Novel Molecule design with POWGAN, a Policy-Optimized Wasserstein Generative Adversarial Networks

Bruno Macedo

up200601848@edu.med.up.pt

University of Porto https://orcid.org/0000-0002-0127-6573

Tiago Taveira-Gomes

Faculty of Medicine, University of Porto

Inês Ribeiro-Vaz

Faculty of Medicine, University of Porto

Article

Keywords:

Posted Date: March 26th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6149551/v1

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

Novel Molecule design with POWGAN, a Policy-Optimized Wasserstein Generative Adversarial Networks

Authors

Bruno Macedo[1,6], Inês Ribeiro Vaz[1,2,3], Tiago Taveira Gomes[1,2,4,5]

Affiliations

[1] Faculty of Medicine, University of Porto, Porto, Portugal
[2] Department of Community Medicine, Information and Decision in Health, Faculty of Medicine, University of Porto, Porto, Portugal
[3] Center for Health Technology and Services Research (CINTESIS), Porto, Portugal
[4] Faculty of Health Sciences, University Fernando Pessoa, Porto, Portugal
[5] SIGIL Scientific Enterprises, Dubai, UAE
[6] MedFacts Lda., Lisbon, Portugal

Corresponding author

Correspondence to Bruno Macedo

Abstract

This study introduces Policy Optimized Wasserstein GAN (POWGAN), a novel generative model that integrates reinforcement learning policy-driven optimization. POWGAN employs a dynamically scaled reward function that adaptively adjusts the training focus, promoting the generation of novel molecules with targeted properties, such as graph connectivity to deliver non-fragmented molecules. The results demonstrated substantial improvements over previous approaches, highlighting the model's ability to achieve nearly 100% connectivity and significantly enhance generative capacity by up to eight-fold, producing more than 10,000 novel molecules. R-MedGAN, utilizing POWGAN's capability to produce structurally diverse molecules, facilitated the exploration of novel chemical regions and substantially expanded the accessible chemical space. These findings underscore the effectiveness of adaptive reinforcement-driven strategies in generative adversarial networks oriented by rewards for molecular discovery.

Introduction (without heading)

The pharmaceutical industry is currently facing a productivity crisis, with the cost of developing a new drug doubling approximately every nine years, a phenomenon known as Eroom's Law, which inversely mirrors Moore's Law in technological development [1,2]. This stagnation is exacerbated by the high attrition rates in drug development, where a large proportion of compounds fail to progress past clinical trials owing to efficacy or safety concerns. Traditional drug discovery methods are not only costly and time-consuming but are also limited by the chemical spaces that can be practically explored using conventional high-throughput screening methods [1].

Generative Artificial Intelligence offers a transformative approach to this challenge by enabling the exploration of vast, uncharted chemical spaces at high speed and lower costs [3]. The demand for novel drugs increasingly leverages the power of computational models, particularly through the application of unsupervised learning techniques, such as Generative Adversarial Networks (GANs). [3–5]

Although GANs have revolutionized fields such as image synthesis and natural language processing, their adoption in drug design is still maturing, particularly for complex tasks such as generating chemically valid and pharmacologically relevant molecules [3]. GANs autonomously learn and generate novel molecular structures, potentially uncovering drug candidates that may not be intuitively obvious through conventional methods. This capability to innovate *in silico* is crucial, as it could substantially reduce the time and financial investments required to identify viable candidates for further development [3]. Moreover, the ability of GANs to generate diverse and complex molecular structures from unlabeled datasets perfectly aligns with the needs of modern drug discovery, where understanding complex disease mechanisms and the molecular basis of drug action is paramount. [3,6]

Despite their potential, the deployment of GANs in drug discovery has been limited by inherent challenges related to their stability, quality of generated samples, and need for extensive computational resources, often leading to inefficiencies on the benefits of using AI in this context [7]. Addressing these challenges is crucial for leveraging the full potential of generative AI in transforming drug discovery processes, thus mitigating the effects of Eroom's law by introducing more cost-effective and innovative therapeutic solutions.

Despite considerable advancements in GANs for drug design in the past years, several critical challenges remain largely unaddressed, which restrict their practical application in the pharmaceutical industry. Traditional GAN frameworks still struggle with training stability and produce high variability in sample quality, which can severely limit their utility in generating chemically valid and pharmacologically viable molecules.[5,7–9]

Significant adaptations from the original GAN architectures have been made to address challenges in molecular design, notably to enhance the novelty, complexity, and diversity of the generated molecules [8,9]. MolGAN integrates reinforcement learning to optimize specific molecular properties, achieving nearly 100% generation of valid compounds, as demonstrated by its application to the QM9 chemical database. However, it works for very small structures, struggles with scalability and is susceptible to mode collapse, often failing to maintain the diversity in the generated molecules [10]. MolGAN excelled in synthesizability with scores of up to 0.99 and drug-likeness but faced limitations in diversity (0.75), as observed in structured evaluations against baselines such as SeqGAN and ORGAN [11,12]. ORGAN, which employs an objective-reinforced approach, optimizes the chemical properties directly but incurs high computational costs and complex training dynamics. It has shown the ability to generate diverse molecular samples with a high validity of up to 96.1% and diversity scores reaching 0.92, yet struggles with balancing objective-specific metrics and the discriminator's feedback, which can slow convergence and escalate resource demands[12]. Our previous work, MedGAN, a Generative Adversarial Network with Graph Convolutional Networks and Wassertein loss for generating quinoline derivatives from graphs data, demonstrated the ability to produce chemically valid structures up to 50 atoms with a high rate of novelty (93% novel molecules) and uniqueness (95% unique molecules), with 25% of its outputs being fully valid molecules obtained from molecular graphs and 62% of them being non-fragmented (connected) [6,8,9]. Despite these advances, the random generation process of MedGAN is not oriented towards specific molecular properties, prompting the need for a more directed and efficient generative process. It achieved quinoline-specific design with significant implications for generating complex molecules of up to 50 atoms, retaining both novelty and diversity, as evidenced by its superior performance in generating novel and unique quinoline molecules that were absent from the training and known datasets (1 million molecules from ZINC15 dataset), but it struggled on generating fully connected molecules effectively [6,13].

Furthermore, previous models have shown limitations in maintaining molecular diversity and complexity without sacrificing other essential properties such as drug-likeness and synthesizability [10,12,14]. These models often struggle to balance the trade-offs between optimizing specific molecular properties and retaining broad generative capacity. This results in overly simplistic molecular structures that lack functional viability, or highly complex molecules that do not align with the desired pharmacological profiles. Moreover, the simultaneous integration and optimization of multiple molecular properties, such as connectivity and bioactivity, without compromising fundamental adversarial learning principles, remains a significant hurdle. These persistent gaps underscore the need for a more refined and dynamic approach to GANs in molecular design, one that can adaptively modulate the adversarial training

process to produce more targeted and functionally relevant molecular outputs.[10,12,15]

This study introduces a novel framework, the Policy-Optimized Wasserstein Generative Adversarial Network (POWGAN), designed to enhance the generative process by systematically addressing these limitations. POWGAN incorporates advanced reinforcement learning strategies that not only guide the generation of molecular structures possessing the desired chemical properties and biological activities but also refine how these properties are integrated during the generative process. The theoretical concept leading to the development of POWGAN is driven by the necessity to control the learning process of GANs to produce novel and functionally targeted molecules. This was inspired by the initial success of MedGAN, which achieved the generation of interesting molecules in a largely stochastic manner [6].

POWGAN leverages the foundational principles of Wasserstein GANs, known for their stability and robust training characteristics, combined with a Relational Graph Convolutional Network (RGCN) [8,9,16]. This combination significantly improves the ability to generate complex molecular structures. By introducing a policy optimization mechanism that dynamically adapts the generative process towards a reward related to the final molecule output, POWGAN ensures that each generated molecule not only adheres to the desired chemical profiles but also aligns with pharmacological targets, effectively bridging a critical gap in automated drug discovery.

Moreover, the introduction of POWGAN underscores a significant shift towards utilizing unsupervised learning not only as a data generation tool but also as a sophisticated framework capable of understanding patterns linked to bioactivity goals. This paradigm shift is crucial in pharmaceutical research, where the potential chemical space is vast and largely unexplored and unsupervised methods can yield unprecedented insights and innovations.

This study details the architecture of POWGAN and the theoretical advancements it brings over traditional GANs on drug discovery. The empirical results highlight the efficacy of POWGAN in producing valid, connected, and complex molecular structures with optimized properties, thereby demonstrating a significant improvement in AI-driven drug discovery.

Methods

Generative Adversarial Networks (GAN). The original GAN framework introduced by Goodfellow *et al.* uses a minimax game between a generator G and a discriminator D [9]. The generator attempts to produce samples that are indistinguishable from real data, while the discriminator attempts to classify inputs as real or fake. The loss function represents an iterative two-player optimization problem, meaning D and G are optimized in alternating steps. The loss is given by:

$$\min_{\mathrm{G}} \max_{\mathrm{D}} \mathcal{L}\,(\mathrm{D},\mathrm{G}) = \mathrm{E}_{x\sim \mathrm{p_{dt}}}[\log \mathrm{D}\,(x)] + \mathrm{E}_{z\sim \mathrm{p_z}}\left[\log\left(1-\mathrm{D}\big(\mathrm{G}(z)\big)\right)\right]$$

Where:

- $D(x)$ is the discriminator's estimate of the probability that data x is real. The discriminator does not necessarily become perfectly accurate. Rather, it remains somewhat uncertain whether the generator will continue to improve.
- $G(z)$ is the generator's output given noise z.
- $D(G(z))$ is the discriminator's estimate of the probability of a fake instance being real.

The discriminator is trained to maximize this entire loss by maximizing $D(x)$ for real samples and minimizing $D\big(G(z)\big)$ for fake samples. This means the generator adjusts its output so that $D\big(G(z)\big)$ increases (closer to 1), making it harder for D to distinguish real from fake.

Wasserstein GAN (WGAN). WGAN replaces the standard GAN loss with the Wasserstein (Earth Mover's) distance, aiming to improve training stability and reduce mode collapse [8,9]. Instead of a discriminator outputting a probability, WGAN uses a critic D that scores how "real" or "fake" a sample is. The WGAN loss is given by the following equation:

$$\min_{\mathrm{G}} \max_{\mathrm{D}\in\mathcal{D}} \mathcal{L}\,(\mathrm{D},\mathrm{G}) = \mathrm{E}_{\mathrm{x}\sim \mathrm{p_{dt}}}[\mathrm{D}(x)] - \mathrm{E}_{\mathrm{z}\sim \mathrm{p_z}}\big[\mathrm{D}\big(\mathrm{G}(\mathrm{z})\big)\big]$$

$$\mathrm{W}\big(\mathrm{P_r},\mathrm{P_g}\big) = \inf_{\gamma\in\Pi(\mathrm{P_r},\mathrm{P_g})} \int |x-\mathrm{y}|\,\mathrm{d}\gamma(x,\mathrm{y})$$

$$\lambda_{\mathrm{GP}} \cdot \mathrm{E}_{x\sim \mathrm{p}_{\tilde{x}}}[(|\nabla_x D(\tilde{x})|_2 - 1)^2]$$

Where[8,9]:

- D represents the set of 1-Lipschitz functions, where the output changes linearly with respect to its input, leading to smooth and well-behaved

gradients, typically enforced through weight clipping or gradient penalty. This distance metric provides more informative gradients and improves generator updates.

- The critic D aims to maximize the difference between its scores on real and generated data, promoting effective gradient flow to the generator.
- The Wasserstein distance $\mathrm{W}(\mathrm{P_r}, \mathrm{P_g})$ measures how much "mass" must be transported to turn $\mathrm{P_g}$ into $\mathrm{P_r}$. Unlike the Kullback-Leibler or Jensen Shannon divergence used in standard GANs, that might provide vanishing or unstable gradients, this provides a meaningful notion of similarity even for distributions with disjoint support, remaining finite and smooth.
- $\tilde{x}$ are interpolated samples between the real and generated data. This prevents the critic from saturating and improves convergence.
- The gradient penalty term ensures that the critic satisfies the 1-Lipschitz constraint by penalizing deviations of $|\nabla_x D(\tilde{x})|_2$ from 1. Unlike weight clipping, which can lead to poor training dynamics, gradient penalty allows the critic to learn more expressive and stable representations.

These formulations show the evolution from the original GAN, which uses a binary cross-entropy loss, to the WGAN, which utilizes a continuous and theoretically grounded loss based on the Wasserstein distance to provide more stable and meaningful training dynamics.The standard GAN loss (cross-entropy) causes gradients vanishing when the discriminator becomes too confident, whereas the Wasserstein loss yields meaningful gradients even when samples are far from real data.[8,9]

WGAN-GP. To address the difficulties arising from weight clipping, the WGAN-GP variant adds a gradient penalty (GP) term that constrains the gradient norm of the critic to be near 1 [8,9]. Weight clipping artificially constrains the critic weights, often leading to poor gradient updates. If the clipping range is too tight, the critic becomes under-expressive; however, if it is too loose, training instability arises. WGAN-GP avoids this by softly penalizing gradient norm violations instead of enforcing hard constraints. This is the basis from which the MedGAN model built [6]. It incorporates the Wasserstein loss with a gradient penalty (GP) to improve the training stability and convergence[6]. The loss then becomes:

$$\mathcal{L}_{\text{WGAN-GP}} = \mathrm{E}_{\mathrm{x}\sim\mathrm{p_{dt}}}[\mathrm{D}(x)] - \mathrm{E}_{\mathrm{z}\sim\mathrm{p_z}}\left[\mathrm{D}\left(\mathrm{G}(\mathrm{z})\right)\right] + \lambda_{\text{GP}} \cdot \mathrm{GP}$$

Where:

$$\mathrm{GP} = \mathrm{E}_{\hat{\mathrm{x}}\sim\mathrm{T}}[(|\nabla \mathrm{D}(\hat{x})|^2 - 1)^2]$$

$$\hat{x} = \alpha x + (1-\alpha)G(z), \alpha \sim U(0,1)$$

The interpolated points $\hat{x}$ are sampled using convex combinations of real and fake samples, ensuring that the gradient penalty is enforced exactly where the critic should be most sensitive—along decision boundaries between real and generated data.

The gradient penalty ensures that D remains a 1-Lipschitz function, which is required for the Wasserstein distance to be valid. If the gradient norm deviates significantly from 1, the critic may miss the ability to approximate meaningful distances between distributions.

Policy-Optimized Wasserstein Generative Adversarial Networks (POWGAN). Integrates a dynamic scaling factor $S(t)$ and reward mechanism $r(G(z))$ to focus training on generating new molecules. A key novelty lies in how $S(t)$ is updated over time to focus more strongly on the reward criterion once the model has converged under the current scaling. The updated loss function is given by

$$\mathcal{L}_{\mathrm{POWGAN}} = \mathrm{E}_{\mathrm{x}\sim\mathrm{p_{dt}}}[\mathrm{D(x)}] - \mathrm{E}_{\mathrm{z}\sim\mathrm{p_z}}\left[\mathrm{D}\left(\mathrm{G(z)}\right)\cdot(\text{baseline or } \mathrm{S(t)})\cdot \mathrm{r}\left(\mathrm{G(z)}\right)\right] + \lambda_{\mathrm{GP}}\cdot \mathrm{GP}$$

Rather than uniformly scaling every sample by $\mathrm{S(t)}$, samples that do not meet a key criterion receive a baseline multiplier of 1.0 and samples meeting the criterion receive the (potentially increasing) factor $\mathrm{S(t)}$.

This ensures that molecules not fulfilling the criteria for reward still contribute to gradients (avoiding total neglect), while successful ones receive proportionately larger updates. This balance proved to be essential for the model to improve its generative capacity through the defined reward, otherwise (for example, applying a 0 multiplier to samples not fulfilling the reward mechanism) led to unsuccessful training and the generative capacity was lost.

In our stepwise “accordion-style” update, each epoch either resets the consecutive-failure counter $c(t)$ if improvement is observed or increments $c(t)$ otherwise. Once $c(t)$ hits a threshold λ_{win}, $S(t)$ increases by a fixed amount λ_{inc}, and the counter resets to zero. $M(t)$ denote the measured value of the target property at iteration t, and M_{max} be the best (maximum) value observed so far:

Initialization:

$$S(0) = 10, \qquad c(0) = 0$$

Update rule:

$$If\ M(t+1) > M_{max} : c(t+1) = 0, S(t+1) = S(t)$$

Otherwise:

$$c(t+1) = c(t) + 1, if\ c(t+1) \geq \lambda_{win}\ , then\ \ S(t+1) = S(t) + \lambda_{inc},$$
$$c(t+1) = 0, else\ S(t+1) = S(t)$$

$S(t)$ increases by λ_{inc} once every λ_{win} consecutive non-improvement criteria, rather than increasing continuously. When $M(t)$ improves, $c(t)$ resets, preventing further expansion of $S(t)$.

The reward function is given by

$$r(G(z)) = \sum_{i=1}^{I} w_i\, f_i\big(G(z)\big),$$

Where:

- w_i are the weights assigned to each property based on their importance to the overall molecular design goals.
- $f_i\big(G(z)\big)$ are the functions evaluating specific properties of the molecule generated by $G(z)$. These might include measures, such as connectivity (non-fragmented molecules obtained from graphs), drug-likeness, synthetic accessibility, predictive bioactivity, and specific pharmacological targets.

The reward function $r\big(G(z)\big)$ incorporates molecular properties. Each property contributes with a weight w_i, allowing the model to prioritize an objective.

This mechanism balances exploration (no sample is discarded outright) with stronger exploitation (greater reward) as soon as the model stagnates, helping to overcome local minima and drive further progress in generating desired molecules. Because $r\big(G(z)\big)$ is scaled by $S(t)$, molecules with higher rewards receive stronger gradients, ensuring that training gradually shifts towards generating optimized molecules. The scaling factor $S(t)$ follows an accordion-style dynamic: it stays constant when training is improving but increases when progress stalls. Each time the generator fails to improve the property metric $M(t)$ for λ_{win} iterations, $S(t)$ jumps by λ_{inc}, increasing the weight of the reward

term in the loss function. This mechanism forces the model to focus on optimizing the desired property until improvement is achieved.

When improvement occurs (the model consistently finds molecules closer to the desired property), the accordion remains compressed (scaling factor remains constant). At this stage there is no need for additional pressure. When improvements stop (the model repeatedly fails to find better molecules), the accordion expands, exerting more pressure (the scaling factor increases). This extra pressure forces the model to try harder to achieve the desired goal.

R-MedGAN. The foundational model in this study was MedGAN[6], leveraging a Wasserstein Generative Adversarial Network (WGAN) combined with Relational Graph Convolutional Networks (R-GCN) to generate quinoline-scaffold molecules. The architecture optimizes molecular structures by utilizing adjacency and feature tensors derived from training data, effectively capturing chemical patterns, trained using ZINC15 dataset containing 1 million quinoline molecules up to 50 atoms and 7 atom types (C, H, N, O, Cl, S, F), refining hyperparameters such as a latent space dimension of 256, an RMSprop optimizer with a learning rate of $1e^{-4}$, and a Generator and Discriminator configuration with 4,092 units each with 63,451,470 and 22,831,617 trainable parameters, respectively. This approach enabled MedGAN to generate molecules with a 93% novelty rate and 95% uniqueness, while ensuring a 92% retention of quinoline scaffolds, achieving 25% valid molecules from graphs and, among those, 62% fully connected (non-fragmented) molecules. R-MedGAN represents the evolution of MedGAN by integrating POWGAN's reinforcement-driven optimization strategy (Figure 1). While MedGAN effectively generates quinoline-scaffold molecules using a WGAN-GP architecture, R-MedGAN overcomes its limitations through a dynamic reward scaling, increasing the focus on high-connectivity molecular structures, and an improved generative capacity of novel, valid, and connected molecules compared to MedGAN, due to its limitation on generating non-fragmented molecules[6]. By leveraging POWGAN's adaptive optimization, R-MedGAN enhances molecular design capabilities, producing a lot more structurally diverse and chemically meaningful molecules suitable for drug discovery applications.

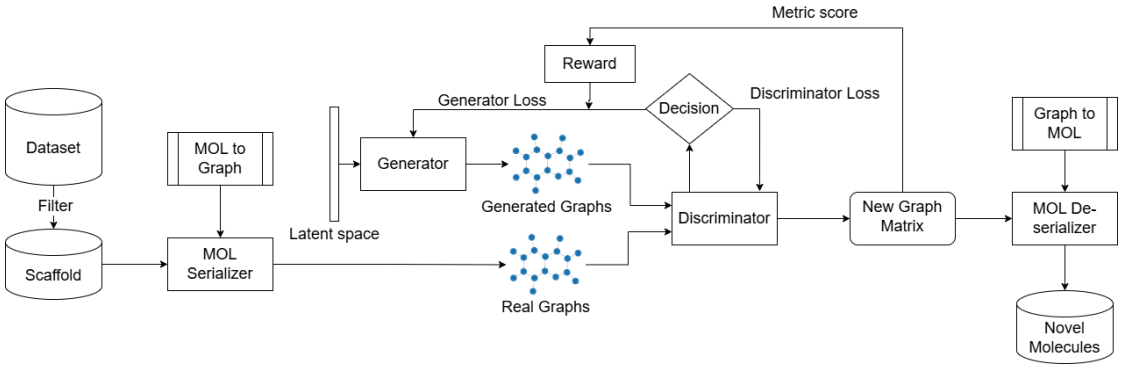


Fig. 1. R-MedGAN architecture with POWGAN algorithm for molecule generation. The generator produces candidate molecular graphs from a latent space evaluated by the discriminator against real molecular graphs. A policy-based reward function guides the training, and the optimized molecular graphs were converted back to their chemical structures, yielding novel molecules.

Results

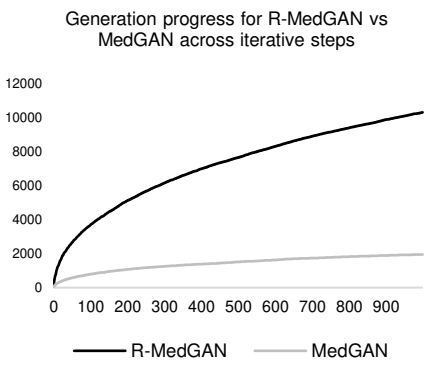
POWGAN introduces a policy-driven optimization mechanism built on WGAN-GP R-GCN by employing a dynamically scaled reward function. Unlike traditional WGAN-GP, which solely minimizes the Wasserstein distance, POWGAN prioritizes the generation of molecules with specific targeted properties by progressively amplifying the loss function whenever training stagnates.

Building upon MedGAN, which demonstrated strong novelty and uniqueness but suffered from limited connectivity and validity rates, R-MedGAN leverages POWGAN's dynamic reward scaling to enhance molecular graph connectivity and generative efficiency (Figure 1).

The reinforcement-driven optimization strategy significantly increased the fraction of fully connected molecules up to 98% while maintaining structural diversity and chemical relevance. The following results illustrate the impact of this optimization, highlighting the increased connectivity, validity, and expanded generative capacity of R-MedGAN (Table 1 and Figures 2–4). Applying the same running principles for MedGAN and R-MedGAN, such as 1,000 generative iterations, the yield for novel quinolines was 1,946 vs 10,291 while quinoline scaffold was present in 92% vs 80%, respectively, leading to an increased capacity for generalizing to different core scaffolds for R-MedGAN (Table 1).

Table 1 | R-MedGAN results. Model performance on molecule generation for each model.

	MedGAN	R-MedGAN
Latent space	256	
Activation	Tanh/ReLU	
Optimizer	RMSprop (1e-4)	
G and D units	4,092	
Training data	1 million quinolines from ZINC15	
POWGAN reward	-	Connectivity
POWGAN incremental λ_{inc}	-	10
Training iterations	300	698
Generative iterations	1,000	
Novel and unique molecules	1.946	10,291
Quinoline scaffold	92.1%	80.4%
Connectivity	0.62	0.98



Improvement in Graph Connectivity. Two strategies were tested to enhance graph connectivity: a strong fixed scaling factor $S(t) = \lambda_{inc}$, where λ_{inc} is set to 100, and an adaptive incremental scaling factor $S(t)$, starting at 10 and increasing by λ_{inc} set to 10 whenever connectivity fails to improve for λ_{win} consecutive epochs, also set to 10.

The strong fixed scaling factor rapidly increased the graph connectivity, reaching close to 100% early in the training. However, connectivity declined in subsequent epochs, ultimately stabilizing at improved levels compared with the original MedGAN model. This fixed-scaling approach doubled the generative yield of MedGAN, generating up to four thousand molecules. In contrast, the incremental scaling factor approach exhibited superior performance. Connectivity reached 98% as the dynamic scaling factor $S(t)$ eventually rose to 80, thereby improving generative capacity by more than eight-fold compared to MedGAN on the same conditions, yielding more than 10,000 novel, valid, and connected molecules (Figures 2-4).

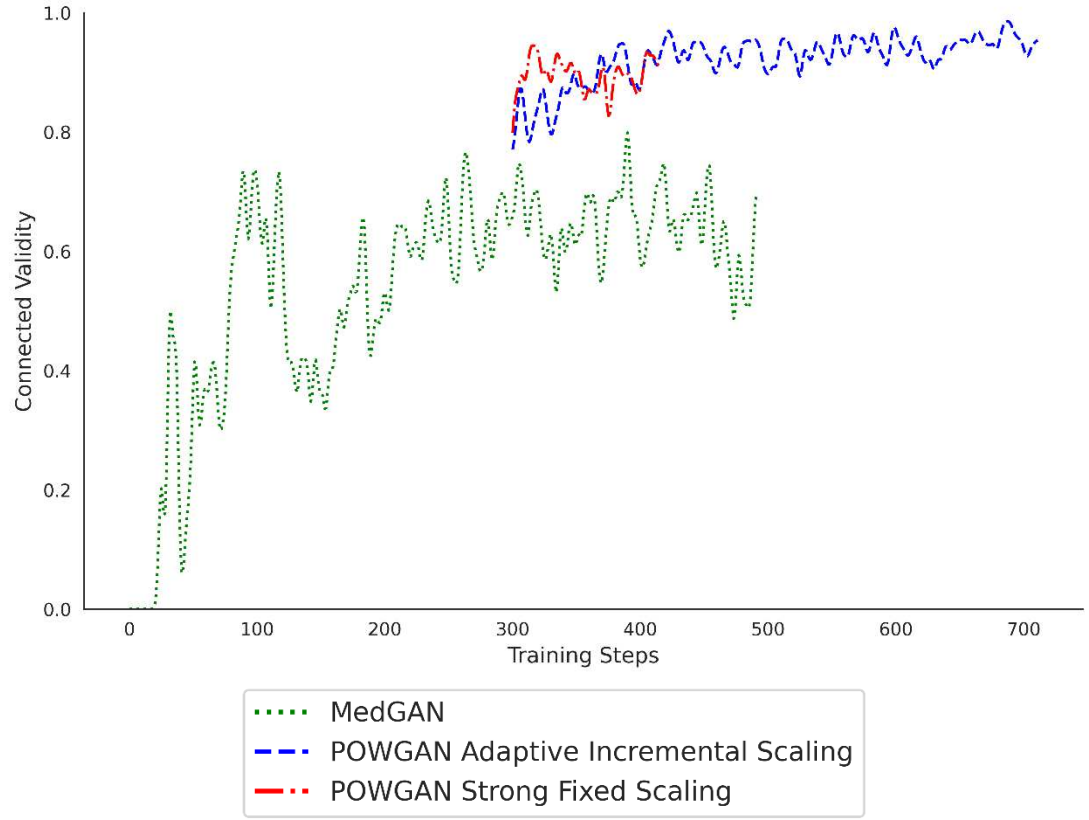
In additional experiments, we investigated what happens if non-fully-connected graphs receiving 0 as a scaling factor (no reward signal). Several runs demonstrated that once connectivity lagged behind the best observed values, the model quickly lost its ability to produce connected molecules, eventually collapsing toward the behavior of a baseline with no reinforcement. Similarly, only rewarding perfectly connected graphs resulted in comparable degradation: the generator stopped improving after a certain point and failed to recover. These observations confirm that a baseline multiplier of 1.0 for molecules non fulfilling the reward criteria is crucial. Without it, the generator is effectively penalized so severely that it ceases to explore or improve, ultimately undermining its overall capacity to produce valid, connected molecules.

For comparison purpose, the original MedGAN model ran additional 200 iterations and it was clear the limited performance increase, where connectivity

did not increase from epoch 300. To reduce noise and improve clarity, a Gaussian smoothing filter was applied to the raw data uniformously.

Connected Validity (connected molecules within the total valid molecules)

A.



B.

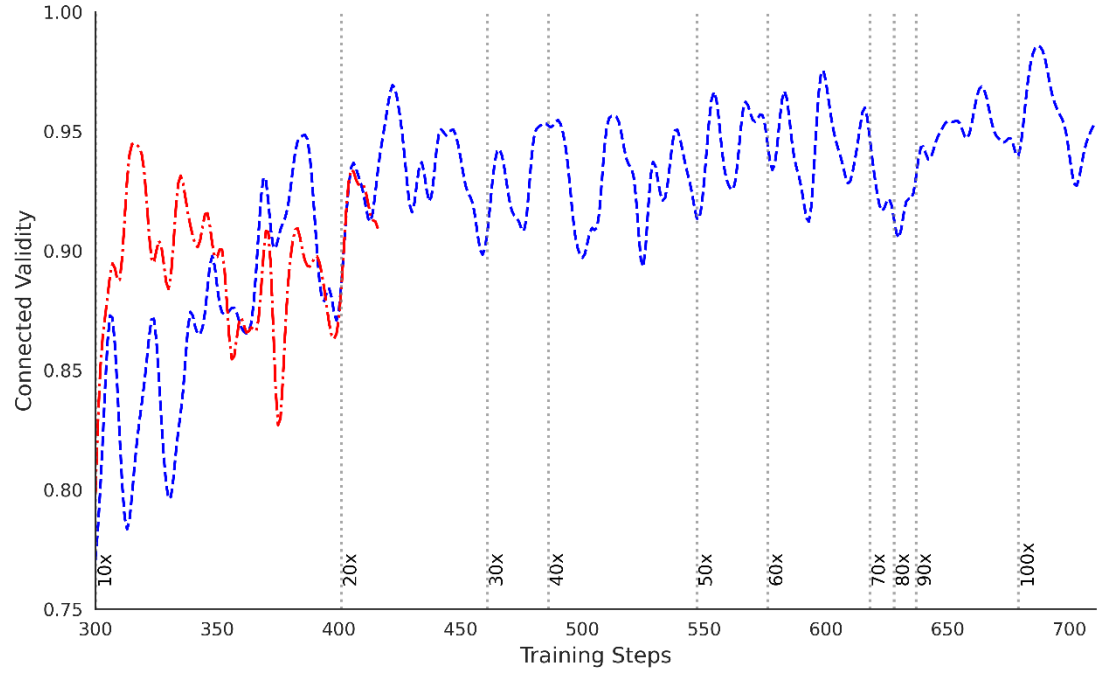


Fig. 2: Connected validity. The incremental scaling approach consistently demonstrated superior connected validity compared to the fixed scaling method and the original MedGAN model (A). The scaling iterations are detailed in B.

Validity (% of valid molecules overall)

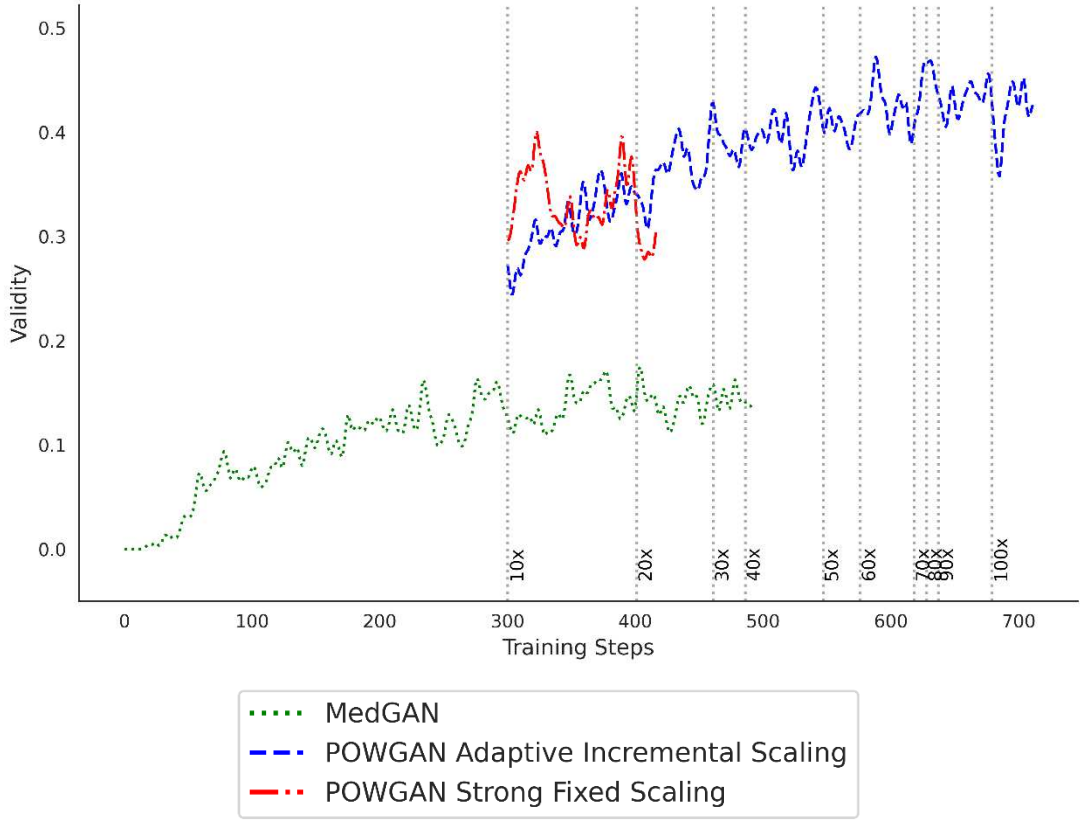


Fig. 3: Overall validity. The incremental scaling strategy maintained higher validity throughout the training than the fixed scaling and the original model.

Connectivity (% of connected molecules overall)

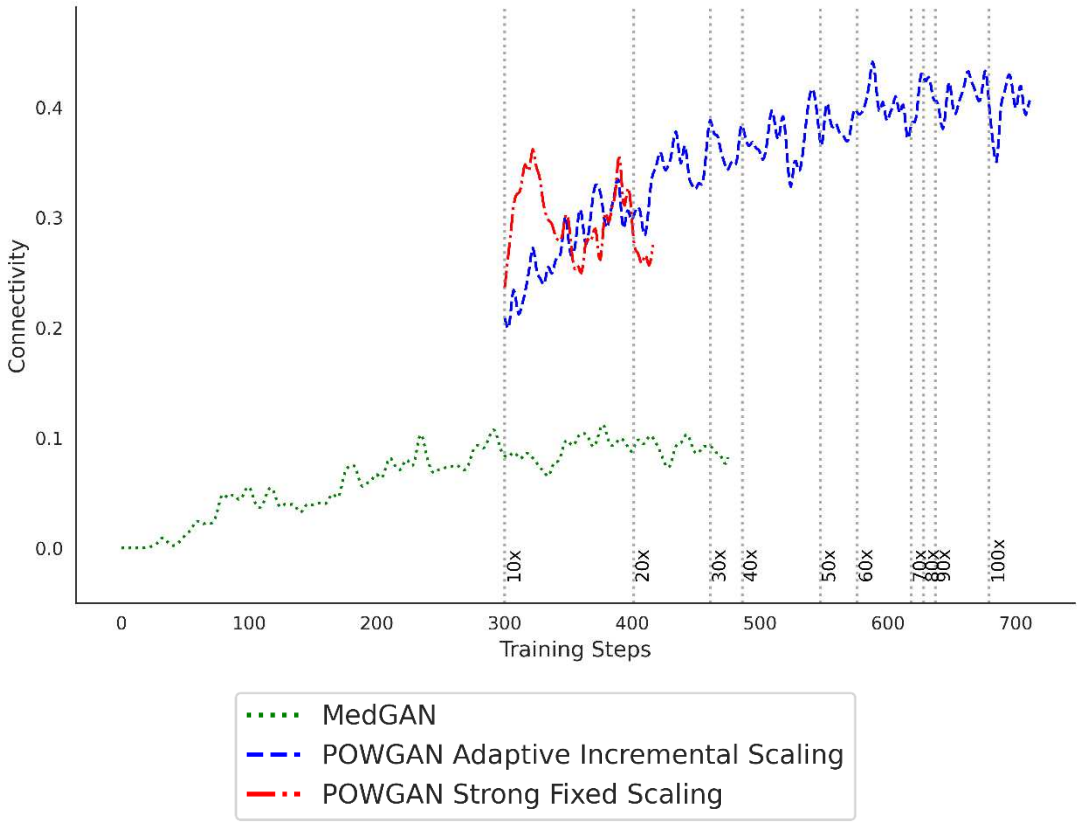


Fig. 4: Overall connectivity. The incremental scaling approach achieved consistently higher overall connectivity, significantly outperforming the original MedGAN and the fixed scaling strategies.

Further analysis using t-SNE visualization illustrated a substantial exploration of novel regions for molecules generated by R-MedGAN (using POWGAN reinforcement learning strategy) unrepresented in the original dataset and MedGAN generated molecules (Figure 5). This demonstrates POWGAN's capacity to generate structurally diverse and chemically meaningful quinoline derivatives suitable for drug discovery applications.

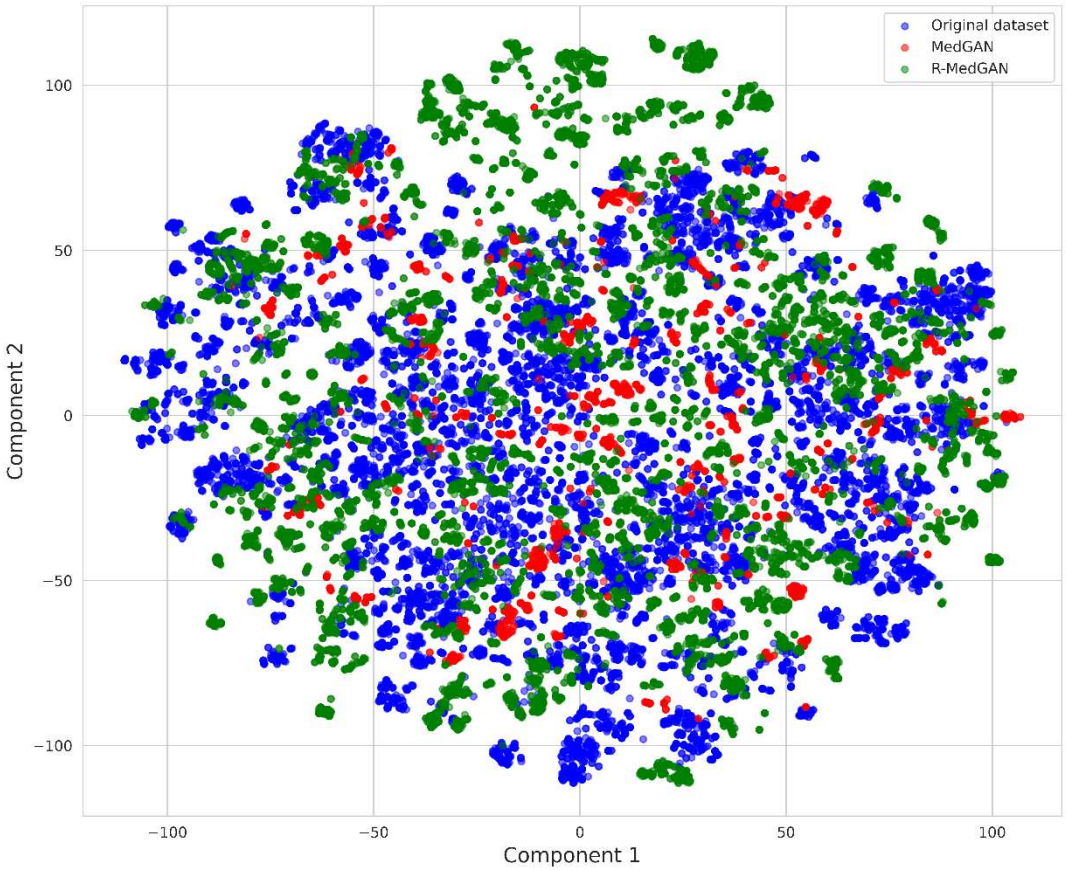
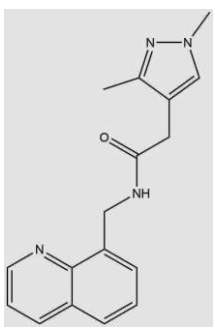
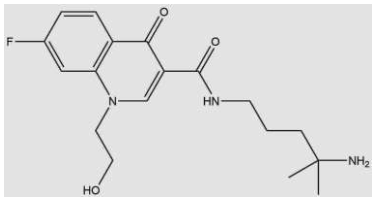


Fig. 5: t-SNE visualization of molecules. A sample of molecules generated by POWGAN (RL strategy, in green) showed proximity with samples of reference dataset (blue) and MedGAN results (red), adding a notable exploration of novel chemical regions.

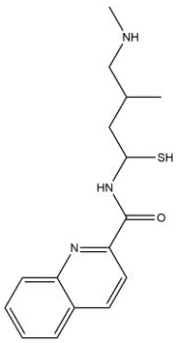
Additionally, molecules present in the R-MedGAN dataset and absent from the original MedGAN dataset confirmed POWGAN's enhanced chemical exploration capabilities, highlighting its potential in discovering novel drug candidates (Figure 6).



R-MG-Q-1-107
Similarity = 0.453



R-MG-Q-14-482
Similarity = 0.288



R-MG-Q-17-19
Similarity = 0.426

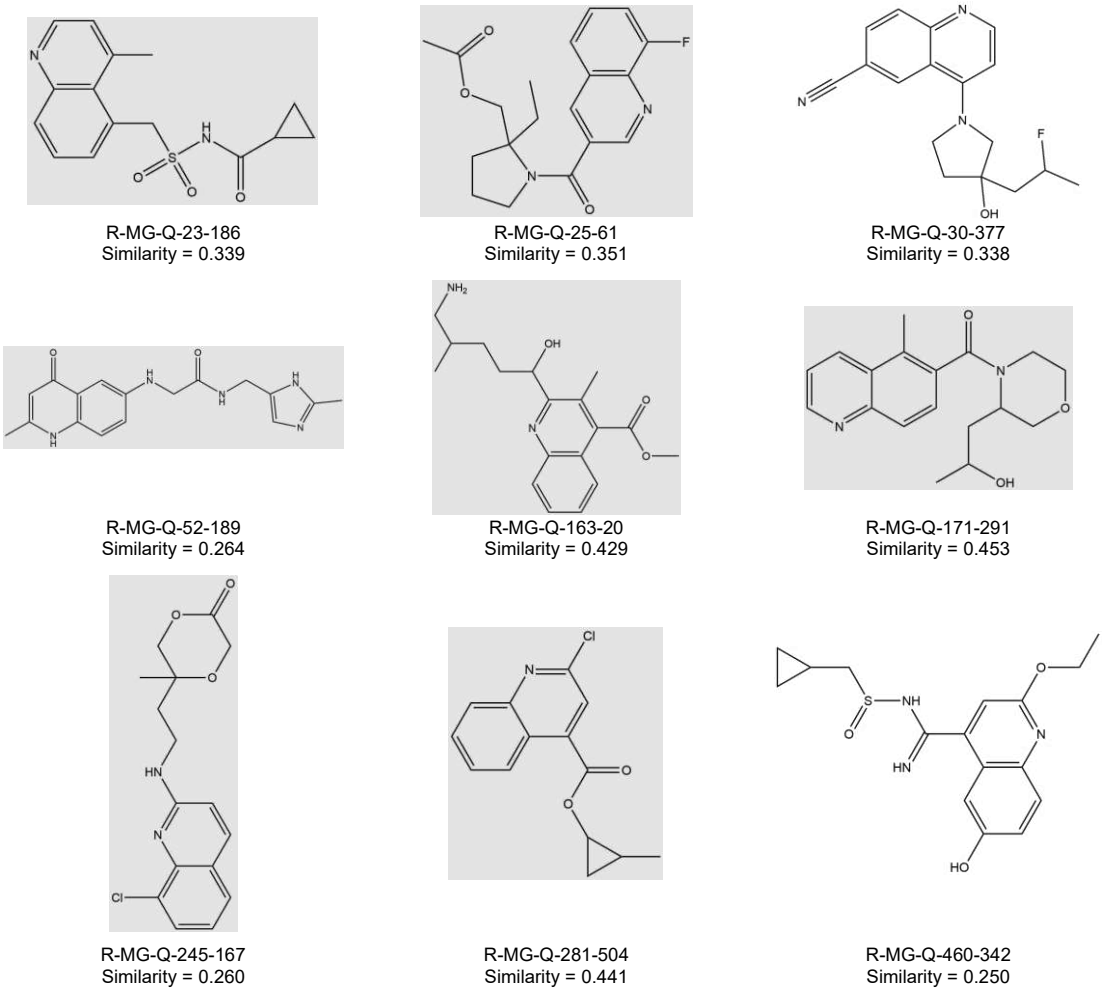


Fig. 6: Examples of R-MedGAN novel molecules. Generated with POWGAN with connectivity as a reward, absent from the original MedGAN dataset, computing the closest similarity to the MedGAN dataset and the expanded chemical diversity and exploration.

Discussion

A Generative Adversarial Network (GAN) utilizes a standard adversarial loss by setting the generator and discriminator in a minimax game. The generator aims to maximize the probability of the discriminator misclassifying its outputs as real. WGAN-GP, the core generative algorithm used in MedGAN, enhances the traditional GAN framework by adopting the Wasserstein distance, which provides robustness and training stability through a gradient penalty. Our proposed algorithm, Policy Optimized WGAN (POWGAN), represents a significant advancement by implementing a scaled and progressively oriented loss function.

Unlike previous reinforcement-driven GAN models, POWGAN effectively maintains multimetric performance through the adaptive modulation of the adversarial training path. The "accordion-style" scaling of the reward mechanism allows the model to iteratively refine local regions of the parameter

space, driven by gradient-based optimization and dynamic loss re-weighting. The key innovations of POWGAN include a policy-based reward function, r(G(z)), promoting desirable molecular properties; a dynamic loss scaling S(t), which progressively adjusts the training emphasis during periods of stagnation; and the maintenance of stable adversarial training using WGAN-GP's gradient penalty.

This unique adaptive scaling mechanism ensures high sample fidelity and diversity, which are essential for molecular generation tasks, while avoiding the difficulties of the mode collapse commonly encountered in GAN training. Convergence criteria based on stagnation of improvement, both in loss reduction and reward enhancement, prompt incremental adjustments in the loss scaling. Such iterative deepening into optimization subspaces pushes the model to explore structurally diverse molecular configurations extensively.

The results demonstrated a substantial improvement in molecular graph connectivity, validating POWGAN's effectiveness. The incremental scaling strategy proved superior to fixed scaling, significantly enhancing generative capacity and connectivity. The t-SNE visualization further corroborates POWGAN's efficacy in exploring chemically relevant and novel regions of quinoline molecular space.

We also observed that continuing to reward partially successful samples, rather than zeroing them out, was essential for robust generation. When a GAN entirely "forgets" data that do not yet meet the connectivity goal, it loses the valuable information those imperfect examples provide and can quickly collapse into producing invalid outputs. By keeping a small but nonzero reward even for molecules that fail the connectivity criterion, we preserve the model's ability to learn from partially valid configurations, maintain coverage of diverse chemical subspaces, and ultimately discover pathways to fully connected molecules. This ensures the generator remains flexible and does not overly discard promising leads that simply need further refinement, an outcome especially crucial when seeking novel structures in a high-dimensional, complex design space.

Overall, POWGAN demonstrates the viability of integrating reinforcement learning adaptive loss scaling with GAN architectures, providing a robust and versatile framework for targeted molecular generation, leading to R-MedGAN focusing on optimizing specific molecular properties, thereby enhancing the practical applicability in specialized domains such as drug discovery.

References

1. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.* **11**, 191–200 (2012).
2. Moore, G. The Future of Integrated Electronics. *Electronics Magazine* (1965).
3. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).
4. Radford, A., Metz, L. & Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *CoRR* **abs/1511.0**, (2015).
5. Goodfellow, I. *et al.* Generative adversarial networks. *NIPS'14 Proc. 27th Int. Conf. Neural Inf. Process. Syst. - Vol. 2* **63**, 139–144 (2014).
6. Macedo, B., Ribeiro-Vaz, I. & Taveira-Gomes, T. MedGAN : Optimized Generative Adversarial Network with Graph Convolutional Networks for Novel Molecule Design. *Nat. Sci. Reports* (2024) doi:10.1038/s41598-023-50834-6.
7. Polykovskiy, D. *et al.* Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Front. Pharmacol.* **11**, (2020).
8. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. & Courville, A. Improved training of wasserstein GANs. *Adv. Neural Inf. Process. Syst.* **2017-Decem**, 5768–5778 (2017).
9. Arjovsky, M., Chintala, S. & Bottou, L. Wasserstein GAN. (2017).
10. De Cao, N. & Kipf, T. MolGAN: An implicit generative model for small molecular graphs. (2018).
11. Yu, L., Zhang, W., Wang, J. & Yu, Y. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. *Proc. AAAI Conf. Artif. Intell.* **31**, (2017).
12. Guimaraes, G. L., Sanchez-Lengeling, B., Outeiral, C., Farias, P. L. C. & Aspuru-Guzik, A. Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models. (2017).
13. Sterling, T. & Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **55**, 2324–2337 (2015).
14. Maziarka, Ł. *et al.* Mol-CycleGAN: A generative model for molecular optimization. *J. Cheminform.* **12**, 1–18 (2020).
15. You, J., Liu, B., Ying, R., Pande, V. & Leskovec, J. Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation. *arXiv e-prints* arXiv:1806.02473 (2018) doi:10.48550/arXiv.1806.02473.
16. Schlichtkrull, M. *et al.* Modeling Relational Data with Graph Convolutional Networks. *arXiv e-prints* arXiv:1703.06103 (2017) doi:10.48550/arXiv.1703.06103.