPT-4o to generate paraphrased variants of this template (for example, we prompted GPT-4o with: *"Here is a prompt for generating cell type descriptions: 'Describe the cell type [cell type].' Please generate 20 similar sentence templates based on this example."*). For each cell type, we randomly selected one instruction from the augmented template pool. The corresponding response was the descriptive annotation of the cell type, as curated from ATCC and GPT-4o (see the previous section).

We followed a similar procedure for genes, except that we generated 50 instruction templates in total. Additionally, because gene names often have multiple synonyms,

we incorporated those synonyms when constructing instruction-response pairs. This resulted in a total of 75,203 distinct instruction–response pairs for gene entities.

The full list of instruction templates is provided in **Table S5**.

Cross-batch consistency analysis

To examine whether text-based representation improves robustness to technical variation, we analyzed Plate 6 and Plate 14 from the Tahoe-100M dataset, a pair of plate-level technical replicates. For each plate, raw UMI counts were normalized to a library size of 10,000 counts per cell, followed by log1p transformation. The top 3,000 highly variable genes were selected and used for downstream analyses.

We quantified the magnitude of each drug-induced perturbation using energy distance (E-distance), calculated between drug-treated cells and DMSO-treated control cells within each plate. Drugs were ranked separately in Plate 6 and Plate 14 according to their E-distance from the DMSO control. To focus on perturbations with robust transcriptional effects, we selected the four drugs with the highest E-distance rankings for further analysis.

For each selected drug, we computed numeric $\log_2$ fold changes in gene expression for each cell line independently in each plate by comparing drug-treated cells with DMSO-treated controls. This produced a plate-specific numeric DEG profile for each drug–cell line pair.

To construct the corresponding text-based representation, we selected, for every drug–cell line–plate combination, the 50 genes with the largest absolute numeric $\log_2$ fold changes as representative perturbation-responsive genes. Genes with positive and negative $\log_2$ fold changes were assigned to up-regulated and down-regulated gene sets, respectively. These gene sets were then converted into a structured text description of the perturbation response and provided to GEMGen. The resulting text-derived DEG profile was then extracted from the GEMGen output.

Negative sample construction for scorer training

By design, the compound–response scoring model functions as a binary classifier, predicting whether a given small molecule is consistent with a specified DGE signature in a particular cell type. However, the observed cell-compound perturbation data only provide positive samples-pairs in which the compound is known to induce the given transcriptional response. To enable effective binary

training, we constructed negative samples using a similarity-based replacement strategy.

For each positive sample, we generated a matched negative example by replacing the compound with a structurally dissimilar molecule. Specifically, we randomly selected the replacement compound from the 100 compounds in the training set with the lowest Tanimoto similarity (based on Morgan fingerprints) to the original compound. All other components, including the cell type and DGE signature, were kept unchanged, and the sample was relabeled as negative.

This strategy ensures that the resulting negative examples are chemically plausible yet biologically implausible, enabling the model to learn to distinguish true perturbation–response relationships from incompatible pairs. The final dataset was balanced, containing one negative sample for each positive example.

Randomly paired compound–response samples for scorer evaluation

To assess GEMGen's performance, we constructed randomly paired compound–response control samples and compared their scores with those of non-hit molecules, defined as generated compounds with Tanimoto similarity below 0.6 to ground-truth compounds. For this purpose, non-hit compounds associated with each test response condition were replaced through the following procedure. First, we randomly selected candidate molecules from the pool of valid compounds generated for other test samples. To ensure structural dissimilarity and avoid trivial matches, candidate replacements were required to satisfy a structural constraint: the maximum Tanimoto similarity between any replacement molecule and the set of molecules originally generated for the target response condition had to be below 0.4. The selected replacement compound was then paired with the target response condition to form a randomly matched control sample.

Baselines

Gex2SGen (46). Since the original implementation of Gex2SGen was not publicly available, we reconstructed the model based on the architectural and training specifications described in the publication. We implemented two variational autoencoders: SMILES-VAE employs an encoder-decoder framework with two stacked GRU layers for processing chemical molecular representations, while p-VAE utilizes a multilayer perceptron architecture to learn latent representations of transcriptomic differential expression profiles, with both latent spaces dimensioned to 256. During pre-training, SMILES-VAE was trained on 1.86 million molecules from the ChEMBL and LINCS databases, and p-VAE was trained on 388,466 differential expression profiles from the LINCS project. To accommodate transcriptomic data

from different test sets, we identified the gene intersection between the LINCS and Tahoe100M datasets, comprising a total of 933 genes. For all molecules involved in pre-training, we filtered them using the maximum sequence length of 208 from the test set as the threshold. In the joint training phase, we utilized 388,466 paired transcriptomic expression profiles and molecular SMILES data, connecting p-VAE's encoder and SMILES-VAE's decoder through a 256-dimensional adapter layer to enable cross-modal generation from gene expression data to molecular structures. The model's loss function combines reconstruction loss with KL divergence, employing a progressive weighting strategy for KL terms to ensure latent space continuity and generation diversity. During testing, differential expression vectors from the test set are used as input to generate 1,000 molecular structures through repeated sampling.

GexMolGen (47). Following the instructions provided in the code repository of GexMolGen, we downloaded and installed the corresponding code package and model weights. To accommodate the omics input requirements of GexMolGen, we performed corresponding data processing for each test set. For the L1000CDS^2 test set, we identified the corresponding pre- and post-perturbation transcriptomic data in the LINCS database based on cell line and perturbation information. Where multiple replicate experiments existed, we averaged their values, utilizing only the 978 truly measured landmark genes. For the Tahoe100M and K562 test sets, we identified cells before and after each perturbation based on the perturbation information. Through sampling, we ensured equal numbers of control and perturbed groups, then computed their respective averages. In this case, we utilized the top 2000 highly variable genes. During testing, we processed one sample at a time, setting the sampling parameter *beam_width* to 1000 and the generation parameter greedy to False. This configuration ensured the generation of 1000 small molecules per sample while preventing excessive repetition in the generated molecular structures.

ChemDiv. We downloaded the entire set of 1.53 million small molecules from ChemDiv's official website (https://www.chemdiv.com/) and employed a random sampling with replacement method to generate 1,000 small molecules per test sample as the output of the virtual screening.

GPT-5. GPT-5 is a widely used general-purpose large language model with strong cross-domain reasoning and generation capabilities. We adopted its dialogue-optimized variant, *gpt-5-chat-latest*, as a baseline due to its reliable instruction-following behavior and stable structured outputs. To ensure a fair and controlled

comparison, the prompt provided to GPT-5 was kept identical to that used for the fine-tuned GEMGen model. To better elicit the capabilities of gpt-5-chat-latest under our task setting, we designed the following system instruction:

> *"You are a SMILES generator. Your only task is to generate exactly one valid SMILES string of drug-like small molecules, based on a provided panel of genes that are either up-regulated or down-regulated, listed in order of regulation magnitude. The generated SMILES string must have the potential to induce the specified gene panel changes.*
>
> *Instructions:*
>
> *• Each molecule must be wrapped individually in <mol> and </mol> tags.*
>
> *• Do not include any explanations or text outside the <mol></mol> tags.*
>
> *• Output exactly one molecule — no more, no less."*

This strict format specification ensured comparable and structurally constrained baseline outputs. For each target transcriptional response, the instruction was executed 1,000 times to generate candidate molecules.

Scaffold analysis

A scaffold reference set was constructed from ChemDiv BMS library (https://www.chemdiv.com/catalog/diversity-libraries/bemis-murcko-clustering-library/) consisting 300 thousand molecules, from which 80,611 unique cores were extracted using MurckoScaffold.GetScaffoldForMol() for exact scaffold and MurckoScaffold.MakeScaffoldGeneric() for generic topology. The eight active molecules were processed using the same scaffold extraction procedure. Each query exact Bemis–Murcko scaffold was matched against the exact reference scaffold set by exact scaffold SMILES matching. Each query generic Bemis-Murcko topology was similarly matched against the generic reference topology set.

Experiments

Cell culture. HEK293T cells, K562 cells, NIH-3T3 cells were purchased from ATCC. LX-2 cells were purchased from the National Collection of Authenticated Cell Cultures. K562 cells were cultured in RPMI 1640 medium (Corning, 10-040-CV) supplemented with 10% fetal bovine serum (CellMax, SA211.02). HEK293T, LX-2 cells, and NIH-3T3 cells were cultured in high glucose DMEM medium (Corning, 10-013-CV) supplemented with 10% fetal bovine serum (CellMax, SA211.02).

Dose-Response Analysis of Flavokawain C. K562 cells and HEK293T cells were treated with flavokawain C (TargetMol) for 24 hours at final concentrations of 40 μM, 20 μM, 10 μM, 5 μM, 2.5 μM, 1.25 μM, and 0.625 μM. Cells treated with DMSO were included as the control group. At study end, total RNA was extracted from the cells (TRIzol, 15596018CN, ThermoFisher) and reverse-transcribed into cDNA (Takara, RR047A). Quantitative PCR (qPCR) (Takara, RR820) was performed to measure the mRNA expression level of *NQO1*. The dose-dependent effects of flavokawain C on *NQO1* expression were analyzed in both K562 and 293T cells.

Transcriptomic analysis of flavokawain C. K562 cells were cultured in 10-cm dishes and treated with flavokawain C or the known NRF2 activator bardoxolone at a final concentration of 10 μM or 100 nM, respectively. Cells treated with DMSO served as the control group. After 24 hours of treatment, total RNA was extracted using TRIzol reagent according to the manufacturer's instructions. The purified RNA samples were submitted to Genewiz for RNA sequencing. The resulting RNA-seq data were subjected to downstream bioinformatics analysis to evaluate gene expression changes induced by flavokawain C and bardoxolone.

Western Blot. HEK293T cells were seeded into 6-well plates and treated with flavokawain C at final concentrations of 5, 10, and 20 μM for 24 hours. Cells treated with DMSO were used as the control group. After treatment, cells were harvested and total cellular protein was extracted using RIPA lysis buffer (Millipore, 20-188). Protein samples were subjected to SDS–PAGE and transferred to membranes for Western blot analysis. Membranes were blocked with 5% BSA in TBST for 1 hour followed by incubation with primary antibodies (NRF2: proteintech, 80593-1-RR, β-actin: proteintech, 66009-1-Ig) in 5% BSA/TBST. Secondary antibodies (Sigma-Aldrich, A9044-2ML, A0545-1ML) were diluted in TBST. NRF2 protein levels were detected, and β-actin was used as the internal loading control.

Dose-Response Analysis of antifibrotic molecules. The antifibrotic effect of generated molecules was evaluated in both LX-2 cells and NIH-3T3 cells. Cells were seeded into 12-well plates and cultured in DMEM supplemented with 10% FBS for 12 hours. Cells were then treated with molecules at annotated concentrations in DMEM containing 0.1% FBS for 12 hours. 10 μM SB431542 and DMSO were used as control groups. Without changing the medium, TGFβ (PeproTech, AF-100-21C-2) was directly added to the culture medium to induce fibrosis at a final concentration of 10 ng/mL followed by further incubation for 48 hours (LX-2 cells) or 24 hours (NIH-3T3 cells). At study end, total RNA was extracted from the cells followed by qPCR to measure mRNA abundance for *FN1* and *Acta2* in LX-2 cells and NIH-3T3 cells, respectively.

Transcriptomic analysis of GEMGen-010/012. LX-2 cells were cultured in 6-well plate and pre-treated with GEMGen-010/012 at 5 μM in DMEM containing 0.1% FBS for 12 hours followed by TGFβ (PeproTech, AF-100-21C-2) stimulation for 48 hours. Cells treated with DMSO served as the control group. Cells omitting TGFβ induction were included as quiescent control. At study end, total RNA was extracted using TRIzol reagent according to the manufacturer's instructions. The purified RNA samples were submitted to GeneWiz (Genewiz) for RNA sequencing. The resulting RNA-seq data were subjected to downstream bioinformatics analysis to evaluate gene expression changes induced by molecules.

Immunofluorescent staining. LX-2 cells were seeded into 12-well plates and cultured in DMEM supplemented with 10% FBS for 12 hours. Cells were then treated with compounds at annotated concentrations in DMEM containing 0.1% FBS for 12 hours followed by stimulation with 10 ng/mL TGFβ for 48 hours. At study end, cells were fixed with a 4% paraformaldehyde solution for 10 min and subsequently permeabilized with 0.5% Triton X-100 in PBS followed by blocking for 1 h with PBS containing 0.1% Triton X-100 and 5% FBS and then staining overnight at 4 °C with primary anti-fibronectin antibody solution (proteintech, 66042-1-Ig, 5% FBS, 0.1% Triton X-100 in PBS). Cells were then stained with a secondary antibody solution containing Hoechst 33342 dye (10 μg/mL) for 1 h at room temperature in the dark. Cells were then washed three times with PBS prior to imaging.

Compound de novo synthesis. The selected 17 molecules were synthesized by BioDuro. The synthetic procedures, LC-MS analysis results and NMR results were provided in the data repository. (https://zenodo.org/records/18739231)

Animal use statement. All experiments were performed in accordance with methods approved by the Institutional Animal Care and Use Committee at the Westlake University under protocol AP#25-093-SSD-11. Mice were maintained in autoclaved cages with free access to food and water.

Pharmacokinetic studies. Pharmacokinetic analysis of GEMGen-010 levels was performed in house using LC–MS-based quantification using authentic standards. Female 9- to 11-wk-old C57BL/6J mice were randomly assigned to experimental groups (n = 3). GEMGen-010 was prepared fresh at 2 mg/mL in dosing vehicle containing 10% DMSO, 40% PEG300, 5% Tween-80, and 45% saline, and was dosed once orally or intraperitoneally at 20 mg/kg, followed by blood collection by retro-orbital bleeding at 1, 3, 8, 12, and 24 hours post dosing. Plasma was separated from blood by centrifugation at 1,500×g for 10 min at 4°C. GEMGen-010 compound was extracted with acetonitrile, followed by LC-MS (Waters) analysis. Chromatographic separation was

performed using ACQUITY UPLC BEH C18 column (1.7 μm; 2.1 × 50 or 100 mm) maintained at 40°C, using water (solvent A) and acetonitrile (solvent B), each containing 0.1% formic acid, at a flow rate of 0.30 mL/min. The gradient program was as follows: 0–1.0 min, 95% A/5% B; 1.0–3.5 min, linear gradient to 5% A/95% B; 3.5–5.5 min, hold at 5% A/95% B; 5.5–7.5 min, re-equilibrate to 95% A/5% B, for a total run time of 7.5 min. Ion detection was carried out on a Waters XEVO TQ-S mass spectrometer operating in ESI positive ion mode. GEMGen-010 was monitored in multiple reaction monitoring (MRM) mode.

CCl_4-induced liver fibrosis model. Male C57BL/6J mice aged 6–8 weeks and weighing 20–25 g were maintained under specific pathogen-free conditions at the Laboratory Animal Center of Westlake University and allowed to acclimatize before the start of the experiments. Mice were assigned to three experimental groups: an olive oil-treated sham control group, a CCl_4-treated model group receiving vehicle, and a CCl_4-treated group receiving GEMGen-010. CCl_4 (MedChemExpress, HY-Y0298) was diluted to 10% (v/v) in olive oil and administered by intraperitoneal injection at 10 mL/kg body weight, corresponding to a CCl_4 dose of 1 mL/kg. CCl_4 was administered twice weekly, on Tuesdays and Fridays, for four consecutive weeks. Sham control mice received an equivalent volume of olive oil according to the same schedule. During the final 2 weeks of CCl_4 administration, mice in the treatment group received GEMGen-010 by daily intraperitoneal injection at 50 mg/kg. GEMGen-010 was formulated in a vehicle consisting of 10% DMSO, 40% PEG300, 5% Tween-80 and 45% saline (v/v/v/v). Mice in the model group received an equivalent volume of vehicle according to the same dosing schedule. Three days after the final CCl_4 injection, mice were euthanized by isoflurane overdose. Blood was collected for serum preparation, and liver tissues were harvested for subsequent biochemical, histological and molecular analyses.

Serum ALT and AST activity test. Whole blood was collected from mice and allowed to clot at room temperature for 30 min. The blood samples were then centrifuged at 1,500 × g for 10 min at 4°C, and the supernatant serum was carefully collected. Serum alanine aminotransferase activity was measured using a GPT/ALT Activity Assay Kit (Solarbio, BC1555). Serum aspartate aminotransferase activity was measured using a GOT/AST Activity Assay Kit (Solarbio, BC1565). Before measurement, serum samples and assay reagents were equilibrated to room temperature. Working solutions were prepared according to the manufacturer's instructions. Serum samples and the corresponding reaction reagents were added to a 96-well plate, mixed, and incubated under the conditions specified in the kit protocols. Absorbance was measured using a microplate reader, and serum ALT and AST activities were calculated according to the manufacturer's instructions.

Table S5. GPT-4o augmented instruction templates

Gene name instructions	Cell type instructions

"What is the function of the *[gene]* gene?", "What role does the *[gene]* gene play in the body?", "Can you explain the purpose of the *[gene]* gene?", "What does the *[gene]* gene do?", "What is the significance of the *[gene]* gene?", "How does the *[gene]* gene work?", "What is the biological function of the *[gene]* gene?", "What is the main function of the *[gene]* gene?", "What is the specific role of the *[gene]* gene?", "What is the function of the *[gene]* gene in simple terms?", "What is the role of the *[gene]* gene?", "How does the *[gene]* gene affect the body?", "What effect does the *[gene]* gene have?", "What is the impact of the *[gene]* gene?", "How does the *[gene]* gene influence biological processes?", "What is the contribution of the *[gene]* gene?", "How does the *[gene]* gene contribute to health?" "What is the role of the *[gene]* gene in disease?", "How does the *[gene]* gene affect development?", "What is the role of the *[gene]* gene in gene regulation?", "Can you summarize the *[gene]* gene?", "What is a brief overview of the *[gene]* gene?", "Can you provide a summary of the *[gene]* gene?", "What are the key points about the *[gene]* gene?", "Can you give a concise summary of the *[gene]* gene?", "What is the main idea of the *[gene]* gene?", "Can you summarize the function of the *[gene]* gene?", "What is the essence of the *[gene]* gene?", "Can you summarize the importance of the *[gene]* gene?", "What is a summary of the *[gene]* gene's role?", "What is the *[gene]* gene?", "Can you explain what the *[gene]* gene is?", "What exactly is the *[gene]* gene?", "What is the definition of the *[gene]* gene?", "What does the *[gene]* gene refer to?", "What is the *[gene]* gene in biology?",	"Describe the cell type *[cell type]*.", "Highlight the characteristics of the cell type *[cell type]*.", "Summarize the details of the cell type *[cell type]*.", "Give an overview of the cell type *[cell type]*.", "Explain the attributes of the cell type *[cell type]*.", "Detail the properties of the cell type *[cell type]*.", "What is known about the cell type *[cell type]*?", "Outline the key aspects of the cell type *[cell type]*.", "Provide a description for the cell type *[cell type]*.", "What defines the cell type *[cell type]*?", "What is the function of the *[cell type]* cell type?", "What role does the *[cell type]* cell type play in cellular processes?", "How does the *[cell type]* cell type contribute to biological functions?", "What are the primary functions of the *[cell type]* cell type?", "What is the biological significance of the *[cell type]* cell type?", "How is the *[cell type]* cell type involved in disease mechanisms?", "Explain the primary applications of the cell type *[cell type]* in research.", "How is the cell type *[cell type]* used in cancer studies?", "What is the origin and history of the cell type *[cell type]*?", "Describe the genetic features of the cell type *[cell type]*.", "Tell me something about the cell type *[cell type]*."

"What is the *[gene]* gene known for?", "What is the *[gene]* gene all about?", "What is the *[gene]* gene in simple terms?", "What is the *[gene]* gene and its function?", "What is the *[gene]* gene used for?", "What is the purpose of the *[gene]* gene?", "How is the *[gene]* gene useful?", "What is the benefit of the *[gene]* gene?", "What is the application of the *[gene]* gene?", "Why is the *[gene]* gene important?", "What is the value of the *[gene]* gene?", "How does the *[gene]* gene help?", "What is the utility of the *[gene]* gene?", "What is the significance of the *[gene]* gene in practical terms?"	

1266

1267

1268

1269

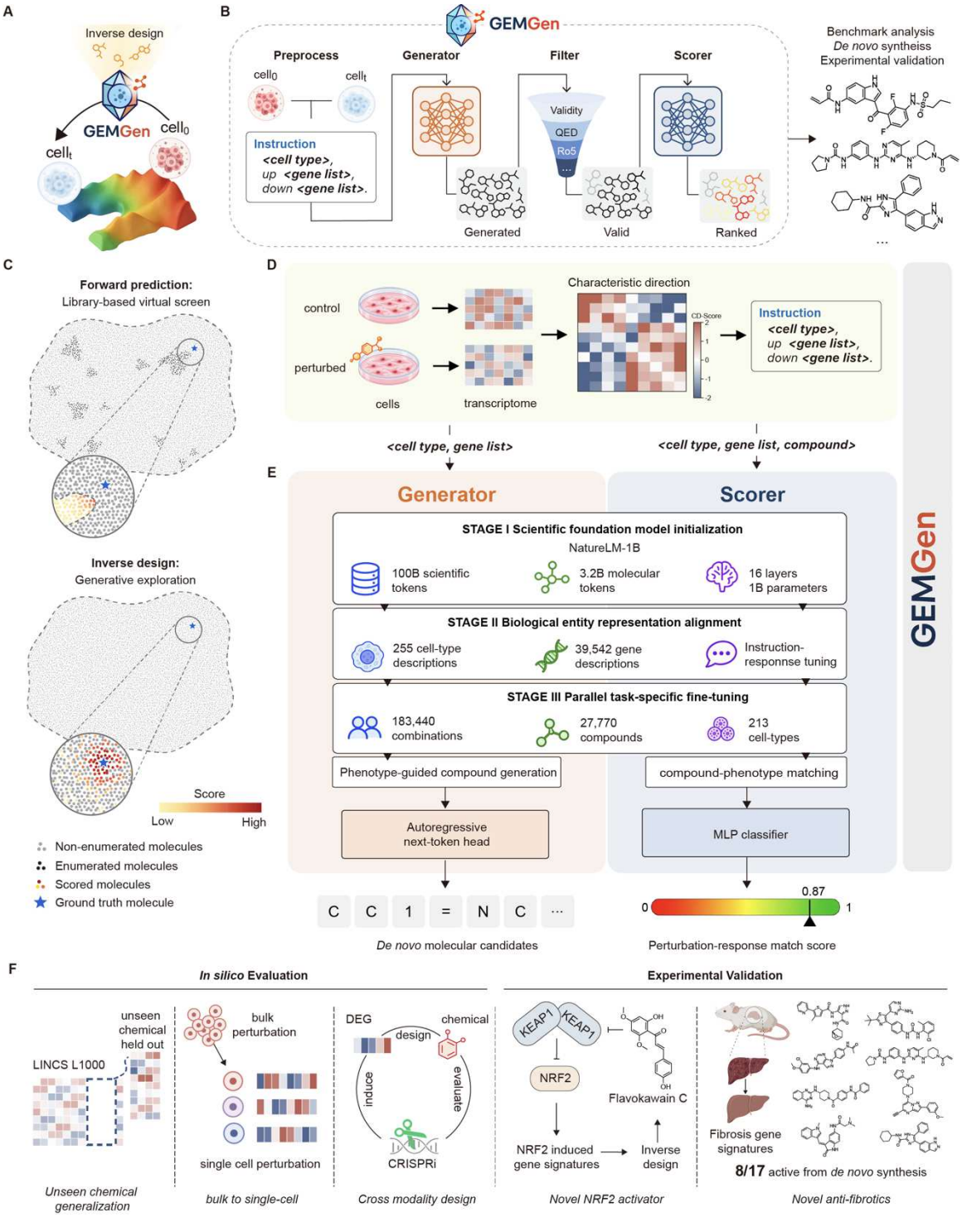

A
Inverse design
GEMGen
cell0
cellt
B
GEMGen
Preprocess
Generator
Filter
Scorer
cell0
cellt
Instruction
<cell type>,
up <gene list>,
down <gene list>.
Validity
QED
Ro5
Generated
Valid
Ranked
Benchmark analysis
De novo syntheiss
Experimental validation
C
Forward prediction:
Library-based virtual screen
Inverse design:
Generative exploration
Score
Low
High
Non-enumerated molecules
Enumerated molecules
Scored molecules
Ground truth molecule
D
Characteristic direction
control
perturbed
cells
transcriptome
CD-Score
Instruction
<cell type>,
up <gene list>,
down <gene list>.
<cell type, gene list>
<cell type, gene list, compound>
E
Generator
Scorer
GEMGen
STAGE I Scientific foundation model initialization
NatureLM-1B
100B scientific tokens
3.2B molecular tokens
16 layers
1B parameters
STAGE II Biological entity representation alignment
255 cell-type descriptions
39,542 gene descriptions
Instruction-responnse tuning
STAGE III Parallel task-specific fine-tuning
183,440 combinations
27,770 compounds
213 cell-types
Phenotype-guided compound generation
compound-phenotype matching
Autoregressive next-token head
MLP classifier
C C 1 = N C ···
0.87
0
1
De novo molecular candidates
Perturbation-response match score
F
In silico Evaluation
Experimental Validation
unseen chemical held out
LINCS L1000
bulk perturbation
single cell perturbation
DEG
chemical
design
induce
evaluate
CRISPRi
KEAP1
KEAP1
NRF2
Flavokawain C
NRF2 induced gene signatures
Inverse design
Fibrosis gene signatures
8/17 active from de novo synthesis
Unseen chemical generalization
bulk to single-cell
Cross modality design
Novel NRF2 activator
Novel anti-fibrotics

Figure 1. GEMGen enables inverse design of chemical perturbations from transcriptomic state transitions. A, Conceptual overview of GEMGen. Given transcriptomic profiles representing an initial cellular state and a desired state, GEMGen generates candidate chemical perturbations predicted to induce the corresponding transition. **B,** End-to-end inverse-design workflow. Paired cellular states are preprocessed into an instruction specifying the cell type and rank-ordered lists of upregulated and downregulated genes. The generator proposes molecular structures, molecular-property filters retain valid and drug-like candidates, and the scorer ranks the remaining molecules according to their predicted ability to reproduce the target transcriptional response. Ranked candidates are subsequently used for benchmark analysis, de novo synthesis and experimental validation. **C,** Conventional forward *in silico* perturbation prediction by library-based virtual screening scores a finite set of enumerated molecules within a predefined chemical collection, whereas the inverse-design approach generates and prioritizes previously non-enumerated molecules, enabling broader exploration of chemical space. **D,** Construction of the phenotype prompt from chemical perturbation data. Control and perturbed transcriptomes are compared using the characteristic direction method, and the resulting directional differential-expression scores are converted into a cell-type-specific instruction containing upregulated and downregulated gene lists. **E,** GEMGen architecture and multi-stage training strategy. NatureLM-1B, an 1-billion-parameter scientific foundation model, is first aligned with biological entity representations and subsequently fine-tuned in parallel to develop specialized generator and scorer modules. GEMGen samples both enumerated and previously non-enumerated molecules and assigns perturbation-response match scores, enabling exploration beyond predefined libraries while prioritizing candidates predicted to induce the desired response. **F,** Overview of *in silico* evaluation and experimental validation. *In silico* analyses assess generalization to unseen chemicals, transfer from bulk to single-cell perturbation data and cross-modality design from CRISPRi-derived signatures. Prospective validation identifies flavokawain C as an NRF2 activator and confirms antifibrotic activity among generated candidates, including 8 of 17 *de novo* synthesized compounds.

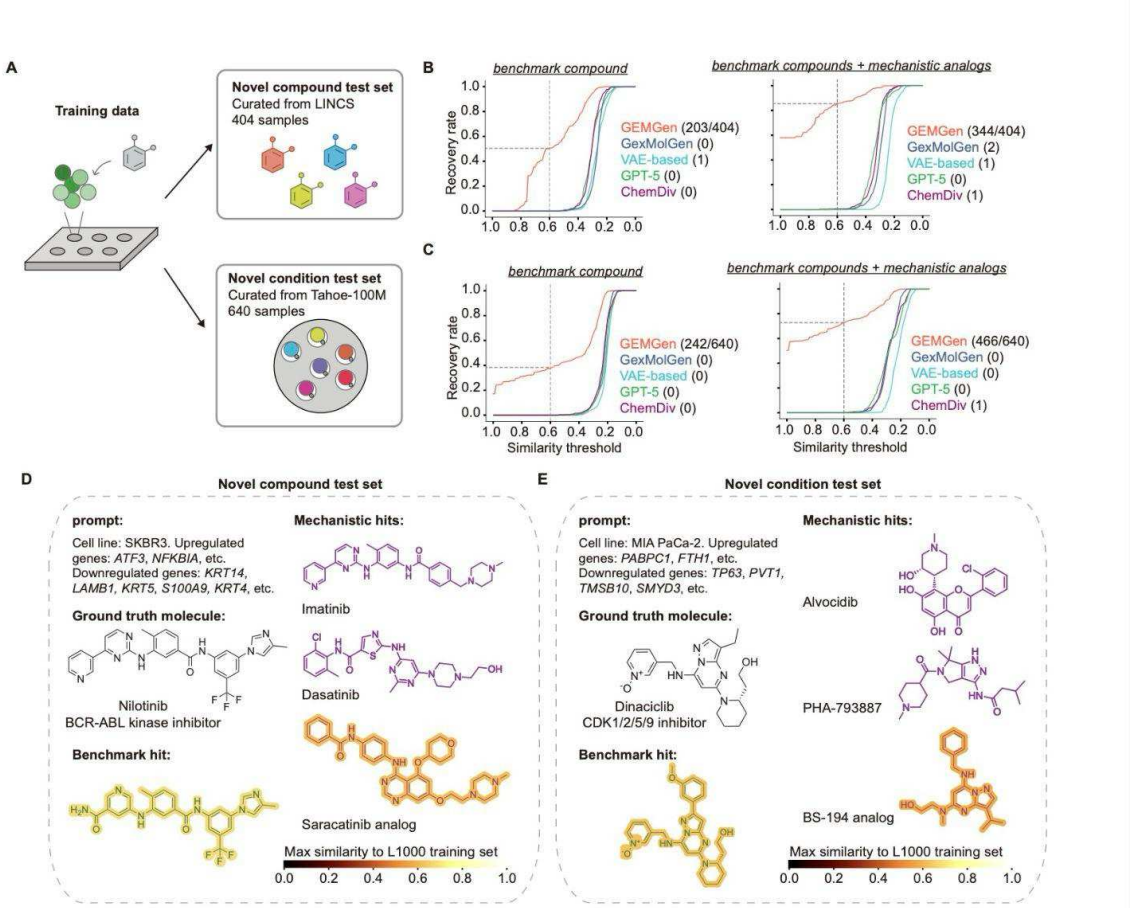


Figure 2. GEMGen recovers chemical perturbations across unseen compounds and conditions. A, Schematic of two complementary evaluation datasets: a novel compound test set (404 samples) curated from LINCS, and a novel condition test set (640 samples) curated from Tahoe-100M. **B-C**, Model performance evaluated by recovery rate (fraction of samples with at least one hit) across varying Tanimoto similarity thresholds. Results are shown for the novel compound (**B**) and novel condition (**C**) sets. Left panels show recovery of the benchmark compounds (compounds used for biological experiments); right panels show recovery when including mechanistic analogs (compounds sharing the same molecular targets or pathways to benchmark compounds). Vertical dashed lines indicate a similarity threshold of 0.6. **D-E,** Representative examples from the novel compound (**D**) and novel condition (**E**) test sets. For each case, generated compounds include both structural analogs of the benchmark compound (green) and mechanistically related molecules with distinct scaffolds (purple). Shading color indicates maximum similarity to the training dataset.

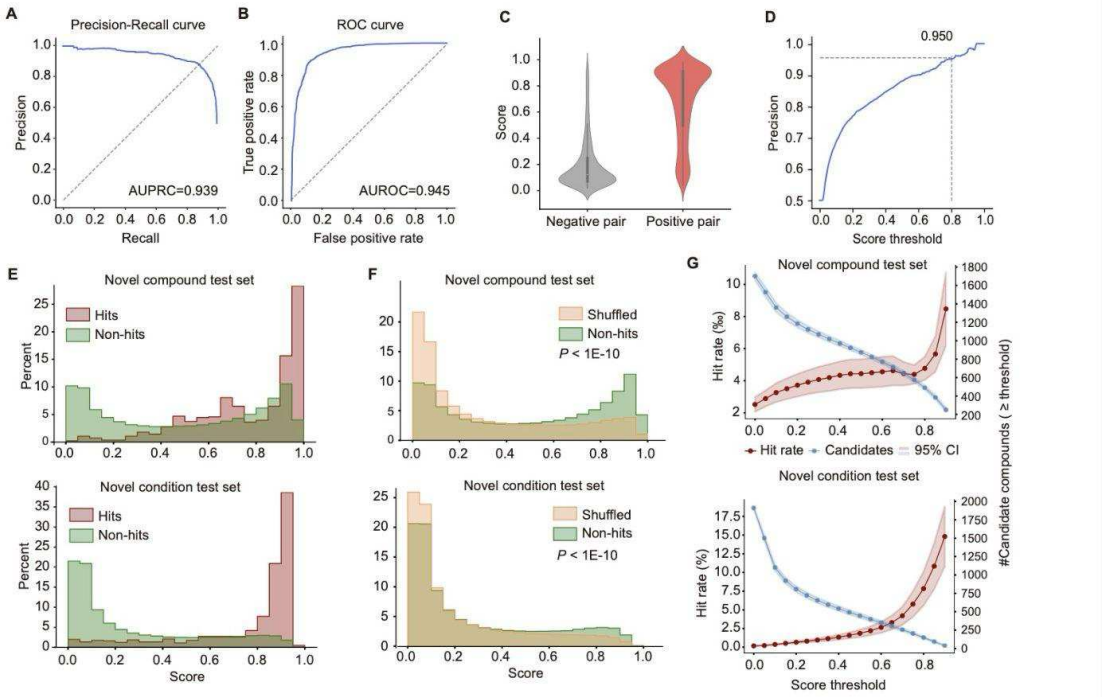


Figure 3. Compound–response scorer distinguishes matched pairs and prioritizes generated candidates. A-B, Performance of the scorer in discriminating benchmark compound–response pairs from randomly shuffled negatives, evaluated by (**A**) precision–recall and (**B**) Receiver Operating Characteristic (ROC) curves. **C,** Score distributions for positive and negative pairs. The scorer assigns significantly higher values to true pairs (two-sided Mann-Whitney *U* test, $P < 1\times10^{-15}$). N=2,088 for both positive and negative pairs. **D,** Precision across varying score threshold. A precision of 0.950 is achieved at a threshold of 0.8 (dashed line). **E,** Score distributions for GEMGen-generated candidates in the novel compound (top) and novel condition (bottom) test sets. “Hits” (Tanimoto similarity > 0.6 to benchmark compounds) receive higher scores than “non-hits”. **F,** Comparison of score distributions between non-hit generated compounds and randomly shuffled pairs. *P*-values are calculated with two-sided Mann-Whitney *U* test. **G**, Effect of score-based filtering on candidate retention (blue) and hit rate (red) in the novel compound (top, N=203) and novel condition (bottom, N=242) test sets. Shaded bands indicate 95% confidence intervals.

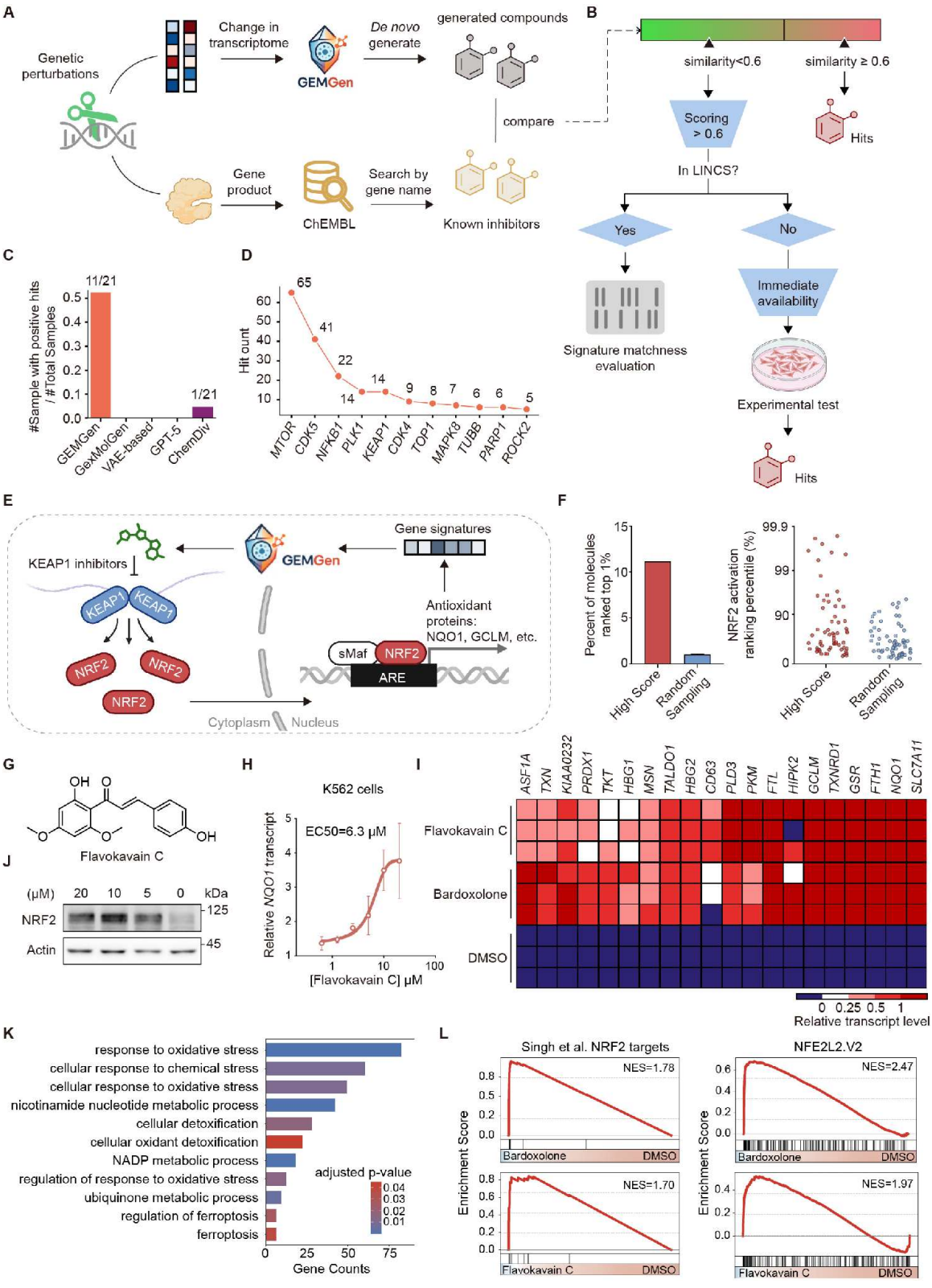

A
Genetic perturbations
Change in transcriptome
GEMGen
De novo generate
generated compounds
compare
Gene product
ChEMBL
Search by gene name
Known inhibitors
B
similarity<0.6
similarity ≥ 0.6
Scoring > 0.6
Hits
In LINCS?
Yes
No
Signature matchness evaluation
Immediate availability
Experimental test
Hits
C
#Sample with positive hits / #Total Samples
11/21
1/21
GEMGen
GexMolGen
VAE-based
GPT-5
ChemDiv
D
Hit count
65
41
22
14
14
9
8
7
6
6
5
MTOR
CDK5
NFKB1
PLK1
KEAP1
CDK4
TOP1
MAPK8
TUBB
PARP1
ROCK2
E
KEAP1 inhibitors
GEMGen
Gene signatures
KEAP1
NRF2
Antioxidant proteins: NQO1, GCLM, etc.
sMaf
ARE
Cytoplasm
Nucleus
F
Percent of molecules ranked top 1%
NRF2 activation ranking percentile (%)
High Score
Random Sampling
G
Flavokavain C
H
K562 cells
EC50=6.3 μM
Relative NQO1 transcript
[Flavokavain C] μM
I
ASF1A
TXN
KIAA0232
PRDX1
TKT
HBG1
MSN
TALDO1
HBG2
CD63
PLD3
PKM
FTL
HIPK2
GCLM
TXNRD1
GSR
FTH1
NQO1
SLC7A11
Flavokavain C
Bardoxolone
DMSO
Relative transcript level
J
(μM)
20
10
5
0
kDa
NRF2
125
Actin
45
K
response to oxidative stress
cellular response to chemical stress
cellular response to oxidative stress
nicotinamide nucleotide metabolic process
cellular detoxification
cellular oxidant detoxification
NADP metabolic process
regulation of response to oxidative stress
ubiquinone metabolic process
regulation of ferroptosis
ferroptosis
adjusted p-value
Gene Counts
L
Singh et al. NRF2 targets
NFE2L2.V2
NES=1.78
NES=2.47
NES=1.70
NES=1.97
Enrichment Score
Bardoxolone
DMSO
Flavokavain C
DMSO

Figure 4. GEMGen designs chemical perturbations that phenocopy genetic perturbation states. A, Schematic overview of the evaluation framework for genetic perturbations. Transcriptional signatures induced by genetic perturbations (e.g., CRISPR interference) serve as input for GEMGen to *de novo* generate candidate molecules. Generated compounds are compared against known inhibitors of the perturbed gene product retrieved from ChEMBL. Compounds achieving a Tanimoto similarity ≥ 0.6 to known inhibitors are classified as hits. **B,** Prioritization workflow for GEMGen-generated compounds. Candidates with structural similarity < 0.6 to known inhibitors are evaluated by the compound–response scorer. Compounds with a score > 0.6 are prioritized as potential functional analogs. Prioritized compounds present in the LINCS library undergo signature-matchness evaluation, while those outside LINCS represent novel candidates for experimental synthesis and testing. **C,** Model performance across 21 genetic perturbations. The bar plot shows the fraction of genetic perturbation conditions for which GEMGen and alternative methods successfully generated at least one positive hit. **D,** Distribution of hit counts for 11 specific gene perturbations where GEMGen successfully recovered ground-truth-like compounds. **E**, Schematic illustration of the KEAP1-NRF2 signaling pathway. KEAP1 inhibition stabilizes NRF2, allowing its translocation to the nucleus and activation of antioxidant response element (ARE)-driven genes, including NQO1 and GCLM. GEMGen leverages gene expression signatures associated with NRF2 activation to generate candidate KEAP1 inhibitors. **F**, High-scoring molecules (score ≥ 0.6) were ranked by induction levels of ARE-driven genes, revealing an enriched fraction within the top 1% of NRF2 activation compared with randomly sampled LINCS molecules. The bar plot (left) shows the percentage of molecules in the top 1% NRF2 activation percentile for high-score versus random sampling. The dot plot (right) presents individual generated molecules positioned by their NRF2 activation ranking percentile. **G,** Chemical structure of flavokawain C, a chalcone derivative identified from GEMGen-generated compounds absent from the LINCS dataset. **H,** qPCR analysis of NQO1 transcript levels in K562 cells treated with different concentrations of flavokawain C. Data are shown as mean ± s.e.m; EC50 = 6.3 μM. **I,** Heatmap of RNA-sequencing results showing relative expression levels of NRF2 target genes following treatment with flavokawain C, bardoxolone, or DMSO control. Colors indicate normalized transcript levels. **J,** Western blot analysis of NRF2 protein level in K562 cells following treatment with indicated concentrations of flavokawain C. β-actin was used as loading control. **K**, Gene ontology analysis of RNA sequencing data from K562 cells treated with 10 μM flavokawain C reveals enrichment in antioxidant response pathways. *P* values were adjusted for multiple testing using the Benjamini-Hochberg method. **L**, Gene-set enrichment analyses of RNA sequencing data from K562 cells treated with flavokawain C or bardoxolone using indicated NRF2 related gene sets. NES, normalized enrichment score.

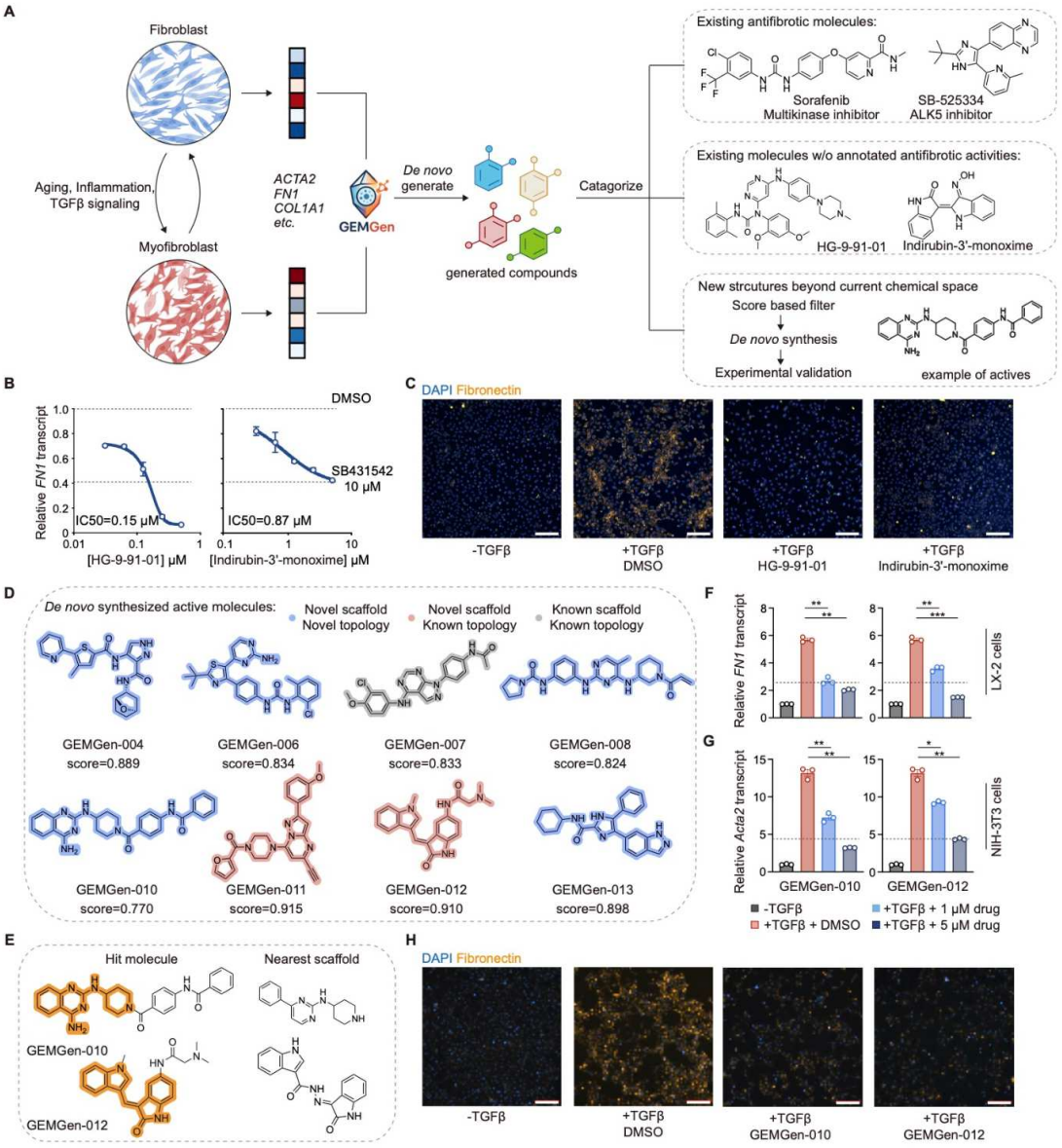

A
Fibroblast
Myofibroblast
Aging, Inflammation, TGFβ signaling
ACTA2
FN1
COL1A1
etc.
GEMGen
De novo generate
generated compounds
Catagorize
Existing antifibrotic molecules:
Sorafenib
Multikinase inhibitor
SB-525334
ALK5 inhibitor
Existing molecules w/o annotated antifibrotic activities:
HG-9-91-01
Indirubin-3'-monoxime
New strcutures beyond current chemical space
Score based filter
De novo synthesis
Experimental validation
example of actives
B
Relative FN1 transcript
DMSO
SB431542
10 μM
IC50=0.15 μM
IC50=0.87 μM
[HG-9-91-01] μM
[Indirubin-3'-monoxime] μM
C
DAPI Fibronectin
-TGFβ
+TGFβ DMSO
+TGFβ HG-9-91-01
+TGFβ Indirubin-3'-monoxime
D
De novo synthesized active molecules:
Novel scaffold Novel topology
Novel scaffold Known topology
Known scaffold Known topology
GEMGen-004 score=0.889
GEMGen-006 score=0.834
GEMGen-007 score=0.833
GEMGen-008 score=0.824
GEMGen-010 score=0.770
GEMGen-011 score=0.915
GEMGen-012 score=0.910
GEMGen-013 score=0.898
E
Hit molecule
Nearest scaffold
GEMGen-010
GEMGen-012
F
Relative FN1 transcript
LX-2 cells
G
Relative Acta2 transcript
NIH-3T3 cells
GEMGen-010
GEMGen-012
-TGFβ
+TGFβ + DMSO
+TGFβ + 1 μM drug
+TGFβ + 5 μM drug
H
DAPI Fibronectin
-TGFβ
+TGFβ DMSO
+TGFβ GEMGen-010
+TGFβ GEMGen-012

Figure 5. GEMGen generates compounds with antifibrotic activity. A, Schematic illustration of fibroblast-myofibroblast state transitions and the application of GEMGen to the fibrosis transcriptional program. Fibroblast activation is marked by upregulation of fibrosis markers including *ACTA2*, *FN1*, and *COL1A1*. The fibrosis-associated transcriptional signature is used as input to GEMGen to generate candidate small molecules predicted to attenuate fibroblast activation. Generated candidate molecules are categorized into three classes, including existing antifibrotic molecules, existing molecules lacking documented antifibrotic activities, and novel structures requiring de novo synthesis. **B**, qPCR analysis of *FN1* transcript levels in TGFβ-stimulated LX-2 cells treated with HG-9-91-01 (left) and indirubin-3′-monoxime (right), normalized to DMSO control levels; SB431542 (10 μM) is included as a positive control, whose level is denoted by the dashed line. IC50 values are indicated. Data are shown as mean ± s.e.m, n=3. **C**, Immunofluorescence analysis of fibronectin deposition in LX-2 cells treated as indicated. Nuclei are stained with DAPI (blue), and fibronectin is shown in yellow. Scale bar = 100 μm. **D**, Chemical structures of *de novo* synthesized molecules confirmed with antifibrotic activities. The GEMGen score of each compound is shown, and colors indicate scaffold classification. Each molecule is colored by scaffold class: gray, known scaffold and topology; red, novel scaffold with known topology; blue, novel scaffold and topology. **E,** Chemical structures of GEMGen-010 and GEMGen-012 with their core scaffolds highlighted in blue. The most similar scaffold from the ChemDiv BMS library for each molecule is shown to the right. **F**, qPCR analysis of *FN1* transcript levels in TGFβ-stimulated LX-2 cells treated with GEMGen-010 (left) and GEMGen-012 (right), SB431542 (10 μM) is included as a positive control, whose level is denoted by the dashed line. **G**, qPCR analysis of *Acta2* transcript levels in TGFβ-stimulated NIH-3T3 cells treated with GEMGen-010 (left) and GEMGen-012 (right), SB431542 (10 μM) is included as a positive control, whose level is denoted by the dashed line. *P*-values are calculated with one-way ANOVA with Dunnett's correction, $*P < 0.05$, $**P < 0.01$, $***P < 0.001$. **H**, Immunofluorescence analysis of fibronectin deposition in LX-2 cells treated as indicated. Nuclei are stained with DAPI (blue), and fibronectin is shown in yellow. Scale bar = 100 μm.

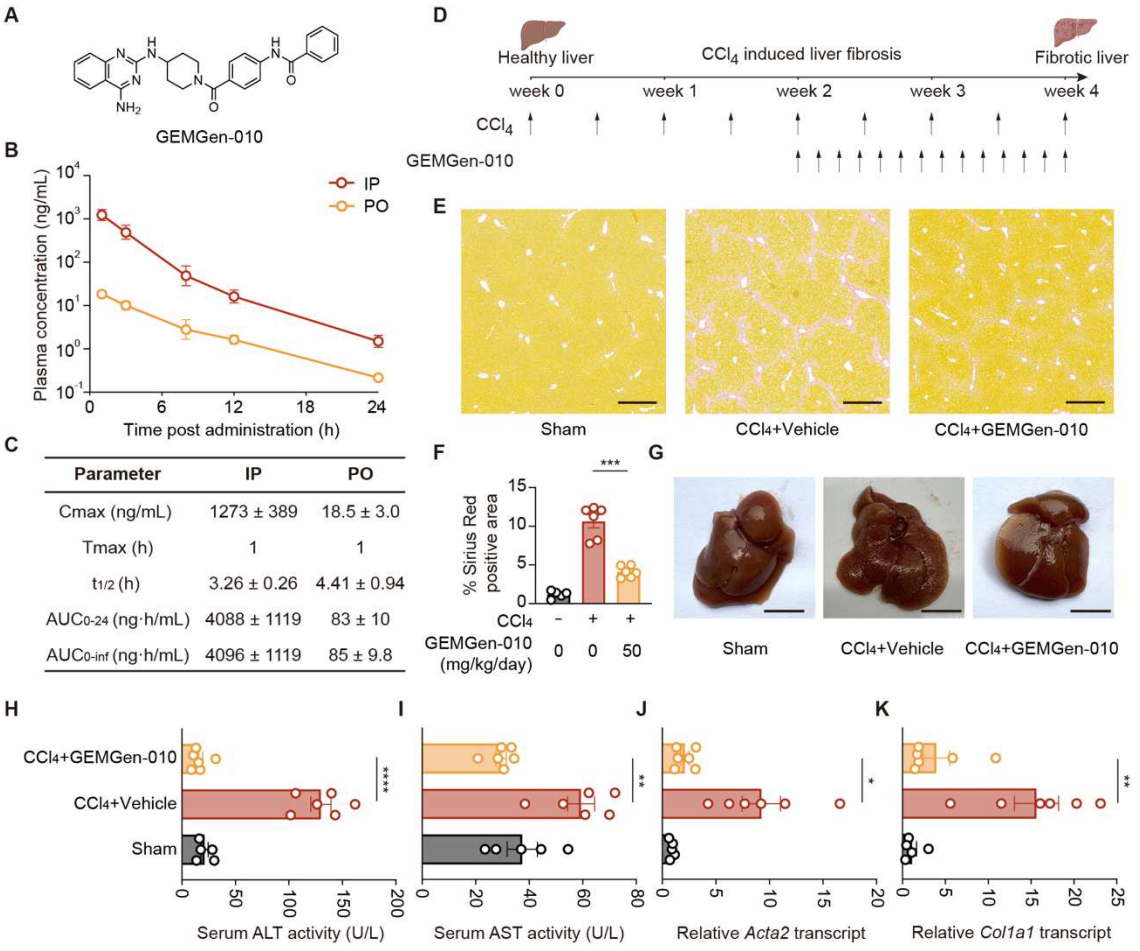


Figure 6. GEMGen-010 alleviates liver fibrosis *in vivo*. **A**, Chemical structure of GEMGen-010. **B**, Plasma concentrations of GEMGen-010 following a single oral (PO) or intraperitoneal (IP) administration at 20 mg/kg in mice. Data are shown as mean ± s.e.m, n=3. **C**, Pharmacokinetic parameters of GEMGen-010 following PO or IP administration at 20 mg/kg. **D**, Schematic of the CCl_4-induced liver fibrosis model. Mice received CCl_4 twice weekly for four weeks. 50 mg/kg GEMGen-010 or vehicle was administered intraperitoneally once daily. **E**, Representative Sirius Red staining of liver sections from mice. Scale bar = 500 μm. **F**, Quantification of Sirius Red–positive area from (**E**). Data are shown as mean ± s.e.m, n = 5-6 mice per group. **G**, Representative gross morphology of livers from the three groups. Scale bar = 1 cm. **H-I**, Serum activities of alanine aminotransferase (**H**) and aspartate aminotransferase (**I**), markers of hepatocellular injury. Data are shown as mean ± s.e.m. **J-K**, qPCR analysis of the fibrosis-associated transcripts *Acta2* (**J**) and *Col1a1* (**K**) in liver tissue, normalized to the sham group. Data are shown as mean ± s.e.m. *P*-values are calculated with one-way ANOVA with Dunnett's correction, *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001.

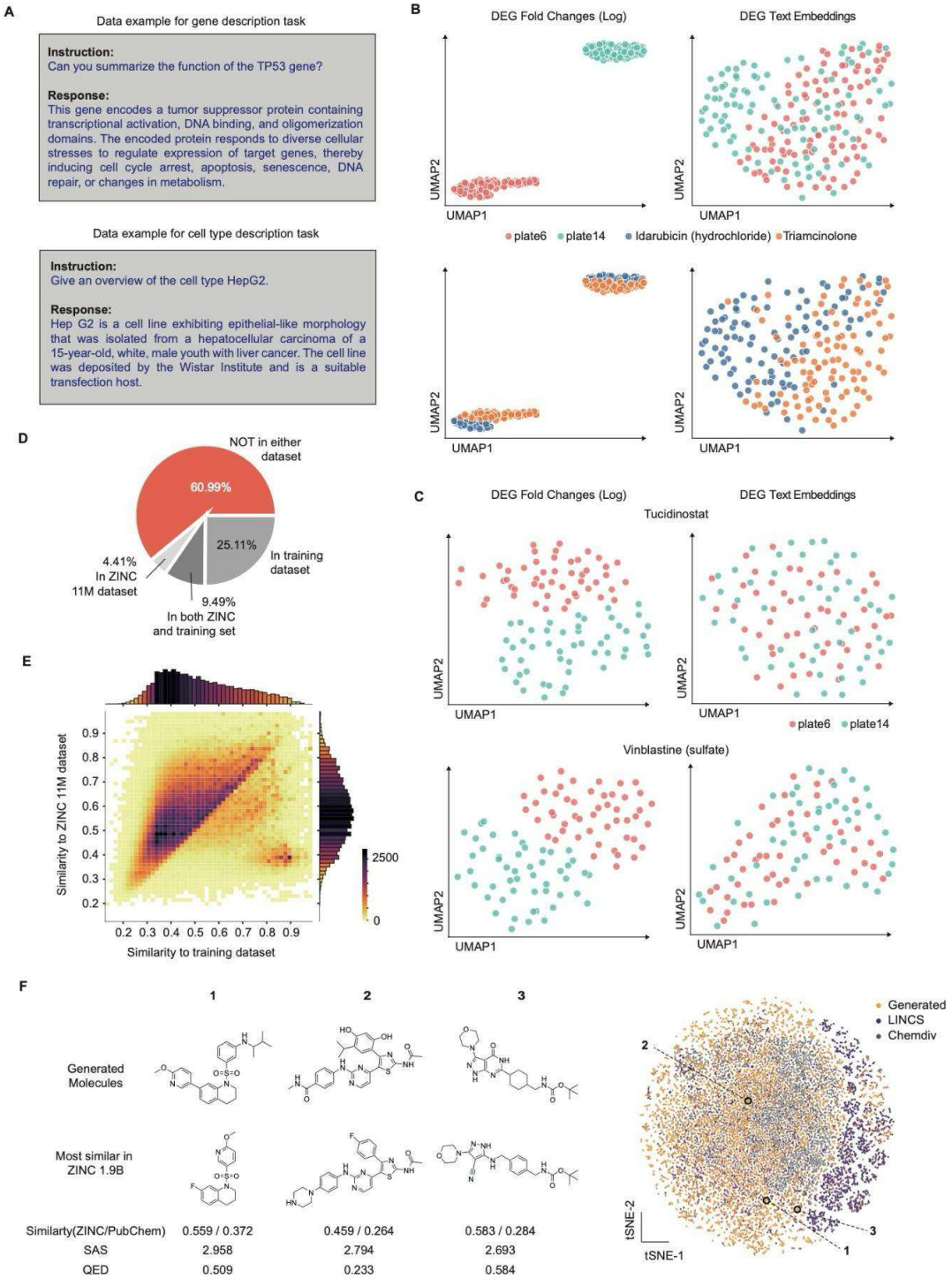

A
Data example for gene description task
Instruction:
Can you summarize the function of the TP53 gene?
Response:
This gene encodes a tumor suppressor protein containing transcriptional activation, DNA binding, and oligomerization domains. The encoded protein responds to diverse cellular stresses to regulate expression of target genes, thereby inducing cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism.
Data example for cell type description task
Instruction:
Give an overview of the cell type HepG2.
Response:
Hep G2 is a cell line exhibiting epithelial-like morphology that was isolated from a hepatocellular carcinoma of a 15-year-old, white, male youth with liver cancer. The cell line was deposited by the Wistar Institute and is a suitable transfection host.
B
DEG Fold Changes (Log)
DEG Text Embeddings
UMAP1
UMAP2
plate6
plate14
Idarubicin (hydrochloride)
Triamcinolone
C
DEG Fold Changes (Log)
DEG Text Embeddings
Tucidinostat
Vinblastine (sulfate)
D
NOT in either dataset
60.99%
25.11%
In training dataset
4.41%
In ZINC 11M dataset
9.49%
In both ZINC and training set
E
Similarity to ZINC 11M dataset
Similarity to training dataset
2500
0
F
1
2
3
Generated Molecules
Most similar in ZINC 1.9B
Similarty(ZINC/PubChem)
0.559 / 0.372
0.459 / 0.264
0.583 / 0.284
SAS
2.958
2.794
2.693
QED
0.509
0.233
0.584
Generated
LINCS
Chemdiv
tSNE-1
tSNE-2

Figure S1. Performance of text-based representation of perturbation responses and characterization of the generated chemical space. A, Representative examples for the bioentity representation generation tasks. **B-C,** UMAP visualization of transcriptomic profiles comparing numerical fold-change values (left) with text-based gene embeddings (right). While numerical data primarily cluster by experimental batch (e.g., Plate 6 vs. Plate 14 from Tahoe-100M dataset), the text-based representation effectively groups samples treated with the same compound, such as Indarubicin, Triamcinolone (**B**), and Tucidinostat, Vinblastine (**C**), regardless of their plate origin. **D,** Distribution of GEMGen-generated molecules relative to the training set and the ZINC 11M reference library. A total of 60.99% of generated compounds represent novel chemical entities absent from both datasets. **E,** Distribution of maximum Tanimoto similarity to the training set and the ZINC 11M dataset for the novel compounds identified in (**D**), reflecting a coverage of chemical regions ranging from structural analogs to entirely distinct chemotypes. **F,** Novelty and diversity of GEMGen-generated compounds. *t*-SNE visualization of the chemical space spanned by GEMGen-generated compounds (yellow), LINCS compounds (purple), and ChemDiv compounds (gray). For each representative generated molecule, the most structurally similar compound retrieved from the ZINC20 database is shown together with similarity scores to ZINC (left) and to PubChem (right), demonstrating that GEMGen explores regions of chemical space beyond its training data.

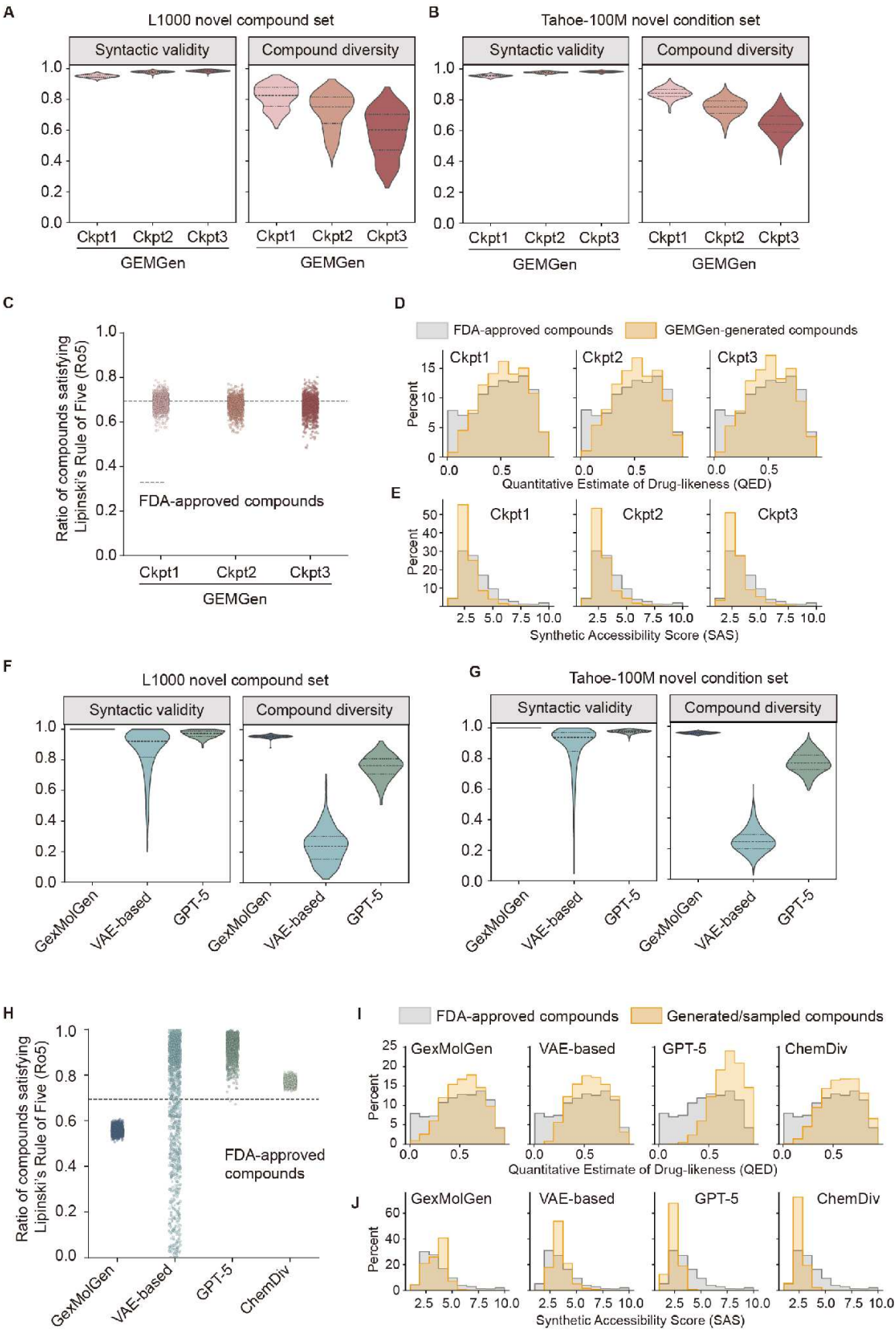

A
L1000 novel compound set
Syntactic validity
Compound diversity
Ckpt1
Ckpt2
Ckpt3
GEMGen
B
Tahoe-100M novel condition set
C
Ratio of compounds satisfying Lipinski's Rule of Five (Ro5)
FDA-approved compounds
D
FDA-approved compounds
GEMGen-generated compounds
Percent
Quantitative Estimate of Drug-likeness (QED)
E
Synthetic Accessibility Score (SAS)
F
GexMolGen
VAE-based
GPT-5
G
H
ChemDiv
I
Generated/sampled compounds
J

Figure S2. Quality and chemical diversity of compounds generated by GEMGen and alternative methods. A-B, Syntactic validity and compound diversity of molecules generated by three GEMGen variants (ckpt1-3), trained for increasing numbers of epochs, on the novel compound **(A)** and novel condition **(B)** test set. Syntactic validity measures the fraction of chemically valid SMILES strings among all molecules generated for each input sample. Compound diversity measures the number of unique valid molecules generated per sample. Dashed lines: the 25th, 50th, and 75th percentiles. **C,** Fraction of generated compounds satisfying Lipinski's Rule of Five (Ro5) compared to FDA-approved compounds (dashed line). **D,** Distributions of the quantitative estimate of drug-likeness (QED) for generated and FDA-approved compounds. **E,** Distributions of the synthetic accessibility score (SAS) for generated and FDA-approved compounds. Lower SAS indicates easier chemical synthesis. **F-G,** Syntactic validity and compound diversity of molecules generated by GexMolGen, VAE-based method, and GPT-5, on the novel compound **(F)** and novel condition **(G)** test set. Syntactic validity measures the fraction of chemically valid SMILES strings among all molecules generated for each input sample. Compound diversity measures the number of unique valid molecules generated per sample. Dashed lines: the 25th, 50th, and 75th percentiles. **H,** Fraction of generated compounds satisfying Lipinski's Rule of Five (Ro5) compared to FDA-approved compounds (dashed line). **I,** Distributions of the quantitative estimate of drug-likeness (QED) for generated and FDA-approved compounds. **J,** Distributions of the synthetic accessibility score (SAS) for generated and FDA-approved compounds. Lower SAS indicates easier chemical synthesis.

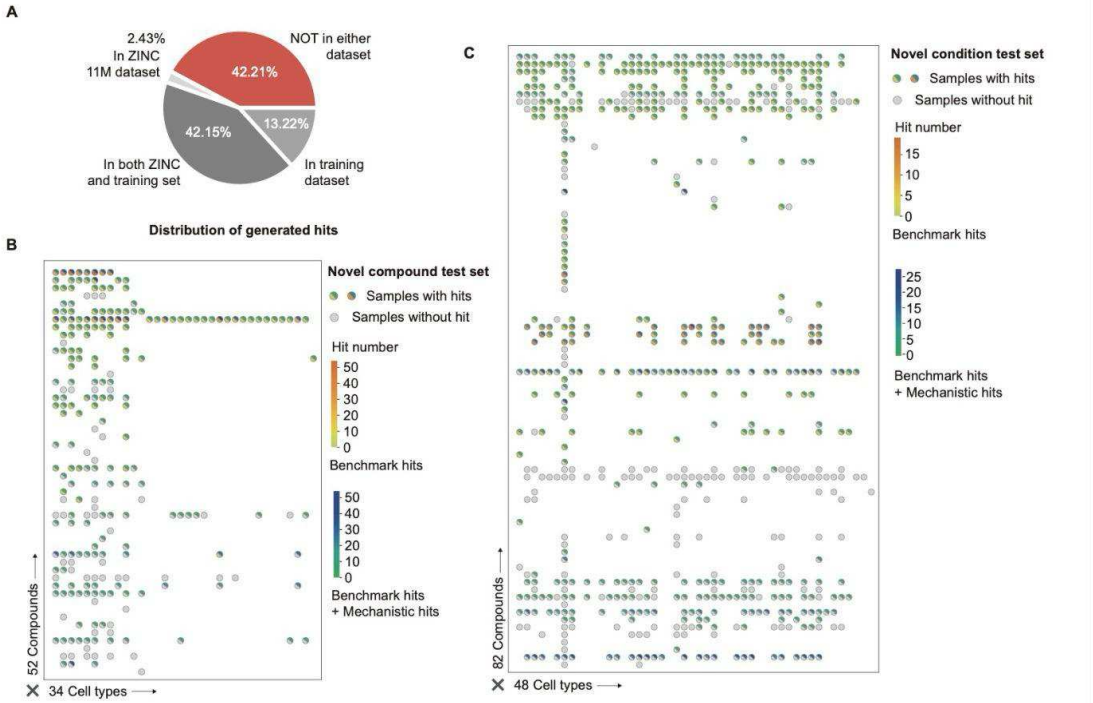
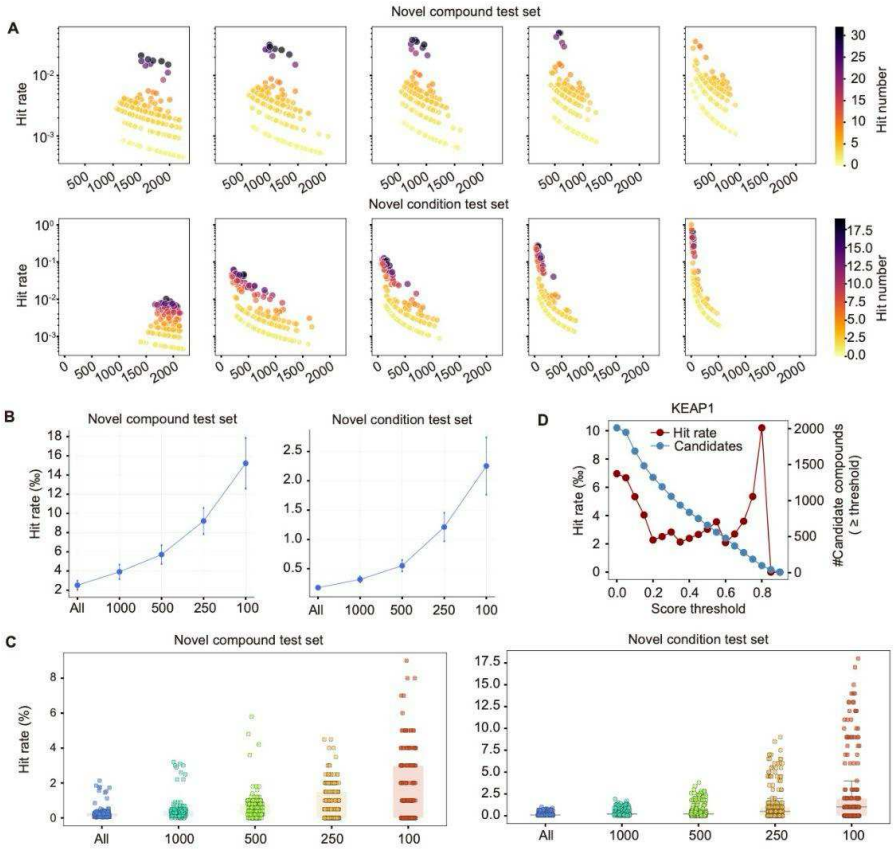


Figure S3. A, Distribution of hit molecules identified at a 0.6 similarity threshold, categorized by their presence in the training set and the ZINC 11M reference library. **B-C**, Per-sample hit distribution across the novel compound (**B**) and novel condition (**C**) test sets. Each point represents a perturbation-response condition, colored by the number of recovered compounds (Hit number). Comparison is shown between benchmark compounds alone (bottom/left) and benchmark compounds combined with mechanistic analogs (top/right).



1520

1521 **Figure S4. The scoring model prioritizes effective candidates. A**, Relationship
1522 between the number of prioritized candidates, hit rate, and hit number across individual
1523 samples in the novel compound (top) and novel condition (bottom) test sets. Color intensity
1524 is proportional to the number of hits, with darker shading indicating a higher number of
1525 hits. Y-axis is log-scaled for better representation. **B**, Average hit rate enrichment when
1526 restricting evaluation to the top-K candidates ranked by score in the novel compound (top,
1527 N=203) and novel condition (bottom, N=242) test sets. Error bars indicate 95% confidence
1528 intervals. **C**, Distribution of hit rates per response condition across different top-K selection
1529 regimes. Centre line, median; box limits, upper and lower quartiles; whiskers, 1.5×
1530 interquartile range. **D,** Effect of score-based filtering on candidate retention (blue) and hit
1531 rate (red) for KEAP1 genetic perturbation sample.

1532

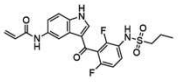
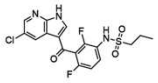
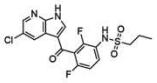
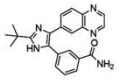
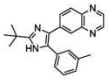
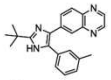
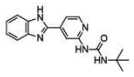
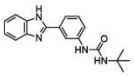
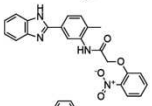
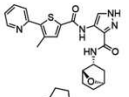
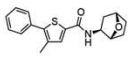
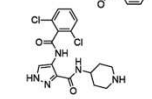
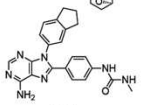
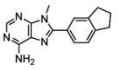
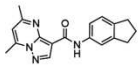
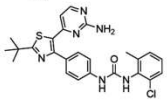
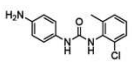
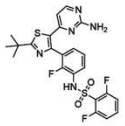
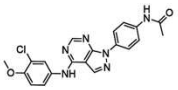
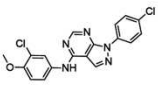
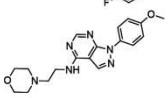
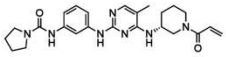
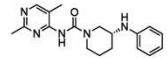
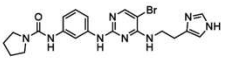
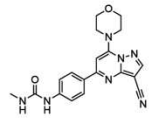
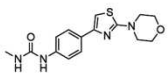
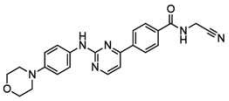
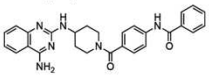
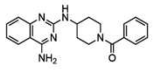
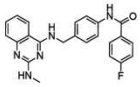
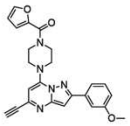
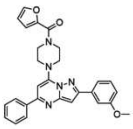
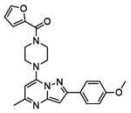
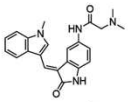
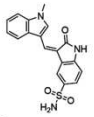
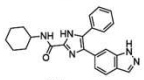
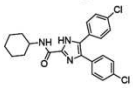
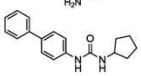
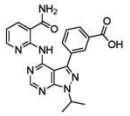
Generated molecule	Structure	Most similar analog in ZINC20-1.9B	Most similar analog in LINCS	Max similarity ZINC LINCS	Cell line	Percent inhibition at 1 μM/5 μM
GEMGen-001				0.606 0.606	LX-2 NIH3T3	-3.5% / 33.8% 16.9% / 23.1%
GEMGen-002				0.596 0.596	LX-2 NIH3T3	2.8% / 12.6% ND / 20.5%
GEMGen-003				0.692 0.417	LX-2 NIH3T3	13.6% / 33.7% 20.0% / 29.6%
GEMGen-004				0.563 0.358	LX-2 NIH3T3	17.4% / 38.3% 44.8% / 77.3%
GEMGen-005				0.524 0.342	LX-2 NIH3T3	-5.3% / 35.3% 25.3% / 29.8%
GEMGen-006				0.492 0.427	LX-2 NIH3T3	28.1% / 39.2% 56.4% / 58.9%
GEMGen-007				0.789 0.382	LX-2 NIH3T3	-5.5% / 75.7% 51.1% / 85.6%
GEMGen-008				0.535 0.458	LX-2 NIH3T3	9.8% / 38.6% 30.2% / 72.8%
GEMGen-009				0.515 0.354	LX-2 NIH3T3	14.9% / -0.5% 31.5% / 33.3%
GEMGen-010				0.827 0.378	LX-2 NIH3T3	64.3% / 78.0% 48.8% / 81.8%
GEMGen-011				0.794 0.677	LX-2 NIH3T3	74.5% / 110.1% 41.7% / 49.7%*
GEMGen-012				0.593 0.575	LX-2 NIH3T3	45.1% / 89.5% 32.1% / 72.0%
GEMGen-013				0.603 0.361	LX-2 NIH3T3	23.3% / 42.2% 35.4% / 47.4%
GEMGen-014				0.545 0.474	LX-2 NIH3T3	-31.4% / -42.9% ND / ND
GEMGen-015				0.690 0.580	LX-2 NIH3T3	8.3% / 17.3% ND / ND
GEMGen-016				0.632 0.600	LX-2 NIH3T3	52.6% / 48.3% 20.3% / -3.5%
GEMGen-017				0.662 0.549	LX-2 NIH3T3	6.3% / 6.4% 11.1% / 13.6%

Figure S5. *De novo* synthesized molecules for experimental testing. Summary of the structural analysis and cellular activities of *de novo* synthesized molecules.

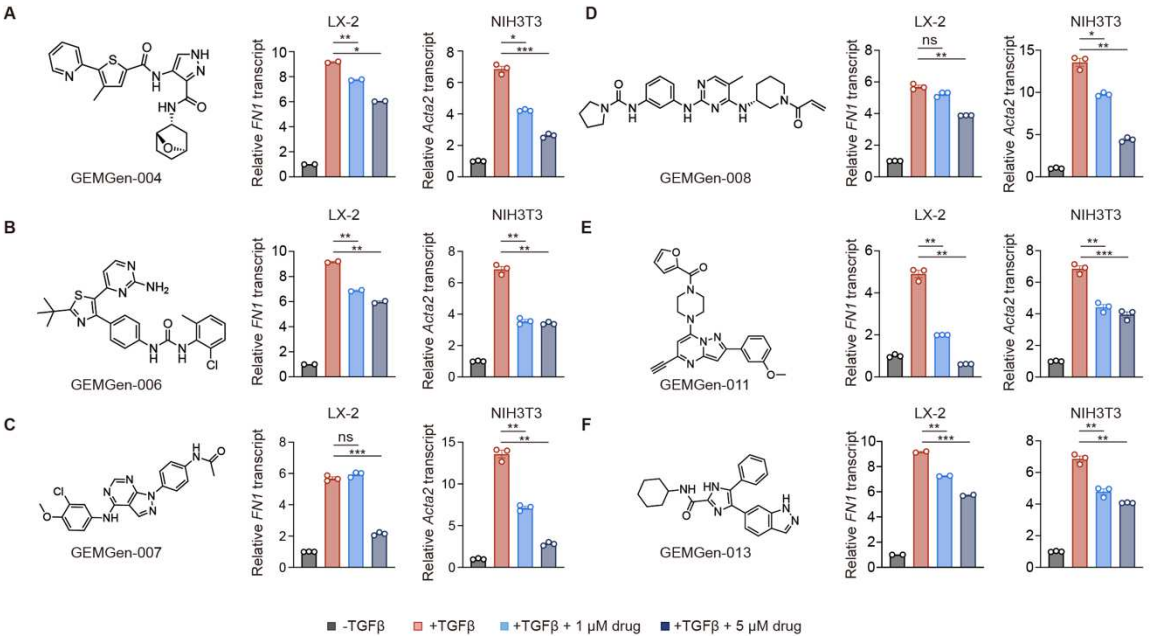


Figure S6. *De novo* synthesized small molecules attenuate fibrosis phenotype. A-F, Evaluation of the effects of active antifibrotic compounds generated by GEMGen by qPCR analysis of *FN1 and Acta2* transcript levels in TGFβ-stimulated LX-2 cells and NIH-3T3 cells, respectively. The structure of each molecule is shown together with the gene expression plot. *P*-values are calculated with one-way ANOVA with Dunnett's correction, **P* < 0.05, ***P* < 0.01, ****P* < 0.001.

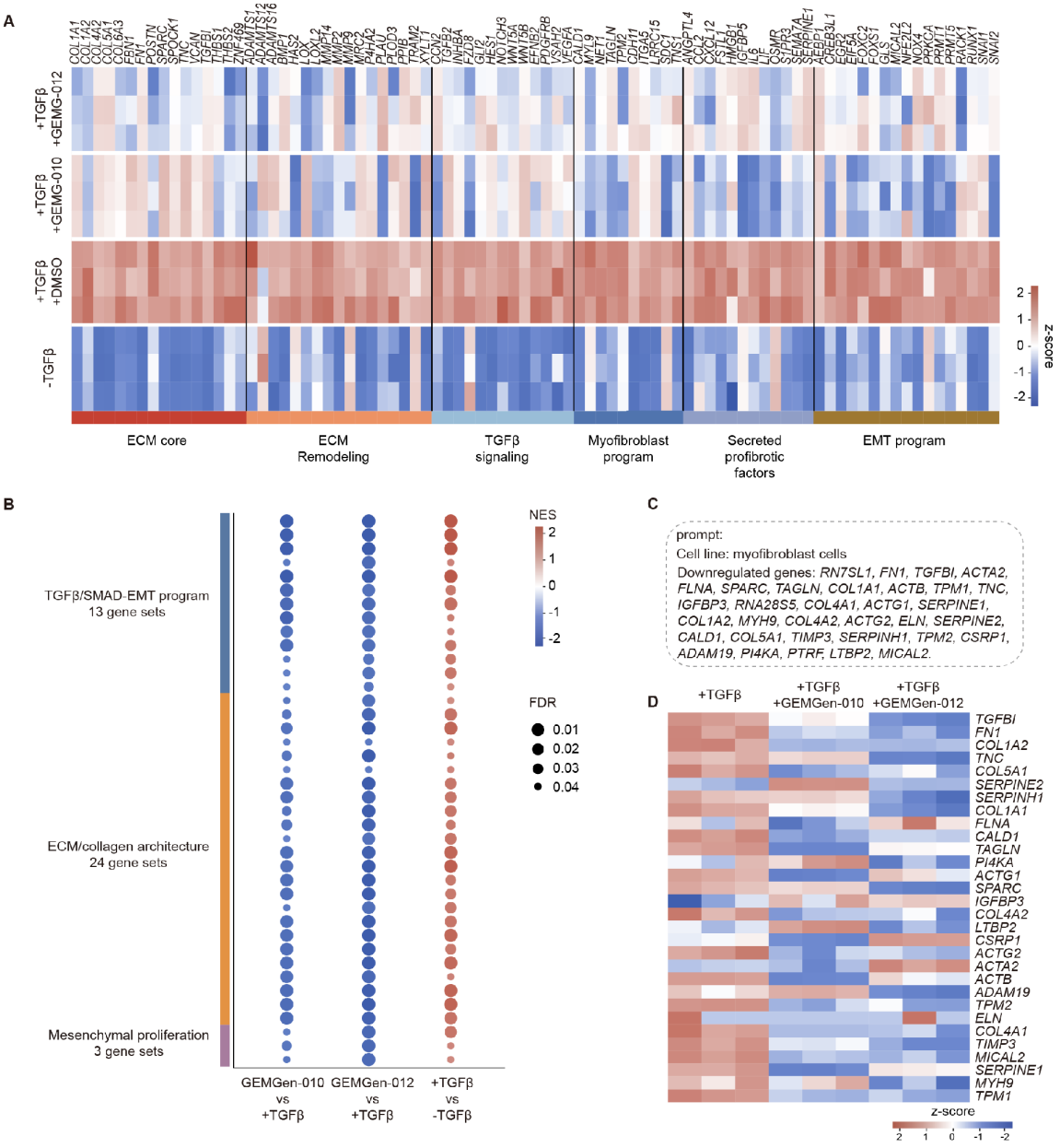


Figure S7. GEMGen-010 and GEMGen-012 suppress the fibrosis-associated transcriptional program. A, Heatmap of RNA-sequencing results showing relative abundance of fibrosis-associated transcripts from TGFβ-stimulated LX-2 cells treated with GEMGen-010, GEMGen-012, or DMSO control. Colors indicate normalized transcript levels. Genes are clustered based on gene function, which is annotated at the bottom. n=3. **B**, Dot plot summary of gene set enrichment analyses from RNA-sequencing results of TGFβ-stimulated LX-2 cells treated with GEMGen-010 (5 μM), GEMGen-012 (5 μM), or DMSO control. Information of gene sets is provided in **Table S4**. **C**, Information used as prompts for GEMGen generation of antifibrotic molecules. **D**, Heatmap of RNA-sequencing results of TGFβ-stimulated LX-2 cells treated with GEMGen-010 (5 μM), GEMGen-012 (5 μM), or DMSO control, showing relative abundance of prompt genes.

Supplementary Files

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

- TableS1.xlsx
- TableS2.xlsx
- TableS3.csv
- TableS4.xlsx