WHEN VALIDITY ISN'T ENOUGH: RELIABILITY GAPS IN MOLECULAR GENERATION & KRAS CASE STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecule generation remains a core challenge in computational chemistry. Practical use of generative models is complicated by strict chemical, structural, and biological constraints: candidate compounds must satisfy physicochemical bounds, avoid reactive or toxic substructures, be synthesizable, and plausibly bind a target. We are the first to perform such comprehensive analysis of modern molecule generators via the Five-Stage Filtering Pipeline, a target-agnostic, practice-oriented benchmark for evaluating *de novo* generators using the following stages: (i) physicochemical descriptors; (ii) structural alerts; (iii) synthesis feasibility; (iv) docking and binding affinity estimation; and (v) blind medicinal chemist review. We compare 18 generators across three families (unconditional, ligand-based, and protein-based), and to make it practically relevant, apply the pipeline to KRAS G12D switch-II pocket for conditional design case study. Less than 1% of molecules pass all stages, exposing a gap between high scores on standard generative metrics and practical medicinal chemistry usage. We release our benchmark, and code to enable reproducible evaluation and to focus future model development on practically useful chemical space.

1 INTRODUCTION

One of the central challenges of biomedicine of the 21st-century is to prevent and treat complex diseases, improve population health, and extend human longevity (Hood et al., 2004; Kirkwood, 2005; Murray et al., 2012). Early drug discovery addresses this challenge through a staged pipeline: (i) identify a biological *target* associated with a disease; (ii) identify a surface, known as a *pocket*, using crystallography or pocket prediction software; and (iii) design a small molecule, known as a *ligand*, that binds and modulates the *target*. Despite decades of progress, identifying high-quality *ligands* remains labor-intensive, time-consuming, and expensive (Paul et al., 2010).

Machine learning is reshaping this landscape by accelerating design–make–test cycles for *de novo* molecular generation (Zhavoronkov et al., 2019). To be considered a viable drug candidate, a molecule must satisfy multiple criteria (Hughes et al., 2011; Waring et al., 2015; Lipinski, 2004). First, physicochemical properties must fall within reasonable bounds, e.g., limited rotatable bonds and polar surface area controlling permeability and oral exposure (Veber et al., 2002). Second, reactive chemotypes and structural alerts associated with toxicity should be removed or flagged, e.g., PAINS substructures filter out potential assay-interference compounds (Baell & Holloway, 2010; Huggins et al., 2011; Sushko et al., 2012). Third, the candidate should be practically synthesizable, which is estimated with heuristic scores and retrosynthesis planning (Ertl & Schuffenhauer, 2009; Coley et al., 2017; Genheden et al., 2020).

However, recent studies show that high scores on popular generative benchmarks often fail to translate into synthesizable, medicinally plausible compounds (Bodenreider et al., 2021). Efficient filtering of generated molecules is therefore essential prior to hit identification and lead optimization (Schneider & Fechner, 2010; Hughes et al., 2011). Without rigorous filtration, computational and experimental resources are wasted on non-viable candidates; with it, molecules meeting chemical, medicinal, and task-dependent criteria proceed further, improving success rates and reducing costs.

To the best of our knowledge, we introduce the first comprehensive, practice-oriented benchmark for evaluating *de novo* molecular generators under realistic medicinal chemistry constraints. We compare three generator families - (i) unconditional generators, (ii) ligand-based models, and

(iii) structure-aware protein-based models - and investigate whether they generate molecules that pass a realistic multi-stage filter cascade. As a biologically relevant case study, we focus on KRAS G12D mutant, for which no approved inhibitors exist; by contrast, KRAS G12C mutant has FDA approved drugs - sotorasib (Blair, 2021; Hong et al., 2020) and adagrasib (Jänne et al., 2022; Canon et al., 2019). For protein-based tasks, we focus generation on the switch-II pocket of KRAS G12D using PDB 7EW9 (PDB ID: pdb_00007ew9), a GDP-bound KRAS G12D structure in complex with TH-Z816. For ligand-based tasks, we condition generation on known KRAS G12D inhibitors (Ghazi Vakili et al., 2025).

Our contributions are:

- We propose the Five-Stage Filtering Pipeline for molecule evaluation with: coarse physicochemical descriptors, medicinal chemistry alerts, synthetic feasibility, docking and binding affinity estimation, and blind medicinal chemistry scoring.
- We propose a standardized target-agnostic filtering and evaluation process applicable to *unconditional*, *ligand-based*, and *protein-based* generators; as a case study, we employ a unified evaluation via docking and binding affinity estimation against KRAS G12D.
- We show that under our pipeline, only a small fraction (less than 1%) of generated molecules pass all filters and remain applicable for future work.
- We demonstrate that unconditional models show the highest pass rates among other families; ligand-based models more often violate coarse descriptor bounds; and protein-based models show the lowest pass rate.

Overall, our benchmark prioritizes stress-testing diverse molecular generators against constraints that matter in drug discovery settings, and shifts evaluation toward actionable chemical space. The protocol is extensible to new targets by swapping the pocket definition and ligand sets while keeping the filter cascade unchanged, enabling reproducible comparisons.

2 Related Work

We categorize molecular generators by generative strategy and architecture because both impose distinct inductive biases - validity and grammar errors for strings, geometry handling for 3D models, pocket alignment for pocket-based models (David et al., 2020; Bilodeau et al., 2022). Table 1 summarizes the mapping and models are described below.

Table 1: Taxonomy of molecular generators considered in our benchmark, by architecture (rows) and generative strategy (columns)

Architecture /Model Type	Unconditional	Ligand-based	Protein-based
Genetic Algorithm	—	MolFinder (Kwon & Lee, 2021)	—
Variational Autoencoder	HierGraphVAE (Jin et al., 2020) JT-VAE (Jin et al., 2018) MoLeR (Maziarz et al., 2021)	GENTRL (Zhavoronkov et al., 2019)	—
Autoregressive	MolGPT (Bagal et al., 2021)	GCPG (Zou et al., 2025) PGMG (Zhu et al., 2023) REINVENT4 (Loeffler et al., 2024)	Dragonfly (Atz et al., 2024) Pocket2Mol (Peng et al., 2022) ResGen (Zhang et al., 2023)
Diffusion	E(3)DM (Hoogeboom et al., 2022) TGM-DLM (Gong et al., 2024)	—	DiffSBDD (Schneuing et al., 2024) ProtoBind-Diff (Mistryukova et al., 2025) TargetDiff (Guan et al., 2023)
Flow matching	—	—	DrugFlow (Schneuing et al., 2025)

Genetic algorithm (GA) GA is a heuristic optimizer that evolves molecules via crossover and mutation operations. **MolFinder** (Kwon & Lee, 2021) applies Conformational Space Annealing to SMILES (Weininger, 1988), and requires no generative model to pretrain for ligand-based design.

Variational autoencoder (VAE) VAE models learn a latent distribution over chemical space with an encoder-decoder pair optimized via the ELBO (Kingma & Welling, 2013). **JT-VAE** (Jin et al., 2018), **HierGraphVAE** (Jin et al., 2020), **MoLeR** (Maziarz et al., 2021) operate on graphs with scaffold-aware decoders. They typically yield high validity and diversity, but as unconditional generators, they do not ensure target relevance. **GENTRL** (Zhavoronkov et al., 2019) is a string VAE

with Reinforcement Learning (RL) fine-tuning, which generates molecules with high similarity to target molecules.

Autoregressive models String models factorize the sequence likelihood as $\prod_i P(t_i \mid t_{<i})$. **MolGPT** (Bagal et al., 2021) is a decoder-only Transformer, with no target protein or ligands hints. REINVENT (Olivecrona et al., 2017) fine-tunes a SMILES *Prior* into an *Agent* via policy gradient to maximize a scoring function. **REINVENT4** (Loeffler et al., 2024) generalizes REINVENT to RNN or Transformer priors, combining transfer learning, curriculum learning, or RL with a multi-component scoring system for goal-directed design.

For structure-based design, 3D autoregressive models condition on a pocket P and learn conditional likelihood of a molecule M as $p_\theta(M \mid P) = \prod_{t=1}^{T} p_\theta(z_t \mid z_{<t}, P)$, where each step z_t adds atom, bond, or coordinates. **Pocket2Mol** (Peng et al., 2022), **ResGen** (Zhang et al., 2023), and **Dragonfly** (Atz et al., 2024) encode pocket geometry with SE(3) or E(3) equivariant encoders, and decode pocket-aware ligands.

Pharmacophore-based models use c as a set of 3D interaction features and geometry, introducing latent z which, via $p(x \mid c) = \int p_\theta(x \mid c, z)p(z)\, dz$, models the many-to-many relationship between pharmacophores and ligands. **PGMG** (Zhu et al., 2023) represents a pharmacophore as a fully connected graph, encodes it with a GNN, and uses a Transformer decoder to generate SMILES; stereochemistry tokens are omitted since the pharmacophore graph lacks stereo information. **GCPG** (Zou et al., 2025) is a Transformer encoder–decoder whose hidden state is modulated by gating on pharmacophore embeddings and user-set targets to property-controlled sampling.

Diffusion models Diffusion models are trained to approximate the reverse process of a predefined forward noising process (Ho et al., 2020). **E(3)DM** (Hoogeboom et al., 2022) is an E(3)-equivariant model that jointly denoise atom coordinates and types. **DiffSBDD** (Schneuing et al., 2024) is an SE(3)-equivariant 3D-conditional model that processes both atomic coordinates and categorical atom features while conditioning on the protein pocket. **TargetDiff** (Guan et al., 2023) conditions the diffusion process on a protein binding site (SE(3)-equivariant), generating ligand coordinates and atom types. Beyond 3D structure, **TGM-DLM** (Gong et al., 2024) denoises token embeddings with non-target specific prompts and post-hoc validity repair. **ProtoBind-Diff** (Mistryukova et al., 2025) is a structure-free diffusion language model, that takes a protein's amino-acid sequence, and generates target-specific ligand candidates.

Flow matching Flow matching models learn continuous-time velocity fields transporting a base distribution to the data distribution. **DrugFlow** (Schneuing et al., 2025) is a pocket-conditioned ligand generation model with flow-based sampling.

3 Benchmark Construction

We evaluate three generator families - *unconditional*, *ligand-based*, and *protein-based* - under a unified, reproducible Five-Stage Filtering Pipeline (Fig. 1). The pipeline thresholds and processing are target-agnostic; for structure-based stages we instantiate experiments on KRAS G12D (switch-II pocket; PDB ID: pdb_00007ew9).

Before any filtering, we standardize molecules with RDKit (Landrum, 2013) and Dimorphite-DL (Ropp et al., 2019). As part of the preparation stage, duplicate molecules were removed within each model's generation set, while duplicates across different models were retained. Validity is checked with RDKit. Then we do the following steps: (i) remove salts and solvents and keep largest organic fragment; (ii) add hydrogens to complete valences; (iii) normalize valence, kekulize, and sanitize; (iv) generate ionization states at $pH 7.4 \pm 0.0$; (v) preserve declared stereochemistry and, where unspecified, generate up to 8 stereocenters; (vi) generate 3D conformers via distance geometry, minimize them, and retain the lowest-energy conformer for each state.

Specific algorithms and configurations for molecule preprocessing are described in Appendix A.

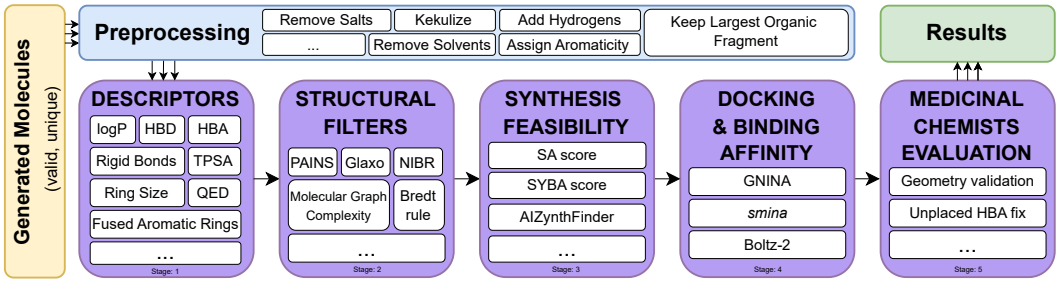


Figure 1: Five-stage filtering pipeline to evaluate generative models. At each stage molecules must satisfy all stage-specific thresholds to proceed. Starting from valid, unique generated molecules, we (i) standardize and generate chemically relevant microstates; (ii) apply physicochemical descriptor filtering; (iii) screen structural and medicinal chemistry alerts; (iv) assess synthesis feasibility and require at least one AiZynthFinder route; (v) evaluate binding compatibility and binding affinity to the KRAS G12D switch-II pocket; (vi) compounds that passed all previous stages receive a blind medicinal chemist review.

3.1 Stage 1: Physicochemical Descriptors

We compute 18 two-dimensional physicochemical descriptors (MW, logP, TPSA, HBDs, HBAs, rotatable bonds, number of rings, f_{sp^3}, QED, etc.) using RDKit after the preprocessing workflow. Rather than applying any single canonical rule set (e.g., strict Lipinski Rule-of-Five (Lipinski, 2004) or Veber's Rules (Veber et al., 2002)), we combined multiple rule sets and extended thresholds to remove clearly outliers and chemically implausible structures while retaining diversity. This approach reflects the fact that no single rule set covers all descriptors, and our aim of a general, task-agnostic filtration. Appendix B.1 reports exact per-descriptor definitions, bounds, and per-model pass rates.

3.2 Stage 2: Structural Filters

Molecules that pass the Descriptors stage are further screened with public structural alert sets and graph sanity check to remove reactive, unstable, and toxic molecules: PAINS (Baell & Holloway, 2010), Glaxo (Hann et al., 1999), Inpharmatica (Emmanuel et al., 2025), SureChEMBL (Papadatos et al., 2016); a molecular graph filter (e.g., removal of molecules containing atoms embedded in multiple 3–4-membered rings); a complexity outlier filter (e.g., Bertz (Bertz, 1981), Whitlock (Whitlock, 1998), SMCM (Allu & Oprea, 2005), TWC (Gutman et al., 2001)); the Novartis hit-triage (NIBR) filter (Schuffenhauer et al., 2020); and a Bredt-rule check (Fawcett, 1950). Implementation details, rule lists, and alert sets samples are provided in Appendix B.2.

3.3 Stage 3: Synthesis Feasibility

We assessed synthesizability with three independent predictors: Synthetic Accessibility score (SA score) (Ertl & Schuffenhauer, 2009), Retrosynthetic Accessibility score (RA score) (Thakkar et al., 2021), and Synthetic Bayesian Accessibility (SYBA) score (Voršilák et al., 2020). SA score combines fragment frequencies (from PubChem (Kim et al., 2023)) to yield a synthetic complexity score from 1 (easy) to 10 (hard). RA score is a classifier predicting probability of being a synthetic path for a compound. SYBA is a fragment-based Bernoulli Naive Bayes classifier - trained on ZINC15 "easy" and Nonpher-generated "hard" molecular sets - that classifies structures as easy or hard to synthesize.

We then attempt route finding with AiZynthFinder (Genheden et al., 2020). AiZynthFinder is a machine-learning-guided retrosynthetic workflow. It performs Monte Carlo Tree Search guided by a neural network policy over reaction templates (extracted from USPTO and applied with RDChiral), stopping when all precursors are in stock or the search depth is exceeded. Each compound was processed independently with a maximum reaction depth of 5 steps and tree search budget of 300 s

per molecule. Implementation details - e.g., exact SA score, RA score, SYBA score threshold, and AiZynthFinder configuration - are provided in Appendix B.3.

3.4 Stage 4: Docking Score and Binding Affinity Estimation

We dock all molecules that passed previous stages into the KRAS G12D switch-II pocket (PDB ID: pdb_00007ew9). Before docking, the protein structure was prepared by removing water molecules and ligands, and by adding hydrogens and charges using AutoDockTools (Forli et al., 2016). Molecular docking was performed using *smina* (Koes et al., 2013), and GNINA (McNutt et al., 2021). We estimated target binding affinity with deep learning approach Boltz-2 (Passaro et al., 2025), with a 100μM threshold.

A ligand passes Stage 4 if its best docking score is not higher than $\tau_{\text{dock}} = -6.5$ kcal/mol in both engines and binding affinity score is less than 100μM. Docking parameters are detailed in Appendix B.4.

3.5 Stage 5: Medicinal Chemists Evaluation

Molecules that passed all previous stages are scored by a senior medicinal chemist, blinded to model identity. We used PoseBusters (Buttenschoen et al., 2024), RDKit (Landrum, 2013), ProLIF (Bouysset & Fiorucci, 2021). The evaluation follows five criteria designed to capture general medicinal chemistry principles and target-specific knowledge:

- (i) Pose validation by geometry using PoseBusters: molecules exhibiting unnatural torsions, distorted bond angles, or severe intramolecular and intermolecular clashes were excluded.
- (ii) Pose validation by conformational energy using PoseBusters: docking programs frequently place ligands in energetically unfavorable conformations in order to maximize local protein–ligand interactions. If the docked pose was substantially higher in energy than alternative conformers, it was deemed unlikely to represent a realistic binding mode and the molecule was deprioritized.
- (iii) Hydrogen bond donors and acceptors using ProLIF and RDKit: unoccupied hydrogen bond donors (HBDs) and acceptors (HBAs) are penalized, as polar groups are energetically favored to remain solvent-exposed. Their presence in a buried pocket is only justified if supported by strong interactions. Particular attention was given to HBDs, whose number is more stringently limited in drug-like compounds, whereas HBAs can be somewhat more tolerated.
- (iv) Pocket burial using RDKit: to ensure that the ligand fits entirely within the binding pocket rather than protruding into solvent, the maximum distance of any ligand atom to the nearest protein atom was measured. Molecules with atoms extending farther than 5 Å were discarded.
- (v) Target-specific interaction with Asp12 using ProLIF: selectivity for KRAS G12D over wild-type KRAS critically depends on interactions with Asp12. Molecules failing to engage Asp12 were deprioritized, as their likelihood of selective binding was considered low.

4 Results

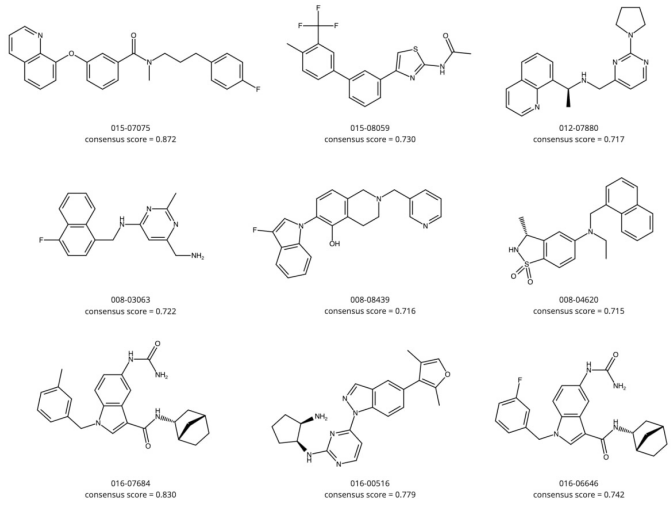
We compare three generator families with six unconditional generators, seven ligand-based generators, and nine protein-based generators. For each model we sample $N_{\text{gen}} = 10{,}000$ molecules. Validity is checked with RDKit; invalid samples are discarded and resampled; duplicates are removed and resampled within each model's batch. This yields 60,000 molecules for unconditional models, 70,000 molecules for ligand-based models, and 80,000 molecules for protein-based models. Table 2, Table 3, and Table 4 report cumulative pass rates after each step of the pipeline for each model.

We frame a set of architecture-informed hypotheses and then test them under the Pipeline. Equivariant diffusion models explicitly model 3D coordinates and Euclidean symmetries, so they are expected to produce geometrically plausible ligand poses and improved docking performance

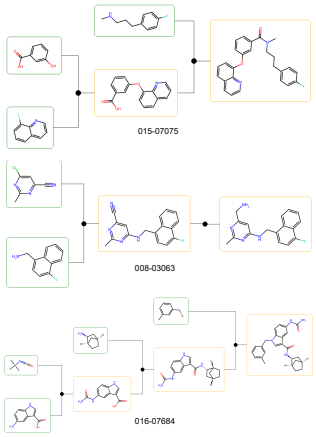
(E(3)DM, DiffSBDD, TargetDiff). Graph-based VAEs with scaffold-aware decoders have shown to yield high validity, but may sample synthetically complex chemotypes that are hard to synthesize without additional constraints (JT-VAE, HierGraphVAE, MoLeR). Autoregressive SMILES models are highly sensitive to the learned prior, which substantially alter novelty, similarity to training set, and downstream filtering rates (REINVENT4). Genetic optimizers can rapidly find high scoring and novel molecules without pretraining but may increase structural-alert incidence and reduce synthetic success (MolFinder). Flow matching approaches have reported stable training and efficient sampling that preserves training distribution fidelity (DrugFlow). Pharmacophore-guided methods explicitly bias generation toward interaction motifs and therefore are expected to increase docking enrichment (GCPG, PGMG). Finally, prior work has repeatedly shown that high performance on common generative benchmarks does not guarantee synthesizability in practice, motivating our explicit retrosynthesis and AiZynthFinder gate.

Overall pass rate is low for all families: 364 molecules (0.607% of 60,000) from unconditional generators, 287 molecules (0.41% of 70,000) from ligand-based generators, and 318 molecules (0.398% of 90,000) from protein-based generators. Unconditional models are the most successful, with 0.607% from initial number of molecules passing all stages, showing that such models are able to capture general molecular constraints much better than conditioned models. This may be due to overfitting to features that do not translate into tractable, candidates acceptable for medicinal chemistry.

Figure 2a shows the top three molecules from each model family. Consensus scores were calculated as the arithmetic mean of the inverted and min-max normalized values of *smina* and GNINA docking scores, and the Boltz-2 binding affinity predictions. Figure 2b shows some synthesis path calculated via AiZynthFinder tool for the top molecules. Rest paths are available in the Appendix B.3.



(a) Top generated molecules among three families. Top: unconditional generators (015 - MolGPT, 012 - JT-VAE), middle: protein-based generators (008 - DrugFlow), bottom: ligand-based generators (016 - GCPG).



(b) Synthesis paths for some of the top molecules. Top: MolGPT 015-07075 molecule; middle: DrugFlow 008-03063 molecule; bottom: GCPG 016-07684 molecule.

Figure 2: The top nine generated molecules with their synthesis paths.

4.1 Unconditional Molecule Generators

We evaluate E(3)DM, HierGraphVAE, JT-VAE, MoLeR, MolGPT, and TGM-DLM. These models do not condition on target ligands or pocket structure. Results are presented in Table 2. VAE models (especially JT-VAE) retain markedly more candidates through structural filters and synthetic accessibility estimation stages than E(3)DM or MolGPT models, showing that these models are able to sample molecules that are valid and not chemically complex; E(3)DM collapses at synthetic acces-

sibility stage and no candidates remain after this stage; TGM-DLM leaves with the least candidates, mostly due to struggle with validity, filtering out most molecules on descriptors stage.

Table 2: Comparison of unconditional models, each with initial number of molecules $N_{\text{gen}} = 10{,}000$

Stage /Model	E(3)DM	HIERGRAPHVAE	JT-VAE	MOLER	MOLGPT	TGM-DLM
Descriptors	3520	3579	**7586**	3193	3474	1216
Structural Filters	75	1176	**2765**	718	1029	100
Synthesis Feasibility	0	975	**1549**	557	679	35
Docking & Binding Aff.	0	477	**816**	323	340	10
Med.Chem. Evaluation	0	53	**181**	65	64	1
Pass	0	**53**	**181**	**65**	**64**	**1**

4.2 LIGAND-BASED MOLECULE GENERATORS

For benchmarking, we compare baselines: GCPG, GENTRL, MolFinder, PGMG, and three different setups of REINVENT4: REINVENT4 (V), REINVENT4 (P), REINVENT4 (TL) described below. We examine multiple REINVENT4 setups because sampling behavior depends strongly on the learned prior and fine-tuning strategy; comparing variants isolates how prior choice and transfer learning affect diversity, novelty, synthesizability, and downstream performance.
REINVENT4 (V) (vanilla) uses the out-of-the-box prior released by the authors, and no further modifications applied to the model.
REINVENT4 (P) (prior) is a similarity-based REINVENT4 prior released by the authors that was trained under a medium Tanimoto similarity sampling mode.
REINVENT4 (TL) (transfer learning) is our transfer-learned prior, fine-tuned on known KRAS G12D inhibitors to bias sampling toward the target chemical space. Training and implementation details are provided in Appendix C.

Results are presented in Table 3. GCPG sustains the highest end-to-end retention, resulting in 110 molecules after medicinal chemists evaluation stage. REINVENT4 (V) yields more success molecules (93) than REINVENT4 (P) and REINVENT4 (TL) with 17 and 32 molecules respectively, showing that a broader prior favored downstream filtering pipeline, although sampling $10{,}000$ molecules for REINVENT4 (V) required more attempts. PGMG underperforms early at the descriptors stage and retains the fewest candidates, however, the fraction of molecules that passed docking and binding affinity estimation stage with respect to synthetic feasibility stage is the highest ($19/22 = 0.864$), indicating that pharmacophore-based models tend to generate molecules, that are indeed likely to capture pocket shape geometry and complementarity, but PGMG struggles with overall molecule validity.

Table 3: Comparison of ligand-based models, each with initial number of $N_{\text{gen}} = 10{,}000$ molecules

Stage /Model	GCPG	GENTRL	MOLFINDER	PGMG	REINVENT4 (V)	REINVENT4 (P)	REINVENT4 (TL)
Descriptors	**6616**	5669	1592	195	4089	936	1204
Structural Filters	**4168**	1925	366	37	1325	593	413
Synthesis Feasibility	**1064**	303	265	22	918	222	276
Docking & Binding Aff.	**648**	238	200	19	518	72	164
Med.Chem. Evaluation	**110**	24	7	4	93	17	32
Pass	**110**	**24**	**7**	**4**	**93**	**17**	**32**

4.3 PROTEIN-BASED MOLECULE GENERATORS

For benchmarking, we compare baselines: DiffSBDD, Dragonfly, Dragonfly biased (b), DrugFlow, Pocket2Mol, ResGen, TargetDiff. We evaluated two different Dragonfly setups to investigate the overall performance of an unmodified model provided by the authors, and a fine-tuned model biased with only one target compound descriptors.
Dragonfly is an out-of-the-box model released by the authors with no modifications applied to the model.
Dragonfly (b) (biased) leverages built-in ability to condition sampling on target compound descriptors. Specifically, bias is applied toward molecular weight, number of rotatable bonds, hydrogen bond donors and acceptors, topological polar surface area, and logP, thereby steering the generation toward molecules with physicochemical properties aligned with the target profile.

Results are presented in Table 4. Although Dragonfly passes the first stage with only 27.79% of initial molecules, the number of molecules that pass medicinal chemists evaluation is the highest among all families, suggesting that Dragonfly is able to sample molecules that are likely to be valid and useful. DiffSBDD, Dragonfly, DrugFlow, Pocket2Mol and TargetDiff strongly dominate other models while passing descriptors stage. However, DiffSBDD, DrugFlow, Pocket2Mol, TargetDiff loses more than a quarter of molecules after structural filters stage, suggesting that those models struggle with synthesis of non toxic and pan-assay-free molecules. TargetDiff does not pass docking and binding affinity estimation stage, and DiffSBDD does not pass medicinal chemistry evaluation stage, while DrugFlow is the second most successful model.

Table 4: Comparison of protein-based models, each with initial number of molecules $N_{\text{gen}} = 10{,}000$

Stage /Model	DIFFSBDD	DRAGONFLY	DRAGONFLY (B)	DRUGFLOW	POCKET2MOL	PROTOBIND-DIFF	RESGEN	TARGETDIFF
Descriptors	3665	2779	1022	**5464**	2657	1466	1080	3444
Structural Filters	197	**1459**	218	1392	682	195	255	136
Synthesis Feasibility	24	**1207**	38	453	137	102	62	4
Docking & Binding Aff.	13	**575**	15	344	69	66	37	0
Med.Chem. Evaluation	0	**227**	4	62	12	7	6	0
Pass	**0**	**227**	**4**	**62**	**12**	**7**	**6**	**0**

5 DISCUSSION AND CONCLUSION

Applying the same five-stage filtration across unconditional, ligand-based, and protein-based models reveals that across $210{,}000$ generated molecules only 969 (0.461%) of generated molecules pass end-to-end screening. Empirically, unconditional models have the highest overall pass rate of 0.607%, producing molecules that correlate with basic requirements of early drug discovery. Ligand-based models achieved moderate retention with 0.41% pass rate. Protein-based models are left with the smallest fraction of passed molecules (0.398%), with Dragonfly achieving the highest final pass rate of 227 molecules despite low initial retention. Across all families, streepest attrition occurs at synthetic feasibility (≈ 0.2501 molecules w.r.t. descriptors filtration) and medicinal chemistry (≈ 0.1535 molecules w.r.t. synthesis feasibility evaluation stage) evaluation stages. This confirms prior findings that benchmark metrics, such as validity is weak predictor of downstream utility. The explicit retrosynthesis gate (AiZynthFinder) is therefore critical to separate benchmark overfitting from true drug-likeness.

Our results highlight several architecture-dependent trends across molecule generators. Equivariant diffusion models (E(3)DM, DiffSBDD, TargetDiff) exceed at encoding 3D symmetries and geometric constraints, yet collapse under synthetic accessibility evaluation, suggesting that geometric fidelity alone is insufficient for practical usage. Graph-based VAEs (JT-VAE, HierGraphVAE) balance validity and synthesizability better than other unconditional models, confirming the hypothesis that scaffold-aware decoders reduce chemical complexity. REINVENT4 is highly sensitive to prior choice: broad priors generalize well through the pipeline, while similarity-based or transfer-learned priors reduce downstream retention. Genetic optimizers (MolFinder) find high-scoring candidates without pretraining but enrich high structural alert rates, highlighting the exploration-safety trade-off. Pharmacophore-based models (GCPG, PGMG) confirm the value of explicit interaction motif bias, yielding high enrichment, although PGMG shows low overall pass-rate.

Our findings emphasize that standard generative benchmarks are not good proxies for real-world performance. Optimizing for validity, synthesis, or pocket fidelity independently is insufficient for actionable chemical space that requires alignment across all objectives simultaneously. That is why evaluation should integrate: (i) multistage filtering pipeline (descriptors, structural alerts, synthesis feasibility, docking and binding affinity, and medicinal chemistry stages); (ii) synthesis-aware metrics beyond SA scores; and (iii) stage failures analysis.

While our pipeline integrates synthesis and docking gates, it does not yet capture long-range pharmacokinetics, ADMET liabilities, or clinical viability. Future work should couple generative models with multiscale predictions, and uncertainty-aware evaluation of generated molecules.

REFERENCES

Courtney Aldrich, Carolyn Bertozzi, Gunda I Georg, Laura Kiessling, Craig Lindsley, Dennis Liotta, Kenneth M Merz Jr, Alanna Schepartz, and Shaomeng Wang. The ecstasy and agony of assay

interference compounds, 2017.

Tharun Kumar Allu and Tudor I Oprea. Rapid evaluation of synthetic and molecular complexity for in silico chemistry. *Journal of chemical information and modeling*, 45(5):1237–1243, 2005.

Kenneth Atz, Leandro Cotos, Clemens Isert, Maria Håkansson, Dorota Focht, Mattis Hilleke, David F Nippa, Michael Iff, Jann Ledergerber, Carl CG Schiebroek, et al. Prospective de novo drug design with deep interactome learning. *Nature Communications*, 15(1):3408, 2024.

Douglas S Auld, Natasha Thorne, Dac-Trung Nguyen, and James Inglese. A specific mechanism for nonspecific activation in reporter-gene assays, 2008.

Jonathan B Baell and Georgina A Holloway. New substructure filters for removal of pan assay interference compounds (pains) from screening libraries and for their exclusion in bioassays. *Journal of medicinal chemistry*, 53(7):2719–2740, 2010.

Viraj Bagal, Rishal Aggarwal, PK Vinod, and U Deva Priyakumar. Molgpt: molecular generation using a transformer-decoder model. *Journal of chemical information and modeling*, 62(9):2064–2076, 2021.

René Barone and Michel Chanon. A new and simple approach to chemical complexity. application to the synthesis of natural products. *Journal of Chemical Information and Computer Sciences*, 41 (2):269–272, 2001.

Steven H Bertz. The first general index of molecular complexity. *Journal of the American Chemical Society*, 103(12):3599–3601, 1981.

G Richard Bickerton, Gaia V Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90–98, 2012.

Camille Bilodeau, Wengong Jin, Tommi Jaakkola, Regina Barzilay, and Klavs F Jensen. Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(5):e1608, 2022.

Hannah A Blair. Sotorasib: first approval. *Drugs*, 81(13):1573–1579, 2021.

James F Blake. Identification and evaluation of molecular properties related to preclinical optimization and clinical fate. *Medicinal Chemistry*, 1(6):649–655, 2005.

Olivier Bodenreider et al. Artificial benchmarks for molecular generation do not correlate with practical drug discovery performance. *Drug Discovery Today*, 26(8):1863–1870, 2021.

Cédric Bouysset and Sébastien Fiorucci. Prolif: a library to encode molecular interactions as fingerprints. *Journal of cheminformatics*, 13(1):72, 2021.

Ruth Brenk, Alessandro Schipani, Daniel James, Agata Krasowski, Ian Hugh Gilbert, Julie Frearson, and Paul Graham Wyatt. Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *ChemMedChem: Chemistry Enabling Drug Discovery*, 3(3):435–444, 2008.

Martin Buttenschoen, Garrett M Morris, and Charlotte M Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024.

Jude Canon, Karen Rex, Anne Y Saiki, Christopher Mohr, Keegan Cooke, Dhanashri Bagal, Kevin Gaida, Tyler Holt, Charles G Knutson, Neelima Koppada, et al. The clinical kras (g12c) inhibitor amg 510 drives anti-tumour immunity. *Nature*, 575(7781):217–223, 2019.

Stephen J Capuzzi, Eugene N Muratov, and Alexander Tropsha. Phantom pains: Problems with the utility of alerts for p an-a ssay in terference compound s. *Journal of chemical information and modeling*, 57(3):417–427, 2017.

Shuan Chen and Yousung Jung. Estimating the synthetic accessibility of molecules with building block and reaction-aware sascore. *Journal of cheminformatics*, 16(1):83, 2024.

Connor W Coley, Luke Rogers, William H Green, and Klavs F Jensen. Computer-assisted retrosynthesis based on molecular similarity. *ACS central science*, 3(12):1237–1245, 2017.

Jayme L Dahlin, Jonathan Baell, and Michael A Walters. Assay interference by chemical reactivity. 2015.

Jayme L Dahlin, Douglas S Auld, Ina Rothenaigner, Steve Haney, Jonathan Z Sexton, J Willem M Nissink, Jarrod Walsh, Jonathan A Lee, John M Strelow, Francis S Willard, et al. Nuisance compounds in cellular assays. *Cell chemical biology*, 28(3):356–370, 2021.

Laurianne David, Amol Thakkar, Rocío Mercado, and Ola Engkvist. Molecular representations in ai-driven drug discovery: a review and practical guide. *Journal of cheminformatics*, 12(1):56, 2020.

Bradley Croy Doak, Björn Over, Fabrizio Giordanetto, and Jan Kihlberg. Oral druggable space beyond the rule of 5: insights from drugs and clinical candidates. *Chemistry & biology*, 21(9): 1115–1142, 2014.

Noutahi Emmanuel, Mary Hadrien, Kovary Kyle M., Whitfield Julien St-Laurent Shawn, Hounwanou Honore, and Craig Michael. datamol-io/medchem: Molecular filtering for drug discovery. 2025. doi: 10.5281/zenodo.14588938. URL https://doi.org/10.5281/zenodo.14588938.

Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics*, 1(1):8, 2009.

Frank S Fawcett. Bredt's rule of double bonds in atomic-bridged-ring structures. *Chemical Reviews*, 47(2):219–274, 1950.

Stefano Forli, Ruth Huey, Michael E Pique, Michel F Sanner, David S Goodsell, and Arthur J Olson. Computational protein–ligand docking and virtual drug screening with the autodock suite. *Nature protocols*, 11(5):905–919, 2016.

Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

Samuel Genheden, Amol Thakkar, Veronika Chadimová, Jean-Louis Reymond, Ola Engkvist, and Esben Bjerrum. Aizynthfinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of cheminformatics*, 12(1):70, 2020.

Mohammad Ghazi Vakili, Christoph Gorgulla, Jamie Snider, AkshatKumar Nigam, Dmitry Bezrukov, Daniel Varoli, Alex Aliper, Daniil Polykovsky, Krishna M Padmanabha Das, Huel Cox Iii, et al. Quantum-computing-enhanced algorithm unveils potential kras inhibitors. *Nature Biotechnology*, pp. 1–6, 2025.

Haisong Gong, Qiang Liu, Shu Wu, and Liang Wang. Text-guided molecule generation with diffusion language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 109–117, 2024.

Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. *arXiv preprint arXiv:2303.03543*, 2023.

Ivan Gutman, Christoph Rücker, and Gerta Rücker. On walks in molecular graphs. *Journal of chemical information and computer sciences*, 41(3):739–745, 2001.

Mike Hann, Brian Hudson, Xiao Lewell, Rob Lifely, Luke Miller, and Nigel Ramsden. Strategic pooling of compounds for high-throughput screening. *Journal of chemical information and computer sciences*, 39(5):897–902, 1999.

Majid M Heravi and Vahideh Zadsirjan. Prescribed drugs containing nitrogen heterocycles: an overview. *RSC advances*, 10(72):44247–44311, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

David S Hong, Marwan G Fakih, John H Strickler, Jayesh Desai, Gregory A Durm, Geoffrey I Shapiro, Gerald S Falchook, Timothy J Price, Adrian Sacher, Crystal S Denlinger, et al. Krasg12c inhibition with sotorasib in advanced solid tumors. *New England Journal of Medicine*, 383(13): 1207–1217, 2020.

Leroy Hood, James R Heath, Michael E Phelps, and Biaoyang Lin. Systems biology and new technologies enable predictive and preventative medicine. *Science*, 306(5696):640–643, 2004.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.

David J Huggins, Ashok R Venkitaraman, and David R Spring. Rational methods for the selection of diverse screening compounds. *ACS chemical biology*, 6(3):208–217, 2011.

James P Hughes, Stephen Rees, S Barrett Kalindjian, and Karen L Philpott. Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249, 2011.

John J Irwin and Brian K Shoichet. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1):177–182, 2005.

John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. Zinc: a free tool to discover chemistry for biology. *Journal of chemical information and modeling*, 52 (7):1757–1768, 2012.

Ajit Jadhav, Rafaela S Ferreira, Carleen Klumpp, Bryan T Mott, Christopher P Austin, James Inglese, Craig J Thomas, David J Maloney, Brian K Shoichet, and Anton Simeonov. Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease. *Journal of medicinal chemistry*, 53(1):37–51, 2010.

Pasi A Jänne, Gregory J Riely, Shirish M Gadgeel, Rebecca S Heist, Sai-Hong I Ou, Jose M Pacheco, Melissa L Johnson, Joshua K Sabari, Konstantinos Leventakos, Edwin Yau, et al. Adagrasib in non–small-cell lung cancer harboring a krasg12c mutation. *New England Journal of Medicine*, 387(2):120–131, 2022.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pp. 2323–2332. PMLR, 2018.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. In *International conference on machine learning*, pp. 4839–4848. PMLR, 2020.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2023 update. *Nucleic acids research*, 51(D1):D1373–D1380, 2023.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Thomas BL Kirkwood. Understanding the odd science of aging. *Cell*, 120(4):437–447, 2005.

David Ryan Koes, Matthew P Baumgartner, and Carlos J Camacho. Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise. *Journal of chemical information and modeling*, 53(8):1893–1904, 2013.

Yongbeom Kwon and Juyong Lee. Molfinder: an evolutionary algorithm for the global optimization of molecular properties and the extensive exploration of chemical space using smiles. *Journal of cheminformatics*, 13(1):24, 2021.

Greg Landrum. Rdkit documentation. *Release*, 1(1-79):4, 2013.

Christopher A Lipinski. Lead-and drug-like compounds: the rule-of-five revolution. *Drug discovery today: Technologies*, 1(4):337–341, 2004.

Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 23(1-3):3–25, 1997.

Hannes H Loeffler, Jiazhen He, Alessandro Tibo, Jon Paul Janet, Alexey Voronov, Lewis H Mervin, and Ola Engkvist. Reinvent 4: modern ai–driven generative molecule design. *Journal of Cheminformatics*, 16(1):20, 2024.

Hannes Löffler. Reinvent4 priors. 2025. doi: 10.5281/zenodo.15641297. URL https://doi.org/10.5281/zenodo.15641297.

Frank Lovering, Jack Bikker, and Christine Humblet. Escape from flatland: increasing saturation as an approach to improving clinical success. *Journal of medicinal chemistry*, 52(21):6752–6756, 2009.

Krzysztof Maziarz, Henry Jackson-Flux, Pashmina Cameron, Finton Sirockin, Nadine Schneider, Nikolaus Stiefl, Marwin Segler, and Marc Brockschmidt. Learning to extend molecular scaffolds with structural motifs. *arXiv preprint arXiv:2103.03864*, 2021.

Andrew T McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. Gnina 1.0: molecular docking with deep learning. *Journal of cheminformatics*, 13(1):43, 2021.

Lukia Mistryukova, Vladimir Manuilov, Konstantin Avchaciov, and Peter O Fedichev. Protobinddiff: A structure-free diffusion language model for protein sequence-conditioned ligand design. *bioRxiv*, pp. 2025–06, 2025.

Christopher JL Murray, Majid Ezzati, Abraham D Flaxman, Stephen Lim, Rafael Lozano, Catherine Michaud, Mohsen Naghavi, Joshua A Salomon, Kenji Shibuya, Theo Vos, et al. Gbd 2010: design, definitions, and metrics. *The Lancet*, 380(9859):2063–2066, 2012.

Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics*, 9(1):48, 2017.

Tudor I Oprea, Andrew M Davis, Simon J Teague, and Paul D Leeson. Is there a difference between leads and drugs? a historical perspective. *Journal of chemical information and computer sciences*, 41(5):1308–1315, 2001.

George Papadatos, Mark Davies, Nathan Dedman, Jon Chambers, Anna Gaulton, James Siddle, Richard Koks, Sean A Irvine, Joe Pettersson, Nicko Goncharoff, et al. Surechembl: a large-scale, chemically annotated patent document database. *Nucleic acids research*, 44(D1):D1220–D1228, 2016.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, pp. 2025–06, 2025.

Steven M Paul, Daniel S Mytelka, Christopher T Dunwiddie, Charles C Persinger, Bernard H Munos, Stacy R Lindborg, and Aaron L Schacht. How to improve r&d productivity: the pharmaceutical industry's grand challenge. *Nature reviews Drug discovery*, 9(3):203–214, 2010.

Bradley C Pearce, Michael J Sofia, Andrew C Good, Dieter M Drexler, and David A Stock. An empirical process for the design of high-throughput screening deck filters. *Journal of chemical information and modeling*, 46(3):1060–1068, 2006.

Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. In *International Conference on Machine Learning*, pp. 17644–17655. PMLR, 2022.

Patrick J Ropp, Jesse C Kaminsky, Sara Yablonski, and Jacob D Durrant. Dimorphite-dl: an open-source program for enumerating the ionization states of drug-like small molecules. *Journal of Cheminformatics*, 11(1):14, 2019.

Gisbert Schneider and Uli Fechner. Lead- and drug-like compounds: the role of structural complexity and synthetic accessibility in drug discovery. *Nature Reviews Drug Discovery*, 9(12):949–962, 2010.

Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom L Blundell, Pietro Lio, et al. Structure-based drug design with equivariant diffusion models. *Nature Computational Science*, 4(12):899–909, 2024.

Arne Schneuing, Ilia Igashov, Adrian W Dobbelstein, Thomas Castiglione, Michael Bronstein, and Bruno Correia. Multi-domain distribution learning for de novo drug design. *arXiv preprint arXiv:2508.17815*, 2025.

Kenji Schorpp, Ina Rothenaigner, Elena Salmina, Jeanette Reinshagen, Terence Low, Jara K Brenke, Jay Gopalakrishnan, Igor V Tetko, Sheraz Gul, and Kamyar Hadian. Identification of small-molecule frequent hitters from alphascreen high-throughput screens. *Journal of biomolecular screening*, 19(5):715–726, 2014.

Ansgar Schuffenhauer, Nadine Schneider, Samuel Hintermann, Douglas Auld, Jutta Blank, Simona Cotesta, Caroline Engeloch, Nikolas Fechner, Christoph Gaul, Jerome Giovannoni, et al. Evolution of novartis' small molecule screening deck design. *Journal of medicinal chemistry*, 63(23): 14425–14447, 2020.

Jonathan Shearer, Jose L Castro, Alastair DG Lawson, Malcolm MacCoss, and Richard D Taylor. Rings in clinical trials and drugs: present and future. *Journal of medicinal chemistry*, 65(13): 8699–8712, 2022.

Iurii Sushko, Elena Salmina, Vladimir A Potemkin, Gennadiy Poda, and Igor V Tetko. Toxalerts: a web server of structural alerts for toxic chemicals and compounds with potential adverse reactions, 2012.

Amol Thakkar, Veronika Chadimová, Esben Jannik Bjerrum, Ola Engkvist, and Jean-Louis Reymond. Retrosynthetic accessibility score (rascore)–rapid machine learned synthesizability classification from ai driven retrosynthetic planning. *Chemical science*, 12(9):3339–3349, 2021.

Natasha Thorne, James Inglese, and Douglas S Auld. Illuminating insights into firefly luciferase and other bioluminescent reporters used in chemical biology. *Chemistry & biology*, 17(6):646–657, 2010.

Daniel F Veber, Stephen R Johnson, Hung-Yuan Cheng, Brian R Smith, Keith W Ward, and Kenneth D Kopple. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of medicinal chemistry*, 45(12):2615–2623, 2002.

Edon Vitaku, David T Smith, and Jon T Njardarson. Analysis of the structural diversity, substitution patterns, and frequency of nitrogen heterocycles among us fda approved pharmaceuticals: miniperspective. *Journal of medicinal chemistry*, 57(24):10257–10274, 2014.

Milan Voršilák, Michal Kolář, Ivan Čmelo, and Daniel Svozil. Syba: Bayesian estimation of synthetic accessibility of organic compounds. *Journal of cheminformatics*, 12(1):35, 2020.

Michael J Waring, John Arrowsmith, Andrew R Leach, Paul D Leeson, Sam Mandrell, Robert M Owen, Garry Pairaudeau, William D Pennie, Stephen D Pickett, Jibo Wang, et al. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature reviews Drug discovery*, 14(7):475–486, 2015.

Wendy A Warr, Marc C Nicklaus, Christos A Nicolaou, and Matthias Rarey. Exploration of ultralarge compound collections for drug discovery. *Journal of Chemical Information and Modeling*, 62(9):2021–2034, 2022.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

HW Whitlock. On the structure of total synthesis of complex natural products. *The Journal of Organic Chemistry*, 63(22):7982–7989, 1998.

Scott A Wildman and Gordon M Crippen. Prediction of physicochemical parameters by atomic contributions. *Journal of chemical information and computer sciences*, 39(5):868–873, 1999.

Jun Xu and James Stevenson. Drug-like index: a new approach to measure drug-like compounds and their diversity. *Journal of Chemical Information and Computer Sciences*, 40(5):1177–1187, 2000.

Odin Zhang, Jintu Zhang, Jieyu Jin, Xujun Zhang, RenLing Hu, Chao Shen, Hanqun Cao, Hongyan Du, Yu Kang, Yafeng Deng, et al. Resgen is a pocket-aware 3d molecular generation model based on parallel multiscale modelling. *Nature Machine Intelligence*, 5(9):1020–1030, 2023.

Alex Zhavoronkov, Yan A Ivanenkov, Alex Aliper, Mark S Veselov, Vladimir A Aladinskiy, Anastasiya V Aladinskaya, Victor A Terentiev, Daniil A Polykovskiy, Maksim D Kuznetsov, Arip Asadulaev, et al. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature biotechnology*, 37(9):1038–1040, 2019.

Huimin Zhu, Renyi Zhou, Dongsheng Cao, Jing Tang, and Min Li. A pharmacophore-guided deep learning approach for bioactive molecular generation. *Nature Communications*, 14(1):6234, 2023.

Yurong Zou, Tao Guo, Zhiyuan Fu, Zhongning Guo, Weichen Bo, Dengjie Yan, Qiantao Wang, Jun Zeng, Dingguo Xu, Taijin Wang, et al. A structure-based framework for selective inhibitor design and optimization. *Communications Biology*, 8(1):422, 2025.