







A genotype-to-drug diffusion model for generation of tailored anti-cancer small molecules

Received: 10 September 2024

Accepted: 22 May 2025

Published online: 01 July 2025

Hyunho Kim ^{1,2}, Bongsung Bae ¹, Minsu Park ¹, Yewon Shin ³,
Trey Ideker ^{4,5,6} ✉ & Hojung Nam ^{1,3} ✉

Despite advances in precision oncology, developing effective cancer therapeutics remains a significant challenge due to tumor heterogeneity and the limited availability of well-defined drug targets. Recent progress in generative artificial intelligence (AI) offers a promising opportunity to address this challenge by enabling the design of hit-like anti-cancer molecules conditioned on complex genomic features. We present Genotype-to-Drug Diffusion (G2D-Diff), a generative AI approach for creating small molecule-based drug structures tailored to specific cancer genotypes. G2D-Diff demonstrates exceptional performance in generating diverse, drug-like compounds that meet desired efficacy conditions for a given genotype. The model outperforms existing methods in diversity, feasibility, and condition fitness. G2D-Diff learns directly from drug response data distributions, ensuring reliable candidate generation without separate predictors. Its attention mechanism provides insights into potential cancer targets and pathways, enhancing interpretability. In triple-negative breast cancer case studies, G2D-Diff generated plausible hit-like candidates by focusing on relevant pathways. By combining realistic hit-like molecule generation with relevant pathway suggestions for specific genotypes, G2D-Diff represents a significant advance in AI-guided, personalized drug discovery. This approach has the potential to accelerate drug development for challenging cancers by streamlining hit identification.

Generative AI is revolutionizing diverse fields by enabling the creation of novel synthetic data across modalities, including images, text, audio, and video¹. This capability allows generative AI to expand the boundaries of what can be efficiently created, with particularly transformative potential in domains that rely on developing new ideas. The evolution from single-modality generative models (e.g., GPTs^{2–4}) to multi-modality models (e.g., DALL-Es^{5–7}) and more complex platforms like GPT-4⁸ further enables customized data generation under diverse and complex conditions.

Current approaches to AI-guided generative drug discovery can be broadly categorized into target-based and phenotype-based methods, each with its own strengths and limitations. The first, target-based approach aims to identify candidate compounds that effectively engage with specific protein targets implicated in a disease. In these studies, molecular generative models are fine-tuned using compounds with known binding to target proteins^{9–13}, or using reinforcement learning (RL) based on predicted binding affinity to

¹Department of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea. ²Division of Advanced Predictive Research, Korea Institute of Toxicology, Daejeon, Republic of Korea. ³AI Graduate School, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea. ⁴Department of Medicine, University of California San Diego, San Diego, CA, USA. ⁵Department of Bioengineering, University of California San Diego, San Diego, CA, USA. ⁶Department of Computer Science and Engineering, University of California San Diego, San Diego, CA, USA.

✉ e-mail: tideker@ucsd.edu; hjnam@gist.ac.kr

generate compounds likely to bind to the desired targets^{14–20}. More recently, following the successful prediction of protein 3D structures by DeepMind's AlphaFold^{21,22}, there has been a proliferation of models capable of generating candidate compounds likely to recognize binding pockets of specific targets based on their 3D structures^{23–28}. However, these target-based approaches face challenges in complex diseases like cancer, where the best targets are often unknown, and off-target activity can be a significant problem^{29,30}.

The second, phenotype-based approach leverages generative AI to identify compounds that achieve a desired phenotype. This approach complements the target-based method by providing an effective solution when prior knowledge of specific targets is limited, focusing directly on desired phenotypic outcomes. Due to its nature, this approach requires simultaneous consideration of multi-modal features such as genetic and chemical features. While extensive research in predictive modeling has established effective frameworks for joint learning of genetic and chemical features^{31–33}, the more challenging task of generative modeling has recently gained attention for its ability to create entirely new therapeutic candidates, moving beyond the limitations of simply searching through existing compound libraries. In this context, emerging generative approaches leverage large-scale multi-modal datasets measuring drug-induced gene expression (e.g., LINCS L1000³⁴) or drug-induced growth response (e.g., GDSC³⁵, CTRP^{36–38}, and NCI60³⁹).

Recent studies have employed various generative models, including diffusion models, to design compounds capable of inducing a desired phenotype^{40–48}. One class of models^{40–45} aims to generate novel compounds with desired gene expression profiles in specific cancer cell lines. These approaches typically focus on generating novel structures with similar effects by using target gene knock-out-induced gene expression profiles as input conditions. However, while these drug-induced phenotype-based models have shown promise, they are not designed to address the critical need for identifying therapeutic candidates based on cancer baseline conditions. This limitation is particularly crucial for developing therapeutic candidates for intractable cancers, where precise therapeutic targets have yet to be identified.

Thus, several studies have attempted to identify or generate novel anticancer hit compounds based on cancer baseline conditions using conditional variational autoencoder (VAE)⁴⁶, or RL-based generative models^{47,48}. However, these approaches also have several limitations. First, the applicability of models trained on cell line data to clinical contexts is limited, as they rely on gene expression data, which is rarely available in real-world clinical settings, making these models impractical for use. Second, gene expression data frequently exhibits biases due to batch effects or other environmental factors⁴⁹. Third, when methods like RL are used to guide compound generation^{47,48}, there is a risk that the drug response predictor may perform poorly for unseen drugs, potentially leading the generator to learn in undesirable directions.

To address these limitations, we introduce Genotype-to-Drug Diffusion (G2D-Diff), a generative model designed to boost the efficiency of the hit-identification stage by streamlining the process of discovering promising drug candidates in phenotype-based drug discovery. G2D-Diff leverages drug response data to learn to generate hit-like candidates for specific cancer samples. In contrast to prior methods using gene expression, G2D-Diff utilizes genetic alteration information from clinically relevant genes, increasing clinical utility and generalizability. Additionally, the model's attention mechanism allows identification of critical genes or pathways related to the desired drug response, enhancing its interpretability. G2D-Diff uses a diffusion-based generative model⁵⁰ as its backbone to directly learn the distribution of hit-like compounds, avoiding the need for a separate predictor during training and generation (Fig. 1a).

Similar to how text-to-image models input a conditional text prompt to generate corresponding images, our model inputs somatic alteration genotypes and the desired drug response (stratified in five response classes: very sensitive, sensitive, moderate, resistant, and very resistant) to generate the appropriate compounds. The model architecture consists of two main components: a VAE to learn a latent representation of chemical compounds (Fig. 1b), and a conditional diffusion model that generates compound latent vectors based on the input genotype and desired response (Fig. 1c). The VAE is pre-trained on a large chemical structure dataset (~1.5 million) of known compounds, producing an efficient chemical latent space. Subsequently, the genotype-to-drug latent diffusion model (Fig. 1c) receives a genotype and a desired response, then processes them using a condition-encoder module, which generates a numerical condition encoding for the given input. This encoding serves as a condition when the diffusion model creates latent vectors of compounds. The generated vector is then decoded into a simplified molecular-input line-entry system (SMILES) format using the previously constructed chemical VAE decoder, making the structure accessible for human verification.

To improve the generalizability of the condition encoder to unseen genotypes, we also introduce a pre-training approach based on contrastive learning (Fig. 1d). Inspired by the CLIP contrastive learning framework⁵¹, our approach aims to enhance the model's generalizability to arbitrary genotypes and responses by ensuring the condition encoding captures not only basic information about the genotype and response but also the condition-matching drugs' structural information. The latent representation evolves over progressive diffusion steps to gradually decrease the predicted AUC, and thus increase the predicted sensitivity (Fig. 1e).

Results

Evaluating the molecular latent space of the chemical VAE

Our chemical VAE demonstrates proficient capability in generating feasible, drug-like compounds, as evidenced by the comprehensive evaluation of its reconstruction and random generation performance.

In the reconstruction task, 1000 molecules were randomly selected from a validation set, the subset of the chemical structure dataset used for training the chemical VAE (Methods). These molecules were encoded into latent vectors by the VAE, then decoded to test if the model can reconstruct the original input. For random generation, molecules were generated by decoding 1000 randomly sampled latent vectors from the normal distribution (Methods). As a result, we found that the chemical VAE achieved 1.00 validity, 1.00 uniqueness, and 0.99 reconstruction success rate in the reconstruction task, versus 0.86 validity, 1.00 uniqueness, 1.00 novelty, and 0.89 diversity in the random generation task. While the random generation shows relatively lower validity compared to reconstruction, this is primarily due to the inherent challenges in SMILES-based molecular representation. Although alternative representations like SELFIES⁵² have shown improved validity, recent studies have demonstrated that the slight decrease in validity does not significantly impact the model's ability to perform conditional generation tasks⁵³.

We next evaluated the randomly sampled molecules' drug-like properties, such as quantitative estimates of drug-likeness (QED)⁵⁴, synthetic accessibility scores (SAS)⁵⁵, and calculated logarithm of the partition coefficient (LogP)⁵⁶. We observed that a similar percentage of molecules (out of 1000) fell within the acceptable ranges across all three quantitative metrics when compared to the molecules of the validation set (Supplementary Fig. 1).

Given the strong performance across the evaluation metrics, we proceeded to integrate this chemical VAE as the foundation for our conditional latent diffusion model. The comprehensive evaluation results indicate that our chemical VAE successfully constructs a well-structured molecular latent space that effectively captures the

essential features of drug-like compounds, making it suitable for the subsequent generative processes in G2D-Diff.

General performance evaluation of G2D-Diff

G2D-Diff demonstrates exceptional performance in generating diverse, feasible, and condition-matching compounds, significantly outperforming existing methods across multiple evaluation metrics.

First, the condition encoder, a key component of G2D-Diff, was pre-trained using tumor cell-line-centric drug response data (Methods). To evaluate its performance, we conducted two main analyses. First, we examined how it generates distinguishable encodings for genotype-response class conditions through principal component analysis (PCA, Fig. 2a, b). The results show that the PC1 axis

distinguishes conditions based on response class (sensitive, moderate, or resistant) (Fig. 2a), while the PC2 axis differentiates conditions according to genotype differences (Fig. 2a, Supplementary Fig. 2), with PC3 providing further stratification within the sensitive and resistant categories (Fig. 2b). Second, we evaluated the model's ability to identify condition-matching drugs from the condition encodings (Fig. 2c). The top 5 and top 10 potential drug candidates were determined by cosine similarity in the shared space, and the odds ratio between the odds of a drug being in the top 5 or top 10 most similar drugs for a given condition and the odds of a drug having the matching condition in the entire dataset was used to measure accuracy. The odds ratios, consistently above zero across all conditions (Fig. 2c, Supplementary Table 1), were exceptionally high for sensitive conditions, highlighting

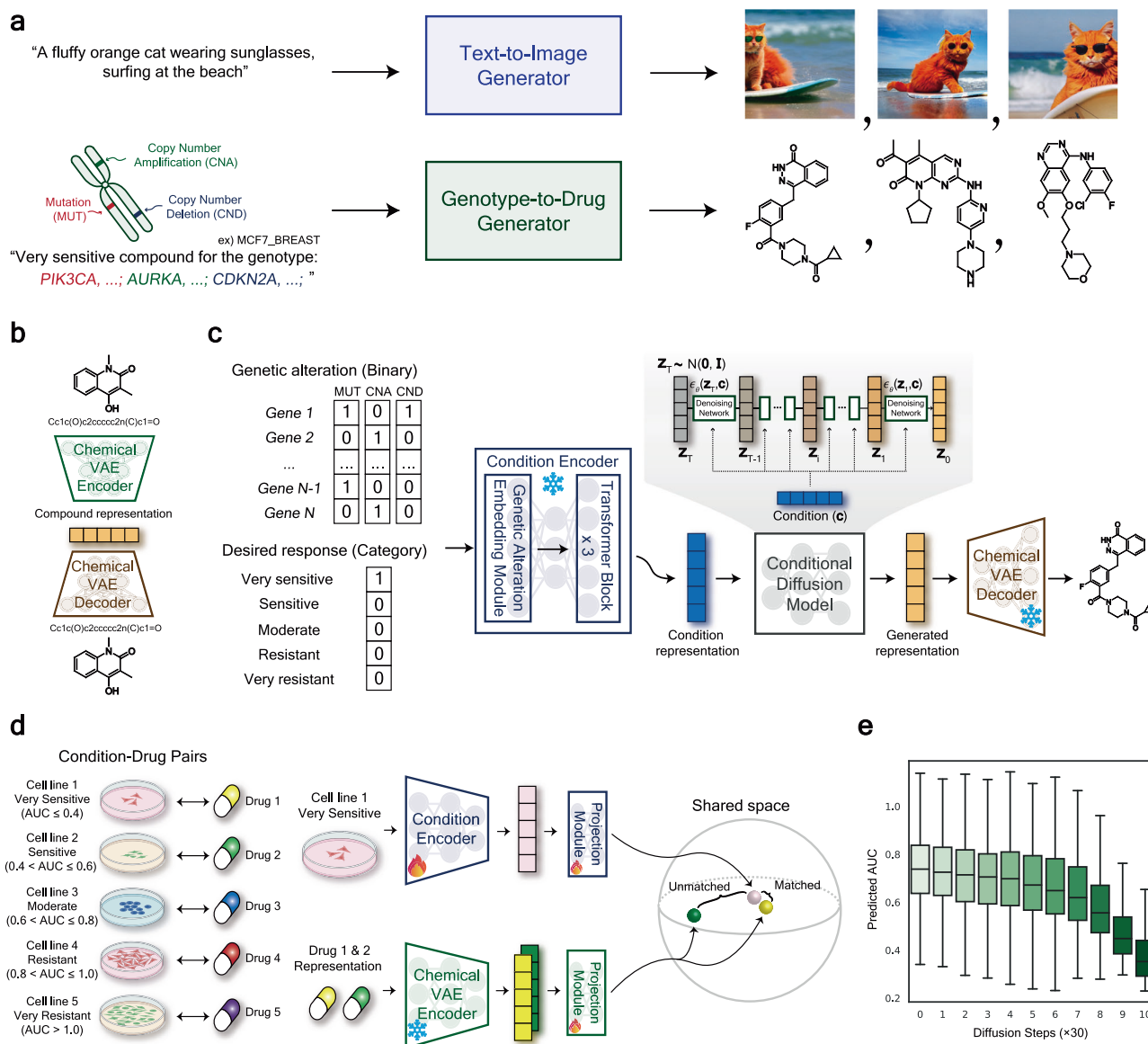


Fig. 1 | Workflow of G2D-Diff. **a** General concept of G2D-Diff aligned with the text-to-image generator. **b** General architecture of the Chemical VAE, which enables the diffusion process in the latent space. **c** Workflow of our core module, the genotype-to-drug latent diffusion model. **d** Detailed graphical explanation of how the condition-drug pair data is constructed. The left part of the panel illustrates how the condition encoder was pre-trained using contrastive learning. **e** Changes in predicted AUC according to diffusion steps when generating with the example condition (MCF7, very sensitive). We generated 1000 molecular latent vectors and

examined the vectors produced at intermediate stages by every 30 steps out of 300 denoising steps. Box plots show median (center line), interquartile range (25th and 75th percentiles, box limits), whiskers (e.g., 1.5 times the interquartile range from box limits). Note that the snowflake icon indicates that the corresponding submodule is frozen during training and inference, while the burning icon signifies that the learnable parameters of the submodule are being optimized through training. Source data are provided as a Source Data file. Created in BioRender. Nam, H. (2025) <https://BioRender.com/bjy6emx>.

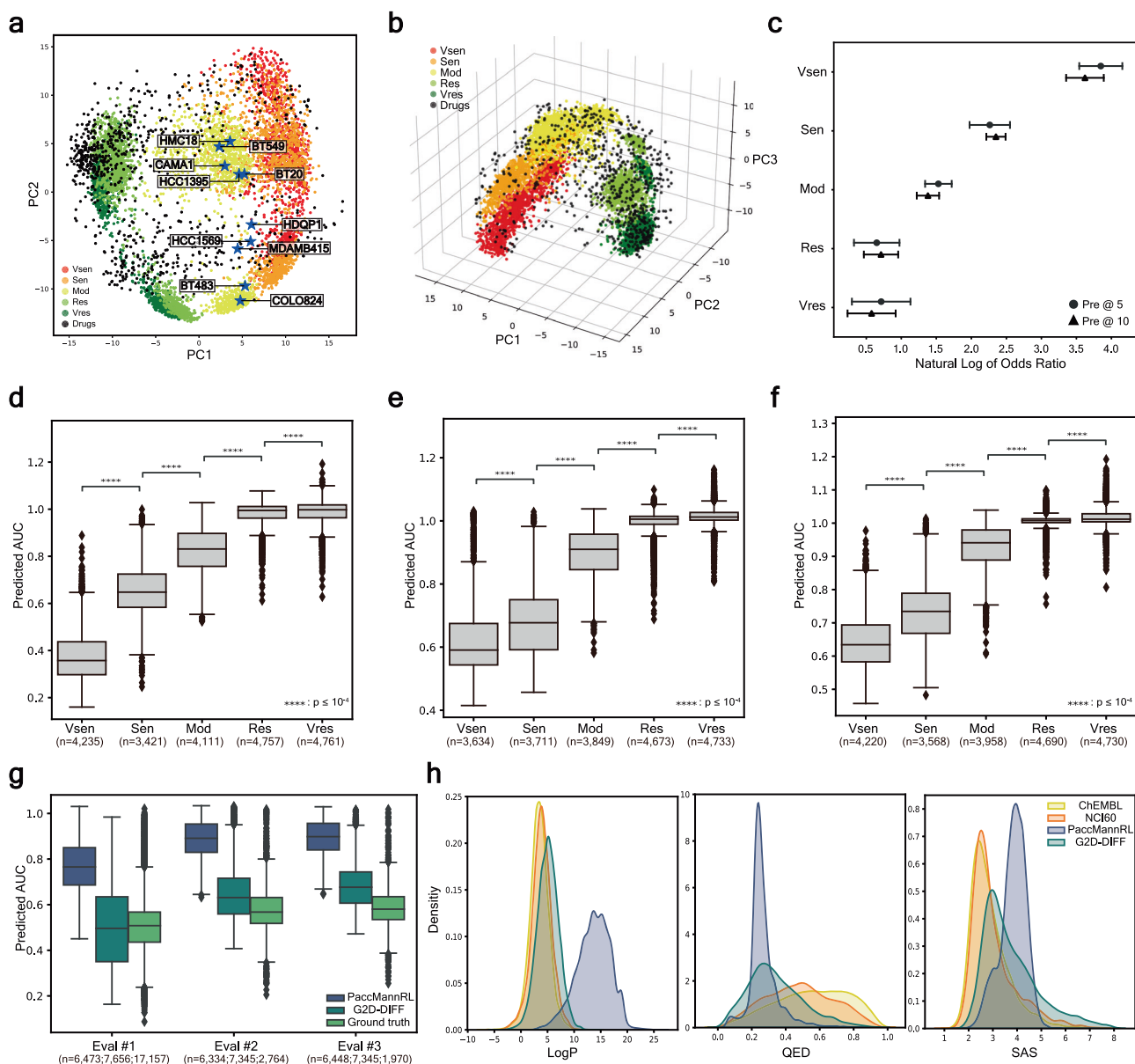


Fig. 2 | Comprehensive performance evaluation results of G2D-Diff. **a** 2D PCA of condition and drug encodings in the shared space. Each colored point represents a condition (cell line with a desired drug response level), while black points denote drugs. The PC1 axis distinguishes the conditions based on the response class: sensitive, moderate, or resistant. In contrast, the PC2 axis differentiates conditions according to the genotype difference. Blue stars mark breast cancer cell lines with different genotypes in the Moderate response class. Other response class results are shown in Supplementary Fig. 2. **b** 3D PCA of condition and drug encodings in the shared space. The PC3 provides further stratification within the sensitive and resistant categories, dividing between very sensitive and sensitive, as well as very resistant and resistant conditions. **c** Natural logarithm of the odds ratio for precision at 5 (circles) and precision at 10 (triangles). The point represents the mean value, and the error bands represent the standard deviation. **d-f** Distribution of the

predicted AUCs for conditionally generated compounds across all conditions in evaluation set 1, evaluation set 2, and evaluation set 3, respectively. Two-tailed Mann-Whitney *U* test was conducted. **g** Distribution of the predicted AUCs for conditionally generated sensitive compounds from PaccMannRL (blue) and G2D-Diff (dark green), with the predicted AUCs of ground truth sensitive compounds (green) used as a positive control. **h** Density of LogP, QED, and SAS for compounds generated from PaccMannRL (blue) and G2D-Diff (dark green), and randomly sampled compounds from ChEMBL (yellow) and NCI60 (orange). Box plots show median (center line), interquartile range (25th and 75th percentiles, box limits), whiskers (e.g., 1.5 times the interquartile range from box limits), and outliers (points outside whiskers). PC principal component, Vsen very sensitive, Sen sensitive, Mod moderate, Res resistant, Vres very resistant, Pre Precision, Eval Evaluation. Source data are provided as a Source Data file.

the precision of the condition encoder in identifying drugs with high efficacy to specific genotypes. These results demonstrate that the condition encoder has successfully learned to generate distinguishable condition representations that capture the relationship between genotypes and drug response levels.

Next, to evaluate the generative performance of G2D-Diff, which was trained using a drug-centric drug response dataset (Methods), we generated 1000 molecular latent vectors for every condition of three

evaluation sets (Methods). We decoded these latent vectors into SMILES strings using the chemical VAE decoder. We then evaluated the two types of performance: basic generation performance and conditional generation performance.

For the basic generation quality assessment, we evaluated the following metrics, such as validity, uniqueness, novelty, diversity, Fréchet ChemNet Distance (FCD)⁵⁷, and Optimal Transport Distance (OTD)¹⁵ (Methods). As a result, we found that G2D-Diff showed robust

Table 1 | Comparison of performance for sensitive molecule (AUC \leq 0.6) generation

Cell lines	Model	Chemical-likeness metrics				Objective-related metrics	
		Validity	Uniqueness	Novelty	Diversity	FCD ^a	OTD ^b
Evaluation set 1	PaccMannRL	0.703	0.915	1.0	0.637	54.317	7.229
	G2D-DIFF	0.764	0.835	0.997	0.870	9.136	4.858
Evaluation set 2	PaccMannRL	0.695	0.912	1.0	0.631	56.130	7.218
	G2D-DIFF	0.765	0.823	0.997	0.866	11.463	4.999
Evaluation set 3	PaccMannRL	0.708	0.911	1.0	0.632	53.516	7.185
	G2D-DIFF	0.757	0.824	0.997	0.865	13.278	5.112

^aFréchet ChemNet Distance. The control scores were measured as follows: Evaluation set 1: 9.332 (\pm 0.057), Evaluation set 2: 15.486 (\pm 0.095), and Evaluation set 3: 17.198 (\pm 0.088).

^bOptimal Transport Distance. The control scores were measured as follows: Evaluation set 1: 4.910 (\pm 0.006), Evaluation set 2: 5.260 (\pm 0.005), and Evaluation set 3: 5.371 (\pm 0.007).

performance across all metrics for every response class category (Supplementary Table 2). The result indicates that the model can produce valid and diverse molecules consistently. We also compared our model with PaccMannRL⁴⁸, a gene expression-based generative model for hit-like anticancer molecules. Our model outperformed PaccMannRL in terms of diversity, FCD, and OTD while exhibiting comparable validity and novelty (Table 1). This indicates that our model can generate not only diverse molecules but also ones similar to real actives.

To evaluate the fitness condition of the generated compounds, we additionally developed a separate genotype-to-drug response prediction model (G2D-Pred, Supplementary Fig. 3a, Methods). The G2D-Pred's robust performance (Supplementary Fig. 3b) made it suitable for assessing all generated compounds. The predicted AUC values for the generated compounds under every condition in three evaluation sets were analyzed (Fig. 2d–f, Supplementary Fig. 4), revealing a progressive increase in the predicted AUC from very sensitive to very resistant classes, with statistically significant differences (two-tailed Mann–Whitney U test, $p < 10^{-4}$). These results demonstrate that G2D-Diff has successfully learned to generate molecular structures that correlate with the desired AUC conditions for arbitrary genotypes. Moreover, in evaluation sets 2 and 3, the model learned the condition-specific sophisticated distributional patterns to generate decent hit-like molecules despite structures of compounds were not being trained on ground truth data (Fig. 2e, f, Supplementary Fig. 5). These results confirm the generalizability of G2D-Diff, proving that our model can effectively respond to unseen genotypes.

We further compared the condition fitness and feasibility of molecules generated by G2D-Diff and PaccMannRL. The distribution of predicted AUC values for compounds generated under sensitive conditions (very sensitive and sensitive) by both models reveals that the compounds generated by our G2D-Diff model showed significantly lower AUC values and were also closely aligned with the ground truth, indicating superior condition fitness (Fig. 2g). To ensure unbiased results, we employed HiDRA, a drug response predictor based on gene expression profile of cell lines⁵⁸, manually re-trained using the same dataset (Supplementary Fig. 3c). We confirmed that the compounds generated by G2D-Diff exhibited better condition fitness than those generated by PaccMannRL when assessed using an independent gene expression-based drug response predictor (Supplementary Fig. 3d, Supplementary Table 3).

Regarding feasibility, we compared the distributions of LogP, QED, and SAS values of the molecules generated by each model (Fig. 2h). The molecules generated by G2D-Diff exhibited closer distributions to control datasets from ChEMBL and NCI60, implying high feasibility (Supplementary Table 4). Furthermore, compared to the PaccMannRL, the molecules generated by G2D-Diff showed generally higher QED and lower SAS values, indicating that G2D-Diff can generate more drug-like and synthetically accessible molecules. In summary, we have confirmed that G2D-Diff exhibits robust performance in generating hit-like anticancer molecules in all aspects, both absolute and relative.

Quality assessment of generated compounds

G2D-Diff demonstrates an exceptional ability to generate condition-matching compounds with unique and diverse chemical scaffolds. We analyze the compounds generated under the conditions for evaluation set 1, which consists of well-trained conditions that the model has extensively learned. We first generated 120,000 molecules (1000 molecules each across 0–11 CFG scales for both very sensitive and sensitive conditions). To assess the quality and relevance of the generated compounds, we compare generated molecules with ground truth compounds in two key aspects: chemical structural features, including scaffold distribution, and functional properties, including pharmacophore and physicochemical characteristics.

First, we examined the scaffold distribution of the generated and ground truth compounds, selecting very sensitive and sensitive ground truth compounds for cell lines in the evaluation set 1. The scaffolds were extracted using the Bemis–Murcko scaffold extraction algorithm⁵⁹. The results show that generated compounds exhibited higher scaffold diversity, considering the number of extracted scaffolds is nearly equal to the total number of compounds in both response classes, compared to the ground truth compounds (Supplementary Table 5). The generated compounds had distinct scaffolds rarely overlapping between response classes or with ground truth compounds, indicating the model's capability to produce diverse, response class-specific scaffolds.

G2D-Diff demonstrates a remarkable ability to generate compounds tailored to specific conditions. The heatmaps visualize scaffold frequency across different cell lines, where the y-axis represents individual scaffolds, the x-axis shows different cell lines, and the color intensity indicates the frequency of each scaffold. As illustrated by the frequency-based cluster maps of scaffolds for very sensitive (Fig. 3a, b) and sensitive compounds (Supplementary Fig. 6a, b), a striking contrast is observed between ground truth and generated compounds: while ground truth compounds show redundant scaffolds shared among multiple cell lines, G2D-Diff generates compounds with unique scaffolds specific to each cell line, rarely overlapping with those of same cell line in the ground truth compounds. This cell line-specificity is further confirmed by the cell line-specific scaffold statistics and structural similarity analysis among generated compounds (Supplementary Table 6, Supplementary Fig. 7a, b). For instance, in the very sensitive class, G2D-Diff generated nearly 12,000 scaffolds (same as the number of generated compounds per condition) for each cell line, but only 29, 18, 26, 24, and 17 scaffolds overlapped with the ground truth scaffolds from the same cell lines. Similar trends are observed in the sensitive class (Supplementary Table 6).

Particularly, we demonstrated that these scaffolds are not merely the result of random molecular generation processes, as evidenced by our additional comparative analysis. Our analysis compared compounds generated by G2D-Diff and a random chemical VAE generator against ground truth compounds in terms of structural similarity and pharmacophore similarity. While G2D-Diff-generated compounds showed marginal increase in structural similarity to ground truth, they

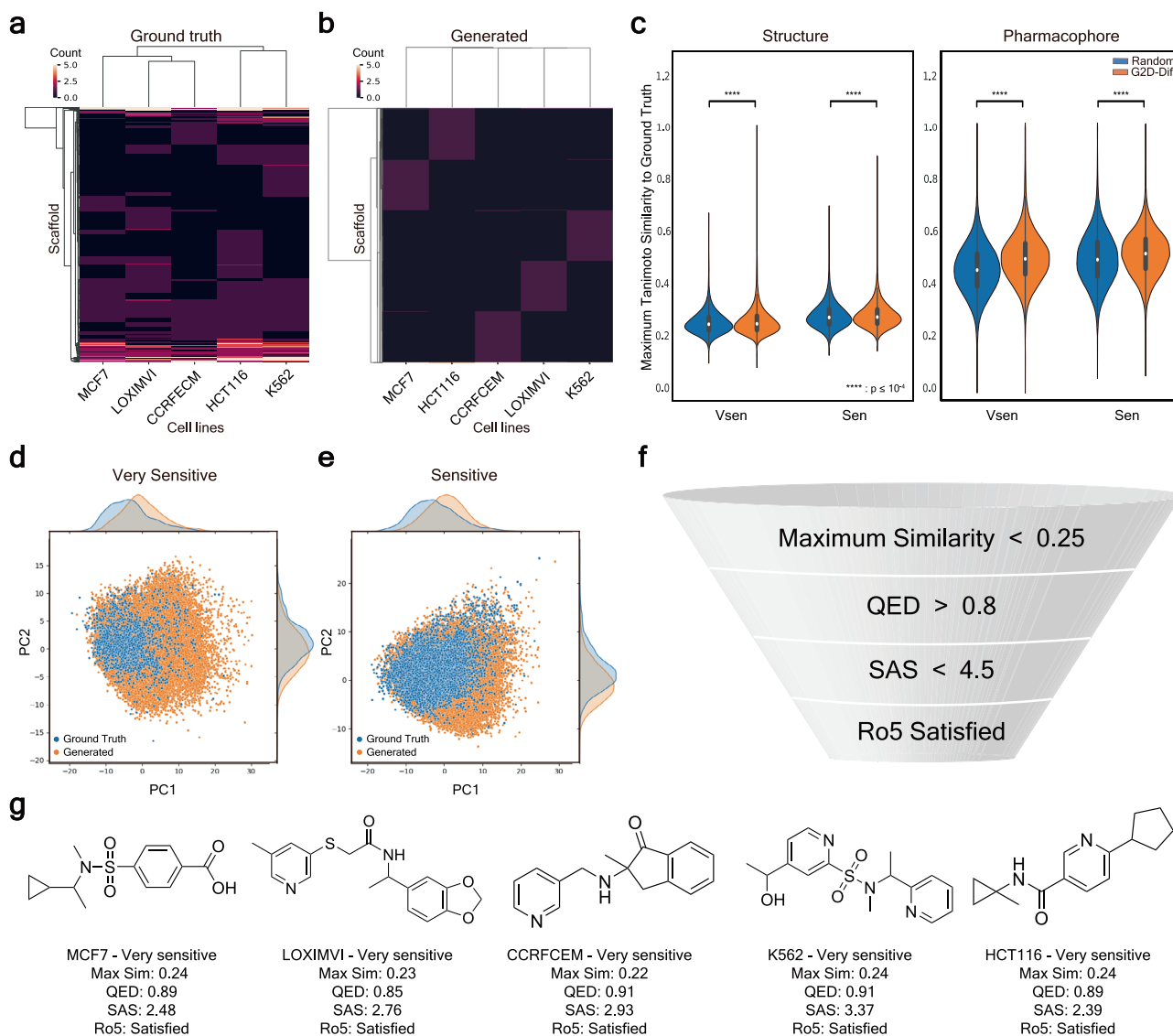


Fig. 3 | Structural analysis of generated compounds by G2D-Diff and potential candidates. **a** Frequency-based cluster map for the scaffolds of the ground truth very sensitive compounds. **b** Frequency-based cluster map for the scaffolds of the generated very sensitive compounds. For (**a**, **b**), each column indicates a cell line in the evaluation set 1, and each row is the frequency count of a unique scaffold. Scaffolds with larger frequencies are represented by brighter colors. **c** Maximum structural and pharmacophore similarity comparison between randomly generated compounds by Chemical VAE and generated compounds by G2D-Diff. Sample size N is 60,000 for each response class, and statistical significance from the one-sided Mann–Whitney U test is indicated as asterisks. Violin plots show the distribution

density (width of the plot) along with median (white dot), interquartile range (25th to 75th percentile; thick bar), and the minimum and maximum values within 1.5 times the interquartile range (thin line). **d** 2D PCA plot of physicochemical properties for very sensitive compounds. **e** 2D PCA plot of physicochemical properties for sensitive compounds. For (**d**, **e**), RDKit descriptors are used for physicochemical properties after standardization. Generated and ground truth compounds are represented as blue and orange colors, respectively. Marginal distributions of samples are depicted in both PC axes. **f** Criteria for selecting potential drug-like hit candidates for each cell line. **g** Selected potential drug-like hit candidates correspond to the query cell lines. Source data are provided as a Source Data file.

demonstrated substantially higher pharmacophore similarity compared to randomly generated compounds (Fig. 3c). This selective preservation of pharmacophoric features, despite generating structurally diverse scaffolds, suggests that G2D-Diff effectively learned to generate high-dimensional chemical patterns rather than merely imitating molecular structures of ground truths. We attribute this sophisticated pattern recognition capability to the model's latent diffusion architecture, which enables effective processing of chemical features in the chemical VAE-derived latent space. In summary, the substantial increase in unique scaffolds for each cell line, while preserving key pharmacophoric features, underscores G2D-Diff's capacity to produce diverse, drug-like, genotype-specific compounds.

Furthermore, we investigated generated compounds using PCA based on physicochemical properties to confirm that the generated

compounds exhibit previously unseen chemical structures while maintaining feasible physicochemical properties. We found that the physicochemical properties of generated compounds exhibit a refined distribution, covering the entire distribution of ground truth compounds (Fig. 3d, e). It demonstrates that the model can generate various structures of compounds with proper properties similar to the known sensitive compounds. Similar results are also shown in the PCA for each cell line (Supplementary Fig. 7c). Furthermore, toxicity assessment using ADMETlab version 3.0⁶⁰ revealed that the generated compounds exhibited significantly lower in vivo toxicity profiles compared to ground truth compounds, while maintaining drug-like physicochemical properties (Supplementary Fig. 8). This indicates that G2D-Diff generates compounds to have targeted anticancer effects while avoiding broad toxicity.

To narrow down realistic hit-like candidates for each cell line, we applied stringent filtering criteria (Fig. 3f). We screened the generated compounds based on the following conditions: structurally distinct from ground truths (Maximum Tanimoto similarity <0.25), enhanced drug-likeness (QED >0.8 , adherence to Lipinski's rule of five), and ease of synthesis (SAS $<4.5^{61}$). We then conducted retrosynthesis prediction (Methods) for the candidates and filtered out compounds with a synthetic route depth greater than four. We identified ~10–30 candidates for each cell line that met all criteria (Supplementary Data 1). Examples of selected compounds that are likely very sensitive to each cell line are shown (Fig. 3g), and predicted synthetic routes of those compounds are also shown (Supplementary Fig. 9). These candidate compounds are realistic, showing both novelty and great drug-like properties simultaneously. These findings suggest that our model is effective in generating viable and innovative drug-like compounds.

Zero-shot generative case studies for triple-negative breast cancer

Triple-negative breast cancer (TNBC) was selected as our case study due to its well-documented poor prognosis and urgent need for therapeutic drug development, making it an ideal candidate cancer type to validate our approach in a real-world setting⁶². Therefore, we conducted two zero-shot case studies to demonstrate the generalizability and practicality of G2D-Diff in TNBC cases. In the first case study, we aimed to generate hit-like candidate molecules for the TNBC cell line HS578T, even in extreme conditions with limited data (Fig. 4a). The second case study focused on applying our model to a real-world scenario, generating compounds that are likely to be sensitive to actual TNBC clinical patients (Fig. 4g, Methods). For each case study, we generated 2000 molecular latent vectors predicted to be sensitive (1000 for very sensitive and 1000 for sensitive conditions, respectively) at each point on the classifier-free guidance scale⁶³, which ranges from 0 to 11. This approach allows for fine-tuned control over the generated compounds, balancing novelty with biological relevance.

A key strength of G2D-Diff lies in its interpretability, achieved through the analysis of attention coefficients extracted from the transformer blocks in the condition encoder. This feature enables the identification of target-related pathways, providing valuable insights into the model's decision-making process. We rigorously validated the credibility of the attention mechanism through comprehensive quantitative analysis (Methods), extracting attention information for every response class and cell line in the cell line-centric dataset. We examined whether the attention highlights the genetic alterations or system-level mutation burden in the NeST (Nested Systems in Tumors) hierarchical ontology, which is the data-driven cancer-related ontology⁶⁴ (Supplementary Fig. 10a). Remarkably, our analysis revealed that G2D-Diff's attention mechanism goes beyond simply focusing on genetically altered genes. It also considers other genes belonging to the same systems as the genetically altered genes through system-level and gene-level attention analysis (Supplementary Fig. 10b). This sophisticated approach allows the model to capture complex biological interactions and dependencies. Furthermore, we observed that the model more likely focuses on genes that belong to the system with mutation burden in sensitive cases (Supplementary Fig. 10c). These results support the credibility of the attention outcomes and suggest that we can further identify critical genes or pathways that are not directly mutated.

In the first case study, we verified whether the generated compounds were consistent with known sensitive compounds to the given TNBC cell line, HS578T, based on the predicted AUC. The distribution of predicted AUC values for compounds generated under the sensitive condition of the HS578T genotype closely resembled those of known sensitive compounds, while significantly differing from the distribution of resistant compounds (Fig. 4b, two-tailed Mann–Whitney U test, $p < 10^{-4}$). Most generated compounds exhibited low maximum

Tanimoto similarity scores when compared to known sensitive compounds, confirming that the generated compounds are likely to be sensitive and exhibit diverse structures (Fig. 4c).

Next, we investigated the attention in the sensitive condition of the HS578T genotype. We extracted genes belonging to the top 10 percent of the attention result and having attention values greater than the uniform value for each attention head. We then conducted a gene set enrichment analysis for each gene set using the NeST ontology. As a result, the PI3K/AKT/PTEN signaling pathway and histone modification pathways, particularly the histone deacetylation pathway, are significantly enriched (Table 2). To determine whether compounds selectively sensitive to the HS578T cell line target the enriched pathways, we examined the compounds in the drug-centric dataset. We found four compounds that showed selective sensitivity to the HS578T cell line and have known molecular targets (Supplementary Table 7). Finally, we identified Fimepinostat (DrugBank ID: DB11891, PubChem CID: 54575456), an orally bioavailable dual inhibitor of PI3K and HDAC (histone deacetylase). Our search for a compound similar to Fimepinostat among the generated molecules yielded a promising candidate, here, temporally named as TNBC-S1 (Fig. 4d). Interestingly, despite the structural difference to Fimepinostat (Tanimoto similarity: 0.305), we observed a distribution of similar polar side-chains based on a central aromatic ring scaffold. Retrosynthesis analysis using SciFinder⁶⁵ revealed that TNBC-S1 could be synthesized at a reasonable cost (Supplementary Fig. 11). This confirms that the queried compound is both structurally distinct and plausible.

Subsequently, we conducted a docking simulation to determine whether TNBC-S1 also likely binds to PI3K α and HDAC1 (Methods). The docking results showed that TNBC-S1 binds strongly to the same docking sites as Fimepinostat on PI3K α and HDAC1 (Fig. 4e, f). We also investigated the interaction profile predicted by PLIP software⁶⁶, and found shared molecular interactions between Fimepinostat and TNBC-S1 (Supplementary Fig. 12).

To further validate our findings, we analyzed the remaining reference compounds. We found that molecular targets of these compounds were all represented in our enriched pathways (Supplementary Tables 7 and 8). Moreover, the generated compounds selected with maximum Tanimoto similarity to those ground truth compounds showed reasonable normalized docking score⁶⁷ distributions compared to the docking scores of reference ground truth compounds (Supplementary Figs. 13 and 14).

Beyond structural analysis, we employed the MoAble⁶⁸ tool to investigate the mechanism of action (MoA) similarity between TNBC-S1 and Fimepinostat, comparing them against a control group of 50 diverse non-cancer drugs. With pathway-level MoA predictions, we first identified enriched pathways (FDR <0.01) and assessed pathway similarities through multiple metrics (Kendall's tau, Jaccard, and Rank-biased overlap). TNBC-S1's predicted MoA was more similar to Fimepinostat than to the control drugs (Supplementary Fig. 15a). This analysis provides additional evidence that TNBC-S1 may share similar biological mechanisms with Fimepinostat.

In our second case study, we employed the identical validation process as in the first case study. We revisited the accuracy of the compounds generated by comparing the predicted AUC values to those of known sensitive compounds. Given the lack of ground truth for clinical cases, we collected sensitive data from all TNBC cell lines (BT549, MDAMB231, MDAMB468, and HS578T) in the drug-centric dataset as known sensitive compounds. The predicted AUC distribution of compounds generated for clinical genotypes was closely aligned with those of known sensitive compounds from all TNBC cell lines. It showed a significant difference from the resistant compounds (Fig. 4h, two-tailed Mann–Whitney U test, $p < 10^{-4}$). Again, most generated molecules exhibited low maximum Tanimoto similarity scores to known sensitive compounds, suggesting that the generated compounds are both credible and unique (Fig. 4i).

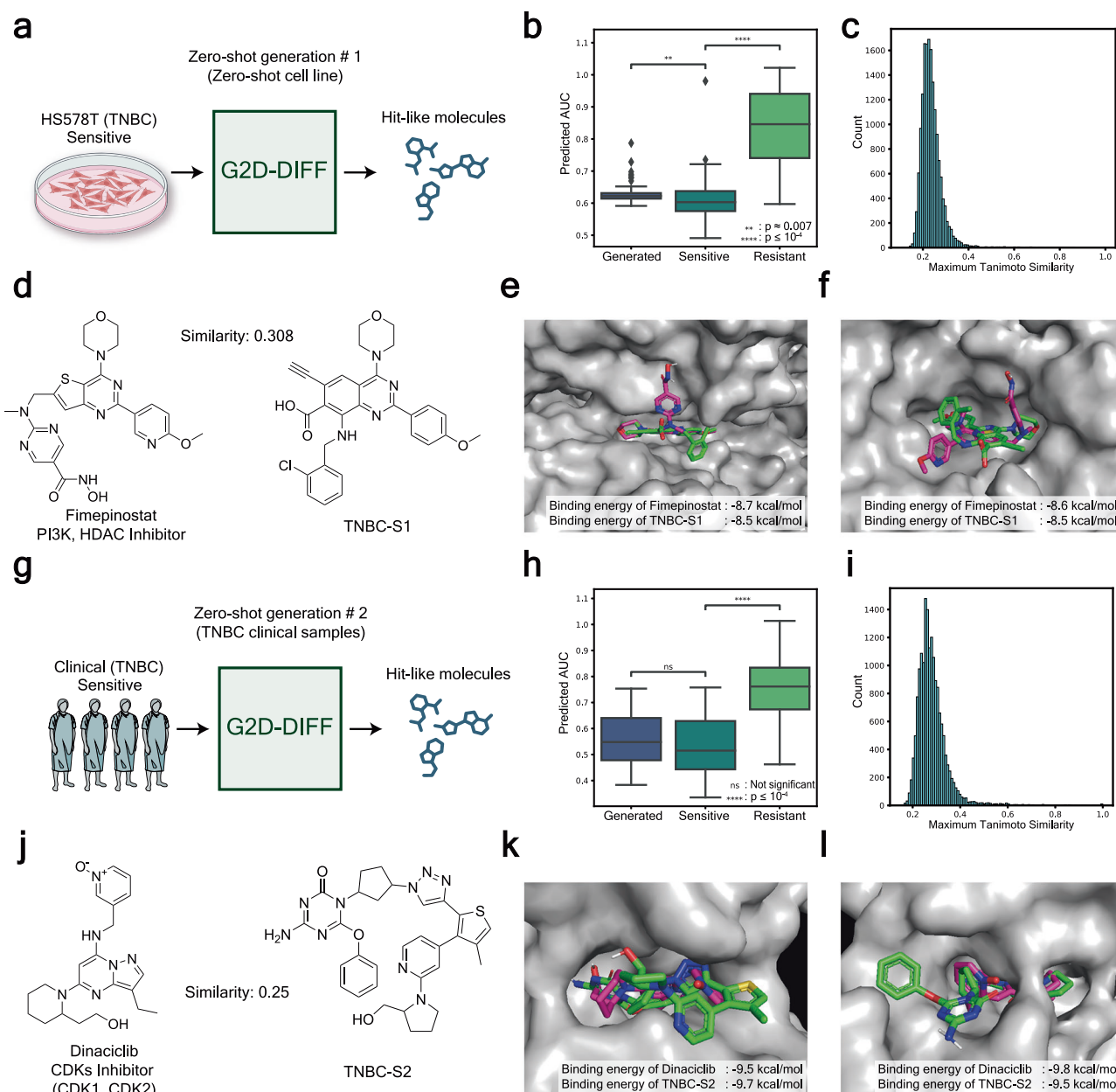


Fig. 4 | Workflow and the results of zero-shot case studies. **a** Workflow of the first zero-shot case study with the unseen TNBC cell line, HS578T. **b** Distribution of predicted AUC of generated compounds, ground truth sensitive, and resistant compounds. Compounds were randomly sampled for each case ($n = 50$). Two-tailed Mann–Whitney U test was conducted. **c** Histogram of maximum Tanimoto similarity between generated compounds and ground truth sensitive compounds. **d** Chemical structures of Fimepinostat and a generated compound named TNBC-S1. **e** Docking simulation result of Fimepinostat (purple) and TNBC-S1 (green) in PI3Ka. **f** Docking simulation result of Fimepinostat (purple) and TNBC-S1 (green) in HDAC1. **g** Workflow of the second zero-shot case study with unseen clinical patients from project GENIE. **h** Distribution of predicted AUC of generated compounds,

ground truth sensitive, and resistant compounds (randomly sampled $n = 50$ for each case). Two-tailed Mann–Whitney U test was conducted. **i** Histogram of maximum Tanimoto similarity between generated and ground truth sensitive compounds. **j** Chemical structures of Dinaciclib and a generated compound named TNBC-S2. **k** Docking simulation result of Dinaciclib (purple) and TNBC-S2 (green) in CDK1. **l** Docking simulation result of Dinaciclib (purple) and TNBC-S2 (green) in CDK2. Box plots show median (center line), interquartile range (25th and 75th percentiles, box limits), whiskers (e.g., 1.5 times the interquartile range from box limits), and outliers (points outside whiskers). Source data are provided as a Source Data file. Created in BioRender. Nam, H. (2025) <https://BioRender.com/sw3mchi>.

As we analyzed the attention, we discovered that CDK-related pathways were significantly enriched in the clinical genotype (Table 2). Consequently, we identified a cancer drug, Dinaciclib (DrugBank ID: DB12021, PubChem CID: 46926350), targeting CDKs (CDK1, CDK2, CDK5, and CDK9) that was sensitive in three out of four TNBC cell lines in the drug-centric dataset. We then searched for the most similar compounds to Dinaciclib among generated compounds and found a suitable match, TNBC-S2 (Fig. 4j). As similar to the previous case study,

TNBC-S2 had a low chemical structure similarity compared to Dinaciclib (Tanimoto similarity: 0.25). Furthermore, as with TNBC-S1, SciFinder retrosynthesis analysis confirmed the feasibility of TNBC-S2 synthesis at a reasonable cost (Supplementary Fig. 16).

To investigate if TNBC-S2 binds to the same target as Dinaciclib, we performed docking simulations on CDK1 and CDK2. Surprisingly, this compound docked to similar sites to Dinaciclib on CDK1 and CDK2 with comparable binding energies (Fig. 4k, l). When investigating the

Table 2 | Enriched pathways of highlighted genes on NeST gene sets for zero-shot case studies

TNBC zero-shot case study	Term	Overlap	P value	Odd ratio
HSS78T cell line	PIK3/AKT/PTEN signaling	4/11	0.018	5.601
	Histone deacetylation	5/16	0.016	4.503
	Histone modification	8/30	0.007	3.658
GENIE clinical	CDK holoenzyme complex I	5/15	0.012	4.939
	CDK holoenzyme complex II	5/17	0.021	4.136
	Regulation of CDK activity I	11/41	0.001	3.780
	Regulation of CDK activity II	8/19	0.0002	7.282

Statistical significance was determined without adjustment for multiple comparisons to explore the biological relevance of the highlighted genes.

interaction profile between Dinaciclib and TNBC-S2, we observed that specific molecular interactions were identical (Supplementary Fig. 17). Similarly, the MoA analysis also revealed that TNBC-S2's predicted MoA was more similar to Dinaciclib than to the control drugs (Supplementary Fig. 15b), providing additional evidence for their mechanistic similarity. In summary, despite the challenges of the two extreme zero-shot scenarios, the model showed remarkably reliable generations along with high interpretability, enabling us to confirm its generalizability and practicality.

Discussion

This research introduces the genotype-to-drug generative model, G2D-Diff, a diffusion-based generative approach designed to create molecules that meet specific conditions, including any cancer genotype and response class. Our findings confirm G2D-Diff's excellent generative performance across various experimental settings, emphasizing its ability to generate realistic and diverse molecules aligned with expected responses. In particular, the model demonstrated robust generative capabilities in both seen and unseen conditions, efficiently utilizing the genotype of clinically relevant genes and pre-training techniques to enhance generalizability.

G2D-Diff showed superior generation quality compared to the competitive method, exhibiting high diversity, feasibility, and condition fitness. Although our model showed relatively lower validity and uniqueness, this was considered manageable due to the trade-offs associated with the level of classifier-free guidance. This guidance allows for adjustments of generated objects during the denoising process, balancing diversity and condition fitness⁶³. Note that the reported results above were obtained at a selected point on the classifier-free guidance scale (CFG scale = 7), where the average predicted AUC of the generated compounds matched that of the ground truth (Supplementary Table 9). Additionally, we have verified that our model can generate diverse and drug-like hit candidates by directly learning the molecular distribution, thereby also enhancing reliability by avoiding bias in predictions.

The zero-shot generative case studies, especially for TNBC, highlight the model's practicality in real-world scenarios where data are scarce, and the model encounters unseen cases. G2D-Diff successfully generated hit-like candidates for these zero-shot cases by fully leveraging the input genotype condition through an attention mechanism, thus enhancing interpretability. In the first case study, G2D-Diff proved that it could generate plausible hit candidates TNBC-S1 by focusing on PI3K and HDAC-related pathways, which are the target proteins of the reference drug. And in the second case study, G2D-Diff showed great adaptability to the clinical TNBC genotype, by generating a great candidate TNBC-S2 by focusing on CDK-related pathways. This systematic validation through ground truth compounds, attention results, and generated compounds demonstrates G2D-Diff's potential for rational drug discovery.

While G2D-Diff has already shown promising results, its performance could be enhanced further by utilizing not only large-scale in vitro drug response data but also in vivo and clinical drug response

datasets. Since we used only the genotype of clinically relevant genes and manually categorized the in vitro drug response into five response classes, our model architecture demonstrates inherent compatibility with various types of drug response data. Beyond data integration, since the model is a latent diffusion model, the model can be further improved with a more precise molecular latent space. In this proof-of-concept study, we developed and utilized a relatively simple and memory-efficient chemical VAE for building molecular latent space, primarily based on SMILES representations. Although this approach showed reasonable performance, it has inherent limitation that the molecular latent space built by our chemical VAE contains ill-defined regions, which subsequently gives invalid chemical structures. Future work could explore the integration of emerging graph-based VAEs and 3D molecular information frameworks that have shown superior performance in capturing molecular structures^{69–71}. These advanced representation approaches, combined with more computational resources and advanced representation learning techniques utilizing large-scale datasets, could significantly enhance G2D-Diff's capability in generating hit-like candidates.

In conclusion, the G2D-Diff showed its ability to generate realistic, hit-like molecular structures under specific genotypic conditions, not only improving our ability to conquer cancer but also pioneering avenues for personalized medicine. By suggesting the drug-like hit candidates and related genes or pathways, particularly for difficult-to-treat cancer types, the drug discovery timeline can be significantly shortened by reducing the trials and errors in target validation and hit identification. This study confidently positions G2D-Diff as an adequate starting point for actual drug discovery, laying the foundation for rapid and targeted therapeutic development.

Methods

Datasets for model training

In this study, we utilized two different types of datasets, each contributing to our G2D-Diff: chemical structure dataset and drug response dataset. The first dataset, which focuses on chemical structure, was used to train the chemical VAE. We collected SMILES strings from GuacaMol's distribution-learning benchmark⁷², which consists of various molecules from the ChEMBL database. To preprocess the collected SMILES strings, we removed entries having characters that are not specified by a set of tokens used in this study. The utilized SMILES tokens include H, C, N, O, F, P, S, Cl, Br, I, c, n, o, p, s, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, (,), [,], -, =, +, #, %, <BEG>, <EOS>, and <PAD>. <BEG> indicates the beginning of the SMILES sequence, <EOS> indicates the end of the sequence, and <PAD> indicates the padding positions coming after <EOS>. Additionally, we removed SMILES strings with sequence length exceeding 126, the maximum length allowed by the model. Then, the SMILES strings were canonicalized using RDKit⁷³, and duplicate strings were removed. The resulting dataset consisted of 1,583,442 unique SMILES strings. We randomly split this dataset using a 9:1 ratio to create training and validation sets. The final training set comprises 1,425,097 molecules, while the validation set contains 158,345 molecules.

The second dataset comprises cell line and drug response data, which was used to train the genotype-to-drug latent diffusion generative model. Initially, we retrieved the drug response data from various databases, such as GDSC, CTRP, and NCI60. Considering the characteristics of the data, GDSC and CTRP have a higher ratio of cell lines to compounds, indicating that these datasets have a larger variety of cell lines. In contrast, NCI60 features a higher ratio of compounds to cell lines, suggesting that this dataset has a larger variety of compounds. Consequently, we combined GDSC and CTRP data to form a cell line-centric dataset and designated NCI60 as a drug-centric dataset. For drug response values, we used the AUC as our metric, where an AUC of 0 signifies total cell elimination, an AUC of 1 indicates no effect on cell viability, and an AUC greater than 1 implies that the drug promotes cell growth. We defined the genotypes of cell lines using a binary format to represent point mutations, copy number amplifications, and deletions across 718 clinical genes, which were used in the Park et al.⁷⁴. To calculate AUC and get the genotype of cell lines, we strictly followed the protocols established in previous drug response prediction studies^{33,74}. For compounds, we canonicalized SMILES strings before inputting them into the chemical VAE encoder to generate numerical representation vectors. To enhance the quality of the dataset, we first excluded compounds that could not be encoded or reconstructed by the chemical VAE. Additionally, we removed compounds that exhibited similar efficacy in more than half of the cell lines in the drug response datasets, as these compounds might introduce detrimental effects, preventing the model from effectively utilizing genotype information. As a result, the cell line-centric drug response dataset includes 1244 cell lines, 803 compounds, and 432,293 cell line-compound pairs. In contrast, the drug-centric drug response dataset contains 62 cell lines, 38,502 compounds, and 811,585 cell line-compound pairs (Supplementary Table 10). We divided the AUC into five discrete response categories: very sensitive, sensitive, moderate, resistant, and very resistant. The AUC category is used to create condition-compound pairs, where each condition combines a cell line with a response class. We used the finalized cell line-centric drug response dataset to pre-train the condition encoder of the G2D-Diff and the drug-centric drug response dataset to train the conditional diffusion generative model.

To evaluate the models, we created three evaluation sets, each containing five cell lines (Supplementary Table 11). Evaluation set 1 contains five data-rich cell lines that have been fully exposed to both the condition encoder and the generative model during training to assess the trustworthiness of the learning process. Evaluation set 2 includes data-scarce five cell lines for which conditions were known and trained, but the generative model was not exposed to the known compounds. This set tests the model's ability to generate condition-meeting compounds under known conditions without any prior information about real compound pairs. Evaluation set 3 also consists of five data-scarce cell lines, but it presents the most challenging collection of cell lines because they were completely unseen during the entire training phase, considered as a zero-shot set.

Model architecture of G2D-Diff

The proposed G2D-Diff model comprises three main components: the chemical VAE, the condition encoder, and the latent diffusion model.

Chemical VAE. First, for the chemical VAE, we enhanced a recurrent neural network-based SMILES VAE⁷⁵, specifically employing long short-term memory (LSTM) networks. This model features an encoder and a decoder of three layers of LSTMs with a hidden dimension of 256. The encoder part takes in SMILES tokens and processes them through LSTM layers, and the final hidden states go through the two fully connected layers with dimension 128, outputting the mean and logarithmic variance of the VAE latent representation, respectively. Subsequently, the decoder takes both the 128-dimensional latent

representation and a SMILES token as input for each time step for the LSTM hidden state update, and the output passes through a fully connected layer with Softmax activation to predict the probabilities for each next token. In practice, the decoder generates the next token using a greedy approach.

Condition encoder. The second component of the G2D-Diff model is the condition encoder, which encodes the genotype and response class into a numerical vector (Supplementary Fig. 18). The encoding process involves two main modules: genetic alteration embedding and transformer-based encoding (Supplementary Fig. 18a). The genetic alteration embedding module begins by generating 128-dimensional embeddings that represent the wild-type genotype for all genes. These embeddings are randomly initialized and optimized during training. When a gene exhibits a genetic alteration, the **gene2sig** module generates the gene-specific genetic alteration signal embedding. For example, let's assume Gene 1 is experiencing a mutation. The **gene2sig** module generates a mutation signal for Gene 1 ($\mathbf{s}_{mut,1}$) through a linear transformation of a randomly initialized trainable mutation basis embedding (\mathbf{basis}_{mut}), employing gene-specific weights and biases. These parameters are derived from a fully connected layer named **gene2wb**, which resides within the **gene2sig** module and takes the initial embedding of Gene 1 (\mathbf{g}_1) as input. After then, the mutation signal embedding for Gene 1 is combined with the initial gene embedding to create the final altered gene embedding corrupted by mutation ($\mathbf{g}_{mut,1}$). The following formulas show the details of the genetic alteration embedding module when Gene 1 exhibits mutation:

$$\mathbf{s}_{mut,1} = \mathbf{gene2sig}(\mathbf{basis}_{mut}, \mathbf{g}_1) \quad (1)$$

$$\mathbf{gene2sig}(\mathbf{basis}_{mut}, \mathbf{g}_1) = \mathbf{w}_{\mathbf{g}_1} \circ \mathbf{basis}_{mut} + \mathbf{b}_{\mathbf{g}_1} \quad (2)$$

$$\mathbf{w}_{\mathbf{g}_1}, \mathbf{b}_{\mathbf{g}_1} = \mathbf{gene2wb}(\mathbf{g}_1) \quad (3)$$

$$\mathbf{g}_{mut,1} = \mathbf{g}_1 + \mathbf{s}_{mut,1} \quad (4)$$

where \circ is the element-wise multiplication. When dealing with multiple genetic alterations (both copy number alteration and mutation) for Gene i , all alteration signal embeddings are generated and added together to the wild-type embedding of Gene i . The following equation presents the generalized formula of the genetic alteration embedding module:

$$\mathbf{g}_{alt,i} = \mathbf{g}_i + \sum_{alt \in \left\{ \begin{matrix} mut, \\ cna, \\ cnd \end{matrix} \right\}} \mathbf{gene2sig}(\mathbf{basis}_{alt}, \mathbf{g}_i) \quad (5)$$

where the \mathbf{g}_i is the wild-type embedding of the Gene i , \mathbf{basis}_{alt} is the basis embeddings of the arbitrary genetic alterations (alt) that can be either mutation (mut), copy number amplification (cna) or copy number deletion (cnd) (Supplementary Fig. 18b). Subsequently, an embedding representing the response class, analogous to a class (CLS) token in large language models, is combined with the final gene embeddings, integrating all genetic alteration signals. The complete set of embeddings is fed into the three-layered transformer encoder⁷⁶ for information exchange within genes. Notably, the first layer's attention matrix is masked to enable information exchange only among genes that belong to the same subsystem in the NeSt hierarchy⁶⁴, modeling the gradual propagation of biological effects of mutations through interacting genes. After processing through a three-layered transformer encoder, an information-integrated response class embedding vector is extracted, which serves as the final condition encoding.

Latent diffusion model. Once the condition encoding is generated, the model inputs this encoding into a diffusion generative model. Here, we refined the noise predictor ϵ_θ in the denoising diffusion probabilistic model (DDPM)⁵⁰ to generate the molecular latent vector that meets the given condition. In detail, the denoising network $\epsilon_\theta(\mathbf{x}_t, \mathbf{c}, \mathbf{t})$ takes the three inputs: the denoised input at time t (\mathbf{x}_t), condition (\mathbf{c}), and time step (\mathbf{t}) embeddings. Each embedding is transformed using its respective multi-layered perceptrons (MLPs) with Gaussian error linear unit (GELU) activation. Then all the transformed embeddings are fed into the condition injection module. This injection module, inspired by adaptive instance normalization⁷⁷, is applied across L layers (here, six layers) in the denoising network. The process in each layer is as follows:

$$\mathbf{x}_t^{l+1} = \mathbf{f}_2^l \left(\text{GELU} \left(\mathbf{w}_c^l \circ \text{norm} \left(\mathbf{f}_1^l \left(\mathbf{x}_t^l \right) \right) + \mathbf{b}_c^l \right) \right) \quad (6)$$

$$\mathbf{w}_c^l, \mathbf{b}_c^l = \text{cond2wb}^l(\mathbf{c}, \mathbf{t}) \quad (7)$$

where \mathbf{x}_t^l is input for the current layer l , \mathbf{f}_1^l is a fully connected layer, **norm** is an instance normalization layer, **cond2wb**^l, and \mathbf{f}_2^l are fully connected layers with GELU activations. After the condition injection process is completed through L layers, the predicted noise ϵ_θ is finally generated by passing through the \mathbf{x}_t^l to the MLPs with GELU activation. With this revised conditional denoising network, the denoising process is conducted as proposed in DDPM during the reverse diffusion process. To control the conditional guidance for generating realistic and condition-meeting molecules, we employed the classifier-free guidance technique⁶³ in the model.

Training details of G2D-Diff

Our proposed G2D-Diff undergoes a three-stage training process to generate hit-like molecules from inputs of genotype and the desired response class. The first stage involves training of the chemical VAE, followed by the pre-training of the condition encoder in the second stage. Lastly, the final stage involves training the latent diffusion model.

To begin, the chemical VAE was trained by minimizing both the reconstruction errors of the SMILES strings and the Kullback–Leibler divergence (KLD) between the distribution formed by the encoder outputs and the standard normal distribution⁷⁸. We adopted the training scheme from beta-VAE⁷⁹ as follows:

$$L(\phi, \beta) = -\mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] + \beta D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) || p_\theta(\mathbf{z})) \quad (8)$$

where θ and ϕ are the parameters for the decoder and encoder networks, respectively. β was used to adjust the relative importance of KLD in the training. Since \mathbf{z} vectors are stochastically sampled in the objective, the reparameterization trick⁷⁸ enabled the back-propagation. We trained the VAE for 250 epochs using the Adam optimizer with a learning rate of 0.0003 and a batch size of 512. To avoid the notorious KLD vanishing problem, we adopted the cyclical annealing schedule for the beta parameter⁸⁰. Initially, β was set to 0 for the first 15 epochs for the model to learn to reconstruct the correct sequences. Then, the cyclical annealing schedule begins, where each cycle consists of 10 epochs. For the first 3 epochs of a cycle, we set $\beta = 2 \times 10^{-8}$. In the later 7 epochs of the cycle, we linearly increased the value to the adjustable maximum, which was initially set to $\beta = 5 \times 10^{-4}$. The maximum β was adjusted manually whenever the KLD value reached a plateau during training.

Secondly, the condition encoder was pre-trained using the cell line-centric drug response dataset via contrastive learning. Initially, we extracted condition encodings and compound latent representations from the condition encoder and the chemical VAE encoder, respectively. These were then normalized and projected into a shared vector

space using trainable MLPs that contain layer normalization, GELU activation, and dropout layers. The training process involved minimizing the refined version of contrastive loss L utilized in CLIP⁵¹, which is as follows:

$$L = (1 - \lambda)L_c + \lambda L_d \quad (9)$$

$$L_c = -\frac{1}{M} \sum_{i=1}^M \log \left(\frac{\exp(\mathbf{c}_i \cdot \mathbf{d}_i / \tau)}{\sum_{k=1}^M \exp(\mathbf{c}_i \cdot \mathbf{d}_k / \tau)} \right) \quad (10)$$

$$L_d = -\frac{1}{M} \sum_{i=1}^M \log \left(\frac{\exp(\mathbf{d}_i \cdot \mathbf{c}_i / \tau)}{\sum_{k=1}^M \exp(\mathbf{d}_i \cdot \mathbf{c}_k / \tau)} \right) \quad (11)$$

where λ is a loss weight, L_c represents the contrastive loss from the perspective of the condition, while L_d represents the same type of loss from the perspective of the drug. M is the batch size, \mathbf{c}_i and \mathbf{d}_i are the i th condition and drug pair in the batch, and τ is the temperature parameter. \mathbf{c}_k , \mathbf{d}_k represent condition and drug data points, respectively, in the batch other than the i th anchor pair. This loss function is designed to reduce the distance between matched condition-drug pairs (i, i) while increasing the distance between unmatched pairs (i, k) from both condition and drug perspectives in the shared vector space. We trained all the trainable parameters with a λ value of 0.1, a batch size of 128, a τ value of 0.3, a dropout rate of 0.2, and for 70 epochs, where a single epoch contains 2000 steps. Inspired by DeepMind's work⁸¹, we sampled a small set of unmatched samples (-10) for each matched pair. We used the Adam optimizer with a learning rate of 5×10^{-5} for the parameter optimization. We also applied a learning rate warm-up over 4000 steps. We also used the weighted random sampling by the number of data points for each response class to tackle the data imbalance. We selected the best epoch by evaluating pair prediction performance using the evaluation set 3 mentioned above.

Lastly, after fully pre-training the condition encoder, we froze it and trained the diffusion model using the drug-centric drug response dataset. We adopted the loss function described in the DDPM paper⁵⁰:

$$L = \mathbb{E}_{t, \mathbf{x}_0, \mathbf{c}, \epsilon} [\|\epsilon - \epsilon_\theta(\mathbf{x}_t, \mathbf{c}, \mathbf{t})\|^2] \quad (12)$$

where ϵ is the random initial noise that follows the normal distribution, ϵ_θ is the noise predictor, and \mathbf{x}_0 is the original data (here, the molecular latent vector). The training protocol included 300 diffusion time steps, a batch size of 512, and utilized the cosine scheduling for the variance of the forward diffusion process. For classifier-free guidance⁶³ learning, 10% of each batch was trained unconditionally. Adam optimizer with a learning rate of 10^{-4} was used to optimize the parameters. The model was trained for 3000 epochs in total, and we selected the best epoch by evaluating the generation performance of the evaluation set 2. Additionally, similar to previous stages, we addressed data imbalance by implementing weighted random sampling based on the number of data points for each response class. Our model was implemented using PyTorch⁸² and HuggingFace⁸³ frameworks, and was trained utilizing four NVIDIA A100 GPUs.

Evaluation metrics to measure the basic generation quality

We employed six evaluation metrics to assess the basic generation quality of the generated compounds, including validity, uniqueness, novelty, diversity, FCD⁵⁷, and the OTD¹⁵. Validity refers to the proportion of generated SMILES strings that are chemically valid. Uniqueness denotes the proportion of non-duplicated compounds among valid compounds. Novelty is defined as the ratio of unique compounds that are different from those in the training set. Diversity is measured by the average pairwise distance among the generated compounds. In detail,

let N as the total number of generated samples, V as the set of valid generations among N , U as the set of unique molecules among V , Z as the set of molecules not present in the pre-training dataset among U , and T as the set of target active molecules in the test or validation set. We define:

$$\text{Validity} = |V|/N \quad (13)$$

$$\text{Uniqueness} = |U|/|V| \quad (14)$$

$$\text{Novelty} = |Z|/|U| \quad (15)$$

$$\text{Diversity} = \frac{1}{|V_{1k}|^2} \sum_{m_1 \in V_{1k}, m_2 \in V_{1k}} 1 - \text{sim}(m_1, m_2) \quad (16)$$

where the molecular similarity is calculated using $\text{sim}(m_1, m_2)$, which represents the Tanimoto similarity of Morgan fingerprints between molecules m_1 and m_2 . Due to computational constraints, we utilize V_{1k} , a random subset of 1000 molecules from V , for pairwise similarity calculations. The FCD, which has become a *de facto* standard in molecular generation evaluation, quantifies the distance between the distribution of generated compounds and real compounds in the ChemNet⁸⁴ feature space. The FCD is calculated as:

$$\text{FCD} = \|\mu_V - \mu_T\|^2 + \text{Tr}(\mathbf{C}_V + \mathbf{C}_T - 2(\mathbf{C}_V \mathbf{C}_T)^{1/2}) \quad (17)$$

where μ_V and μ_T are the means of the feature vectors generated by the ChemNet model, using V and T respectively, while \mathbf{C}_V and \mathbf{C}_T are the corresponding covariances of these vectors. Lastly, to address potential bias in FCD due to its reliance on the SMILES-based ChemNet model, we additionally incorporate the OTD metric proposed by Bae et al.¹⁵. The OTD measures the minimal transportation distance between generated and reference molecular distributions based on molecular fingerprints-based similarities. Based on discrete optimal transport principles⁸⁵, the optimal transport ω is determined by solving:

$$\argmin_{\omega \in R} \sum_{x_i \in V, y_j \in T} T_{ij} \text{dist}(x_i, y_j) \text{ s.t. } T_{ij} \in \left\{0, \frac{1}{c}\right\}, \sum_{x_i \in V} T_{ij} = \frac{1}{c}, \sum_{y_j \in T} T_{ij} = \frac{1}{c} \quad (18)$$

where T_{ij} represents the transport mass from point i to point j , and R denotes the set of all possible one-to-one mappings from set V to set T . The distance metric used for OTD calculation is defined as:

$$\text{dist}(x_i, y_j) = 10^{1 - \text{sim}(x_i, y_j)} - 1 \quad (19)$$

We divided those basic generation quality metrics between chemical-likeness metrics (validity, uniqueness, novelty, diversity) and objective-related distance metrics (FCD, OTD). All calculations for these metrics were performed using the LOGICS package developed by Bae et al.

Genotype-based drug response predictor (G2D-Pred) for general evaluation

To validate the molecules generated from our G2D-Diff and ensure they meet the input conditions, we developed a genotype-based drug response predictor. This predictor takes as inputs the genotype and a latent representation of the compound to predict drug responses. We revised the condition encoder architecture to generate the encoding of the genotype. In detail, we replaced the response class embeddings with a randomly initialized general CLS token embedding, transforming it into a general genotype encoder rather than a response class-specific one. Importantly, the response class category is completely

excluded from G2D-Pred's inputs, preventing any potential data leakage. The genotype encoding is processed through the MLPs and then concatenated with the final compound encoding generated from the initial compound latent representation using the other MLPs. The final concatenated vector is fed into the predictor MLPs to predict the AUC value of the compound-genotype pair.

Since the generative model was trained on drug-centric data, the prediction model was also trained using the same dataset. The MLPs are composed of GELU activation and Dropout, with a dropout rate of 0.1. Parameter optimization was conducted using the Adam optimizer, with a learning rate warm-up for one epoch, followed by exponential decay⁷⁶; the initial learning rate was set at 10^{-6} and increased to 10^{-4} after warm-up steps. To address data imbalance, weighted random sampling based on the absolute z-scores of the AUC values was implemented, and the total training spanned 200 epochs. The optimal epoch was determined using a validation set comprising ~8000 pairs. Subsequently, the performance was evaluated on a distinct test set containing about 8000 pairs. We generated the five models trained with different random seeds and then used them as an ensemble predictor for the final prediction.

Evaluation settings for the fair performance comparison to the comparative method

Among the cancer baseline condition-based approaches^{46–48} mentioned in the Introduction section, PaccMannRL⁴⁸ represents the only benchmarkable generative model. Therefore, in this study, we use PaccMannRL as our sole baseline model for performance comparison. For the fair comparison, we re-trained the key components of PaccMannRL, namely the PaccMann predictor and the PaccMann generator, using the same dataset we used in the training generative model. For the training of the PaccMann predictor, we directly utilized the gene expression data of cell lines, which had been pre-processed and used in the original PaccMannRL study, excluding a small number of cell lines that lacked gene expression data. Furthermore, for the consistent experimental settings, we excluded the cell lines in evaluation set 3 from the PaccMann predictor training set. Also, we removed the cell lines in evaluation set 2 from the PaccMann generator training set. All hyperparameters were employed as mentioned in the original publication. The performance of the PaccMann predictor, which is crucial for the training of the PaccMann generator by RL, was validated on a subset of the total data, ~7000 unseen pairs. The validation results showed a Pearson correlation of 0.730 and an RMSE of 0.203, indicating reasonable predictive performance for RL training. For the PaccMann generator, it was trained only for one epoch since it generated critical mode-collapsed results after one epoch of training. After training the PaccMann generator using RL, we generated 2000 SMILES strings predicted as sensitive (AUC < 0.6) for each cell line in all evaluation sets. For the G2D-Diff model, we combined the generations generated under very sensitive and sensitive conditions across all evaluation sets, a total of 2000 generations for each cell line.

Triple-negative breast cancer (TNBC) clinical data processing

To generate compounds using the genetic alteration information from actual TNBC patients, we parsed AACR project GENIE⁸⁶ data and generated a representative clinical genotype that is compatible with the input of G2D-Diff. In detail, we analyzed the “Metastatic Breast Cancer: 2013-2016 (DFCI, CCR 2020)” cohort from cBioPortal⁸⁷. We selected patients with lethal TNBC based on the following clinical information criteria: (1) patients with primary TNBC; (2) patients initially diagnosed as Stage 1 or 2; and (3) patients whose survival time from initial diagnosis was within 2 years. From this selection, four patients were identified. For these patients, we unionized the mutation, copy number amplification, and copy number deletion information across 718 genes used as input for the model to create a representative clinical genotype. Upon unionization, the total mutation count was 22, the

copy number amplification count was 75, and the copy number deletion count was 12.

Docking simulation of the generated anticancer compound for TNBC genotypes

For the docking simulation of selected compounds that are supposed to be hit candidates against TNBC, we parsed the protein 3D structures of four key proteins: PI3Ka (PDB ID: 4L2Y), HDAC1 (PDB ID: 4BKX), CDK1 (PDB ID: 6GU6), and CDK2 (PDB ID: 4KDI) from PDB database. Docking simulations were conducted using QuickVina 2.0⁸⁸, with docking sites for each protein structure determined either from literature sources (4BKX⁸⁹: $x = -62.66$, $y = 17.19$, $z = -4.97$) or based on compounds interacting in the PDB database (4L2Y: $x = -31.45$, $y = 46.01$, $z = -41.29$, 4KDI: $x = 55.56$, $y = 79.86$, $z = 28.02$, 6GU6: $x = 23.64$, $y = 21.82$, $z = -1.90$). Additionally, the size of the search box was fixed to $30 \times 30 \times 30$, and search exhaustiveness was fixed to 300 for all cases. After the docking process, visualizations were performed using PyMOL software⁹⁰, and the interaction profiles were predicted and analyzed using PLIP software⁶⁶.

Retrosynthesis prediction

To assess the feasibility of synthesizing the designed molecules in a laboratory setting, we analyzed their retrosynthetic pathways using AiZynthFinder software⁹¹. The software configurations were adjusted slightly to align with the objectives of this study. In addition to the default dataset of 17,422,831 ZINC⁹² compounds, we incorporated 1,371,304 compounds from the Enamine Building Blocks catalog (accessed in November 2024)⁹³, expanding the stock library used by AiZynthFinder's search algorithm. The Monte Carlo Tree Search algorithm employed reaction templates from USPTO⁹⁴ for the expansion policy, with a search time limit of 120 s per synthesis tree and a maximum tree depth of eight. If multiple synthetic routes were identified for a single compound, the first route discovered was selected as the representative solution. All other settings remained as per the default configuration of AiZynthFinder version 4.3.1⁹⁵.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The processed datasets used in this study are deposited in a GitHub repository (<https://github.com/GIST-CSBL/G2D-Diff>⁶⁶). The raw datasets are all publicly available. GuacaMol benchmark dataset [<https://github.com/BenevolentAI/guacamol>] ChEMBL (used via GuacaMol, release 24) [<https://www.ebi.ac.uk/chembl/>]. GDSC1 [https://www.cancerrxgene.org/downloads/bulk_download]. GDSC2 [https://www.cancerrxgene.org/downloads/bulk_download]. CTRPv1 [<https://portals.broadinstitute.org/ctrp.v1/>]. CTRPv2 [<https://portals.broadinstitute.org/ctrp.v2.1/>]. NCI60 Growth Inhibition Data (accessed July 2022) [<https://wiki.nci.nih.gov/display/NCIDTPdata/NCI-60+Growth+Inhibition+Data>]. DepMap (22Q1) [<https://doi.org/10.6084/m9.figshare.19139906.v1>]. Project GENIE (mbc_genie_2020) [https://genie.cbioportal.org/study/summary?id=mbc_genie_2020]. 4L2Y. 4BKX. 4KDI. 6GU6. Enamine Building Blocks catalog (accessed November 2024) [<https://enamine.net/building-blocks/building-blocks-catalog>] Source data are provided with this paper.

Code availability

The Python-based source code of G2D-Diff and trained model parameters are available on our GitHub repository (<https://github.com/GIST-CSBL/G2D-Diff>⁶⁶) under the PolyForm Noncommercial License 1.0.0. Portions of the source code were adapted from Phil Wang's repository (<https://github.com/lucidrains/denoising-diffusion-pytorch>), which is licensed under the MIT License. Additionally, all

trained model parameters and data generated using our code are released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0).

References

- Bandi, A., Adapa, P. V. S. R. & Kuchi, Y. E. V. P. K. The power of Generative AI: a review of requirements, models, input-output formats, evaluation metrics, and challenges. *Future Internet* **15**, 260 (2023).
- Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
- Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. Preprint at https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (2018).
- Radford, A. Language models are unsupervised multitask learners. OpenAI Blog. Vol. 1, 9 (2019).
- Betker, J. et al. Improving image generation with better captions. Technical Report (OpenAI, 2023). Available at: <https://cdn.openai.com/papers/dall-e-3.pdf>.
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical text-conditional image generation with CLIP latents. Preprint at <https://arxiv.org/abs/2204.06125> (2022).
- Ramesh, A. et al. *Proc. International Conference on Machine Learning* 8821–8831 (PMLR, 2021).
- Achiam, J. et al. GPT-4 technical report. Preprint at <https://arxiv.org/abs/2303.08774> (2023).
- Gupta, A. et al. Generative recurrent networks for de novo drug design. *Mol. Inform.* **37**, 1700111 (2018).
- Renz, P., Van Rompaey, D., Wegner, J. K., Hochreiter, S. & Klambauer, G. On failure modes in molecule generation and optimization. *Drug Discov. Today Technol.* **32**, 55–63 (2019).
- Segler, M. H., Kogej, T., Tyrchan, C. & Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **4**, 120–131 (2018).
- Song, T. et al. DNMG: Deep molecular generative model by fusion of 3D information for de novo drug design. *Methods* **211**, 10–22 (2023).
- Wang, Y., Zhao, H., Sciabola, S. & Wang, W. cMolGPT: A conditional generative pre-trained transformer for target-specific de novo molecular generation. *Molecules* **28**, 4430 (2023).
- Atance, S. R., Diez, J. V., Engkvist, O., Olsson, S. & Mercado, R. De novo drug design using reinforcement learning with graph-based deep generative models. *J. Chem. Inf. Model.* **62**, 4863–4872 (2022).
- Bae, B., Bae, H. & Nam, H. LOGICS: Learning optimal generative distribution for designing de novo chemical structures. *J. Cheminform.* **15**, 77 (2023).
- Munson, B. P. et al. De novo generation of multi-target compounds using deep generative chemistry. *Nat. Commun.* **15**, 3636 (2024).
- Goel, M., Raghunathan, S., Laghuvarapu, S. & Priyakumar, U. D. Molegular: Molecule generation using reinforcement learning with alternating rewards. *J. Chem. Inf. Model.* **61**, 5815–5826 (2021).
- Guo, J. & Schwaller, P. Augmented memory: sample-efficient generative molecular design with reinforcement learning. *JACS Au* **4**, 2160–2172 (2024).
- Pereira, T., Abbasi, M., Ribeiro, B. & Arrais, J. P. Diversity oriented deep reinforcement learning for targeted molecule generation. *J. Cheminform.* **13**, 21 (2021).
- Thomas, M., O'Boyle, N. M., Bender, A. & De Graaf, C. Augmented Hill-Climb increases reinforcement learning efficiency for language-based de novo molecule generation. *J. Cheminform.* **14**, 68 (2022).
- Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
- Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

23. Huang, L. et al. A dual diffusion model enables 3D molecule generation and lead optimization based on target pockets. *Nat. Commun.* **15**, 2657 (2024).
24. Luo, S., Guan, J., Ma, J. & Peng, J. A 3D generative model for structure-based drug design. *Adv. Neural Inf. Process. Syst.* **34**, 6229–6239 (2021).
25. Ragoza, M., Masuda, T. & Koes, D. R. Generating 3D molecules conditional on receptor binding sites with deep generative models. *Chem. Sci.* **13**, 2701–2713 (2022).
26. Wang, L. et al. A pocket-based 3D molecule generative model fueled by experimental electron density. *Sci. Rep.* **12**, 15100 (2022).
27. Xu, M., Ran, T. & Chen, H. De novo molecule design through the molecular generative model conditioned by 3D information of protein binding sites. *J. Chem. Inf. Model.* **61**, 3240–3254 (2021).
28. Zhung, W., Kim, H. & Kim, W. Y. 3D molecular generative framework for interaction-guided drug design. *Nat. Commun.* **15**, 2688 (2024).
29. Nagarajan, N. et al. Application of computational biology and artificial intelligence technologies in cancer precision drug discovery. *BioMed Res. Int.* **2019**, 8427042 (2019).
30. Pereira, T. et al. Deep generative model for therapeutic targets using transcriptomic disease-associated data—USP7 case study. *Brief. Bioinform.* **23**, bbac270 (2022).
31. Liu, X. et al. GraphCDR: a graph neural network method with contrastive learning for cancer drug response prediction. *Brief. Bioinform.* **23**, bbab457 (2022).
32. Jiang, L. et al. DeepTTA: a transformer-based model for predicting cancer drug response. *Brief. Bioinform.* **23**, bbac100 (2022).
33. Kuenzi, B. M. et al. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* **38**, 672–684.e6 (2020).
34. Subramanian, A. et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452.e17 (2017).
35. Yang, W. et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41**, D955–D961 (2012).
36. Basu, A. et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* **154**, 1151–1161 (2013).
37. Rees, M. G. et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat. Chem. Biol.* **12**, 109–116 (2016).
38. Seashore-Ludlow, B. et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov.* **5**, 1210–1223 (2015).
39. Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* **6**, 813–823 (2006).
40. Das, D., Chakrabarty, B., Srinivasan, R. & Roy, A. Gex2SGen: designing drug-like molecules from desired gene expression signatures. *J. Chem. Inf. Model.* **63**, 1882–1893 (2023).
41. Li, C. & Yamanishi, Y. *Proc. AAAI Conference on Artificial Intelligence* 13455–13463 (AAAI Press, 2024).
42. Wang, C., Ong, H. H., Chiba, S. & Rajapakse, J. C. GLDM: hit molecule generation with constrained graph latent diffusion model. *Brief. Bioinform.* **25**, <https://doi.org/10.1093/bib/bbae142> (2024).
43. Liu, Y. et al. TransGEM: a molecule generation model based on transformer with gene expression data. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btae189> (2024).
44. Méndez-Lucio, O., Baillif, B., Clevert, D.-A., Rouquié, D. & Wichard, J. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nat. Commun.* **11**, 10 (2020).
45. Pravalphruekul, N., Piriyaikitakonkij, M., Phunchongharn, P. & Piyayotai, S. De novo design of molecules with multi-action potential from differential gene expression using variational autoencoder. *J. Chem. Inf. Model.* **63**, 3999–4011 (2023).
46. Joo, S., Kim, M. S., Yang, J. & Park, J. Generative model for proposing drug candidates satisfying anticancer properties using a conditional variational autoencoder. *ACS Omega* **5**, 18642–18650 (2020).
47. Park, S. & Lee, H. A molecular generative model with genetic algorithm and tree search for cancer samples. Preprint at <https://arxiv.org/abs/2112.08959> (2021).
48. Born, J. et al. PaccMannRL: De novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning. *iScience* **24**, 102269 (2021).
49. Whitehead, A. & Crawford, D. L. Variation in tissue-specific gene expression among natural populations. *Genome Biol.* **6**, 1–14 (2005).
50. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **33**, 6840–6851 (2020).
51. Radford, A. et al. *Proc. International Conference on Machine Learning* 8748–8763 (PMLR, 2021).
52. Krenn, M. et al. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn.: Sci. Technol.* **1**, 045024 (2020).
53. Skinnider, M. A. Invalid SMILES are beneficial rather than detrimental to chemical language models. *Nat. Mach. Intell.* **6**, 437–448 (2024).
54. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **4**, 90–98 (2012).
55. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **1**, 1–11 (2009).
56. Wildman, S. A. & Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **39**, 868–873 (1999).
57. Preuer, K., Renz, P., Unterthiner, T., Hochreiter, S. & Klambauer, G. Fréchet ChemNet distance: a metric for generative models for molecules in drug discovery. *J. Chem. Inf. Model.* **58**, 1736–1741 (2018).
58. Jin, I. & Nam, H. HiDRA: hierarchical network for drug response prediction with attention. *J. Chem. Inf. Model.* **61**, 3858–3867 (2021).
59. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **39**, 2887–2893 (1996).
60. Fu, L. et al. ADMETlab 3.0: an updated comprehensive online ADMET prediction platform enhanced with broader coverage, improved performance, API functionality and decision support. *Nucleic Acids Res.* **52**, W422–W431 (2024).
61. Voršilák, M., Kolář, M., Čmelo, I. & Svozil, D. SYBA: Bayesian estimation of synthetic accessibility of organic compounds. *J. Cheminform.* **12**, 35 (2020).
62. Masci, D. et al. Recent advances in drug discovery for triple-negative breast cancer treatment. *Molecules* **28**, 7513 (2023).
63. Ho, J. & Salimans, T. Classifier-free diffusion guidance. Preprint at <https://arxiv.org/abs/2207.12598> (2022).
64. Zheng, F. et al. Interpretation of cancer mutations using a multiscale map of protein systems. *Science* **374**, eabf3067 (2021).
65. SciFinder®, <https://scifinder-n.cas.org/>.
66. Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F. & Schroeder, M. PLIP: fully automated protein–ligand interaction profiler. *Nucleic Acids Res.* **43**, W443–W447 (2015).
67. Pan, Y., Huang, N., Cho, S. & Mackerell, A. D. Consideration of molecular weight during compound selection in virtual target-based database screening. *J. Chem. Inf. Comput. Sci.* **43**, 267–272 (2003).
68. Jang, G. et al. Predicting mechanism of action of novel compounds using compound structure and transcriptomic signature coembedding. *Bioinformatics* **37**, i376–i382 (2021).

69. Dollar, O., Joshi, N., Pfaendtner, J. & Beck, D. A. Efficient 3d molecular design with an e(3) invariant transformer VAE. *J. Phys. Chem. A* **127**, 7844–7852 (2023).
70. Rigoni, D., Navarin, N. & Sperduti, A. Rgcvae: relational graph conditioned variational autoencoder for molecule design. *Mach. Learn.* **114**, 47 (2025).
71. Wu, H., Ye, X. & Yan, J. QVAE-mole: the quantum VAE with spherical latent variable learning for 3-D molecule generation. *Adv. Neural Inf. Process. Syst.* **37**, 22745–22771 (2025).
72. Brown, N., Fiscato, M., Segler, M. H. & Vaucher, A. C. GuacaMol: benchmarking models for de novo molecular design. *J. Chem. Inf. Model.* **59**, 1096–1108 (2019).
73. RDKit: open-source cheminformatics, <https://www.rdkit.org/> (2025).
74. Park, S. et al. A deep learning model of tumor cell architecture elucidates response and resistance to CDK4/6 inhibitors. *Nat. Cancer* **5**, 996–1009 (2024).
75. Dollar, O., Joshi, N., Beck, D. A. & Pfaendtner, J. Attention-based generative models for de novo molecular design. *Chem. Sci.* **12**, 8362–8372 (2021).
76. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 6000–6010 (2017).
77. Huang, X. & Belongie, S. *Proc. IEEE International Conference on Computer Vision* 1501–1510 (IEEE Computer Society, 2017).
78. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. Preprint at <https://arxiv.org/abs/1312.6114> (2013).
79. Higgins, I. et al. beta-vae: learning basic visual concepts with a constrained variational framework. *Proc. Int. Conf. Learn. Represent.* <https://openreview.net/forum?id=Sy2fzU9gl> (2017).
80. Fu, H. et al. Cyclical annealing schedule: a simple approach to mitigating KL vanishing. *Proc. 2019 Conf. North American Chapter Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 240–250 (Association for Computational Linguistics, 2019).
81. Mitrovic, J., McWilliams, B. & Rey, M. Less can be more in contrastive learning. *PMLR* **137**, 70–75 (2020).
82. Paszke, A. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Proc. 33rd International Conference on Neural Information Processing Systems* **721**, 8026–8037 (2019).
83. Wolf, T. et al. Transformers: State-of-the-Art Natural Language Processing. *Proc. 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45 (Association for Computational Linguistics, 2020).
84. Goh, G. B., Siegel, C. M., Vishnu, A. & Hodas, N. O. ChemNet: a transferable and generalizable deep neural network for small-molecule property prediction. (Pacific Northwest National Lab. (PNNL), 2017).
85. Solomon, J. Optimal transport on discrete domains. *Proc. Symposia in Applied Mathematics* (2018).
86. Pugh, T. J. et al. AACR Project GENIE: 100,000 cases and beyond. *Cancer Discov.* **12**, 2044–2057 (2022).
87. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, p11 (2013).
88. Alhossary, A., Handoko, S. D., Mu, Y. & Kwok, C.-K. Fast, accurate, and reliable molecular docking with QuickVina 2. *Bioinformatics* **31**, 2214–2216 (2015).
89. Adewole, K., Ishola, A. & Olaoye, I. In silico profiling of histone deacetylase inhibitory activity of compounds isolated from *Cajanus cajan*. *Beni Suef Univ. J. Basic Appl. Sci.* **11**, 1–22 (2022).
90. DeLano, W. L. PyMol: an open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr.* **40**, 82–92 (2002).
91. Genheden, S. et al. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J. Cheminform.* **12**, 70 (2020).
92. Irwin, J. J. et al. ZINC20—a free ultralarge-scale chemical database for ligand discovery. *J. Chem. Inf. Model.* **60**, 6065–6073 (2020).
93. Grygorenko, O. O. Enamine Ltd.: The Science and Business of Organic Chemistry and Beyond. *Eur. J. Org. Chem.* **47**, 6474–6477 (2021).
94. Thakkar, A., Kogej, T., Reymond, J.-L., Engkvist, O. & Bjerrum, E. J. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chem. Sci.* **11**, 154–168 (2020).
95. Saigiridharan, L. et al. AiZynthFinder 4.0: developments based on learnings from 3 years of industrial application. *J. Cheminform.* **16**, 57 (2024).
96. Hyunho Kim, B. B., Park, M., Shin, Y., Ideker, T. & Nam, H. A genotype-to-drug diffusion model for generation of tailored anti-cancer small molecules. G2D-Diff, <https://doi.org/10.5281/zenodo.15265967> (2024).

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00334990), and supported by a grant of the Korea Machine Learning Ledger Orchestration for Drug Discovery Project (K-MELLODDY) funded by the Ministry of Health & Welfare and Ministry of Science and ICT, Republic of Korea (RS-2024-00460010). Additionally, H.K. was supported by the Korea Institute of Toxicology, Republic of Korea (2710008763), and T.I. was supported by the Bridge2AI program of the National Institutes of Health (OD032742). We appreciate the high-performance GPU computing support of the Artificial Intelligence Industrial Convergence Cluster Development Project funded by the Ministry of Science and ICT (MSIT, Korea) & Gwangju Metropolitan City, and HPC-AI Open Infrastructure via GIST SCENT. We would like to express our gratitude to Dr I. Lee, Dr S. Park, and Dr A. Singhal for their valuable suggestions and for providing the training data.

Author contributions

H.K., T.I., and H.N. conceptualized the study. H.K. implemented the whole methodological pipeline. B.B. and M.P. implemented and evaluated Chemical VAE. B.B. and M.P. conducted retrosynthesis predictions. Y.S. analyzed the generated chemical structures. H.K. and H.N. interpreted the results. H.K., B.B., M.P., and Y.S. wrote the initial manuscript. T.I. and H.N. reviewed and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-60763-9>.

Correspondence and requests for materials should be addressed to Trey Ideker or Hojung Nam.

Peer review information *Nature Communications* thanks Khalid Raza and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025