

A MOLECULES PREPARATION

Unless otherwise stated we use RDKit (Landrum, 2013) (version 2025.03.4), Dimorphite-DL (Ropp et al., 2019) (version 2.0.2), and AiZynthFinder (Genheden et al., 2020) (version 4.4.0). To compare models faithfully, we use the preprocessing pipeline. Each model was asked to generate 10,000 unique and valid molecules, and validity was checked with RDKit. After that each molecule was standardized with RDKit and Dimorphite-DL, following these steps below.

```

from rdkit import Chem
from rdkit.Chem import AllChem
from rdkit.Chem.SaltRemover import SaltRemover
from rdkit.Chem.MolStandardize import tautomer, rdMolStandardize
from rdkit.Chem.EnumerateStereoisomers import EnumerateStereoisomers, StereoEnumerationOptions
from dimorphite_dl import protonate_smiles

mol = Chem.MolFromSmiles("CNCC/C=N/[C@H]1CCc2ccccc21")

# (i) Remove salts or solvents and keep largest organic fragment
remover = SaltRemover()
mol = remover.StripMol(mol)
mol = rdMolStandardize.LargestFragmentChooser(preferOrganic=True).choose(mol)

# (ii) Add hydrogens to complete valences
mol = Chem.AddHs(mol)

# (iii) Normalize valence, kekulize, and sanitize
mol = rdMolStandardize.Normalize(mol)
Chem.Kekulize(mol, clearAromaticFlags=True)
Chem.SanitizeMol(mol)

# (iv) Protomer enumeration at pH 7.4 ± 0.0
smiles = Chem.MolToSmiles(mol)
protomers = protonate_smiles(smiles, ph_min=7.4, ph_max=7.4, max_variants=8)

# (v) Preserve declared stereochemistry
# and, where unspecified, enumerate stereocenters under a cap
cap = 8
opts = StereoEnumerationOptions(onlyUnassigned=True, maxIsomers=cap)
stereo_variants = []
for p in protomers:
    group = []
    for t in p:
        Chem.AssignStereochemistry(t, cleanIt=True, force=True)
        group.extend(list(EnumerateStereoisomers(t, options=opts)))
    stereo_variants.append(group)

# (vi) Generate 3D conformers via distance geometry
# and minimize with a small-molecule force field,
# and retain the lowest-energy conformer per state
def lowestE_conf(m):
    params = AllChem.ETKDGv3()
    cids = AllChem.EmbedMultipleConfs(m, numConfs=20, params=params)
    mmff_props = AllChem.MMFFGetMoleculeProperties(m, mmffVariant='MMFF94')
    if mmff_props: results = AllChem.MMFFOptimizeMoleculeConfs(m)
    else: results = AllChem.UFFOptimizeMoleculeConfs(m)

    energies = [e for (_, e) in results]
    best_idx = min(range(len(energies)), key=lambda i: energies[i])
    best_cid = cids[best_idx]
    best = Chem.Mol(m)
    best.RemoveAllConformers()
    best.AddConformer(m.GetConformer(best_cid), assignId=True)
    return best, energies[best_idx]

# Select the lowest-energy conformer
final_states = []
for variants in stereo_variants:
    for iso in variants:
        iso = Chem.AddHs(iso, addCoords=True)
        best3d = lowestE_conf(iso)
        final_states.append(best3d)
final_states.sort(key=lambda x: x[1])

# Result
best_mol, best_energy = final_states[0]

```

B DETAILED MOLECULAR FLOW WITHIN FIVE-STAGE FILTERING PIPELINE

B.1 DESCRIPTORS DEFINITIONS AND THRESHOLDS

Unless otherwise stated we use RDKit (Landrum, 2013) (version 2025.03.4). This section extends Descriptors 3.1 in the main paper. All descriptors are computed after the ligand-preparation stage. We compute 18 physicochemical descriptors which are all two-dimensional with no 3D dependence

using RDKit and apply literature-based thresholds to remove chemically out-of-scope candidates while retaining structural diversity.

The descriptor panel includes: allowed elements ("Chars"), number of atoms (#atoms), number of heavy atoms (#heavy atoms), number of hetero atoms (#hetero atoms), number of nitrogen atoms (#Nitrogen atoms), molecular weight (MW), logP, ring size, number of rings (#rings), number of aromatic rings (#aroma rings), number of fused aromatic rings (#fused aroma rings), number of rigid bonds (#rigid bonds), number of rotatable bonds (#RotB), hydrogen bond donors (HBDs), hydrogen bond acceptors (HBAs), f_{sp^3} , topological polar surface area (TPSA), quantitative estimation of drug-likeness (QED).

Our bounds are based on established medicinal chemistry rule sets and surveys, reflecting realistic constraints. Mostly we have relied on: the Rule-of-Five (Lipinski et al., 1997), the Beyond Rule-of-Five (Doak et al., 2014), Rule-of-Xu (Xu & Stevenson, 2000), Oprea rule (Oprea et al., 2001), Veber's Rules (Veber et al., 2002), ZINC subset filters (Irwin & Shoichet 2005; Irwin et al., 2012), Generative Design choices based on a review of generative chemistry methods (Warr et al., 2022), QED thresholds based on main paper (Bickerton et al., 2012), and "Escape from Flatland" (Lovering et al., 2009). Descriptive surveys (Heravi & Zadsirjan, 2020; Shearer et al., 2022; Vitaku et al., 2014) are used for guiding toward realistic constraints on those descriptors, that are not presented in rule sets above. We combined and extended rule sets and surveys because (i) no single rule covers all descriptors we use, and (ii) our goal is a general, non-task-specific clean-up to avoid clearly out-of-scope chemistry rather than aggressive optimization for a specific ADME profile.

We restrict to the common small-molecule drug alphabet {C, H, N, O, S, F, Cl, Br}, following common ZINC database Filters (Irwin & Shoichet, 2005; Irwin et al., 2012) and surveys of approved drugs (Irwin et al., 2012; Vitaku et al., 2014; Heravi & Zadsirjan, 2020); heavy metals are excluded. Number of hetero atoms lower bound is set to 2, as this prevents trivial all-carbon hydrocarbons from passing the filter (although such molecules are synthetically valid, they are essentially outside the space of relevant chemotypes). We follow descriptive surveys reporting that nearly all oral drugs contain between 0 and 10 nitrogen atoms (Vitaku et al., 2014; Shearer et al., 2022). The original Oprea rule (Oprea et al., 2001) sets a lower bound of 2 for HBAs, but we set this to ≥ 3 to exclude hydrophobic candidates lacking sufficient polarity. Molecules with $f_{sp^3} \geq 0.25$ are known to exhibit improved three-dimensionality and associated with better synthesisability metrics, following the "Escape from Flatland" hypothesis (Lovering et al., 2009) we use a permissive lower bound of 0.15. Molecules with very low QED values (< 0.3) are strongly enriched for exotic, non-druglike structures (Bickerton et al., 2012).

```
from rdkit import Chem
from rdkit.Chem import Lipinski, rdMolDescriptors, QED, Descriptors
from medchem.rules._utils import n_fused_aromatic_rings

mol = Chem.MolFromSmiles(smiles)

chars = list(set(atom.GetSymbol() for atom in mol.GetAtoms() if atom.GetSymbol())) # --> ZINC Filters
n_atoms = Chem.AddHs(mol).GetNumAtoms() # --> ZINC Filters
n_heavy_atoms = rdMolDescriptors.CalcNumHeavyAtoms(mol) # --> Rule-of-Xu
n_het_atoms = sum(1 for atom in mol.GetAtoms() if atom.GetAtomicNum() not in (1, 6)) # --> Generative Design
n_n_atoms = sum(1 for atom in mol.GetAtoms() if atom.GetAtomicNum() == 7) # --> Descriptive surveys
molWt = Descriptors.MolWt(mol) # --> Beyond Rule-of-Five
logP = Descriptors.MolLogP(mol) # --> Beyond Rule-of-Five
rings_size = [len(x) for x in mol.GetRingInfo().AtomRings()] # upper bound --> ZINC Filters
n_rings = mol.GetRingInfo().NumRings() # --> extended ZINC Filters
n_aroma_rings = rdMolDescriptors.CalcNumAromaticRings(mol) # --> Generative Design
n_fused_aromatic_rings = n_fused_aromatic_rings(mol) # --> Generative Design
n_rigid_bonds = mol.GetNumBonds() - rdMolDescriptors.CalcNumRotatableBonds(mol) # --> Generative Design
n_rot_bonds = rdMolDescriptors.CalcNumRotatableBonds(mol) # --> Oprea Rule
hbds = Lipinski.NumHDonors(mol) # --> Rule-of-Five
hbas = Lipinski.NumHAcceptors(mol) # upper bound --> Generative Design, lower bound --> extended Rule of Oprea
fsp3 = rdMolDescriptors.CalcFractionCSP3(mol) # --> Extended "Escape from Flatland"
tpsa = rdMolDescriptors.CalcTPSA(mol) # upper bound --> Veber's Rules, lower bound --> Generative Design
qed = QED.qed(mol) # --> Unfavorable acc. to QED paper
```

Each cell in Table 5, Table 6 and Table 7 reports the number of generated molecules that satisfy the corresponding descriptor threshold (bold = best, underline = second best). "Pass" counts molecules meeting all thresholds simultaneously. Figure 3 shows the descriptor distributions for the three model families alongside those of known inhibitors.

While rule-based filtering is useful for triaging large generative samples, it is not definitive. Notably, some known KRAS G12D inhibitors fall outside predefined boundaries. We therefore treat molecule generation filtering as an early stage of exploration. Molecules that pass filters are good starting

points. During later stages of hit identification and lead optimization added rigidity, new polar groups, or increased molecular weight often improve potency and selectivity, even if such changes reduce formal drug-likeness by classical rules (Schneider & Fechner, 2010; Hughes et al., 2011). This reflects the inherent trade-offs of medicinal chemistry.

Notably, known inhibitors violate some of these rules, as bioactivity does not follow the strict compliance with rules (Doak et al., 2014).

Table 5: Comparison of unconditional molecular generators by descriptor pass rates

Metric	Threshold	E(3)DM	HIERGRAPHVAE	JT-VAE	MoLeR	MOLGPT	TGM-DLM
Initial		10000	10000	10000	10000	10000	10000
Chars	$[C, N, S, O, F, Cl, Br, H]$	9982	9879	9910	9716	9860	8178
#atoms	[10, 100]	9999	9961	9996	9978	9982	8430
#heavy atoms	[10, 50]	9658	9687	9963	9921	9993	8211
#hetero atoms	[2, 15]	9774	9757	9982	9902	9985	8182
#Nitrogen atoms	[0, 10]	9973	9998	10000	9999	10000	9979
MW	[100, 1000]	9938	9954	9992	9992	10000	9394
logP	[-2, 10]	9690	9974	9985	9982	10000	7986
ring size	[3, 12]	9523	10000	10000	9995	9984	9884
#rings	[0, 8]	9983	9971	9997	9887	9872	9875
#aroma rings	[0, 5]	10000	9979	10000	9925	9997	9994
#fused aroma rings	[0, 2]	10000	10000	10000	9998	10000	10000
#rigid bonds	[0, 30]	9687	9494	9968	8299	8057	8477
#RotB	[2, 8]	6848	8674	9836	7260	8191	5023
HBDs	≤ 5	9968	9975	9992	9913	10000	7978
HBAs	[3, 12]	8549	7047	9043	8821	8914	6395
f_{sp^3}	[0.15, 0.8]	8145	7199	9278	7611	7790	5977
TPSA	[40, 140]	7799	7414	8700	8089	8118	5550
QED	[0.3, 1]	8355	9530	9978	8231	9041	5975
Pass Stage 1		3520	3579	7586	3193	3474	1216

Table 6: Comparison of ligand-based molecular generators by descriptor pass rates

Metric	Threshold	GCPG	GENTRL	MOLFINDER	PGMG	REINVENT4 (V)	REINVENT4 (P)	REINVENT4 (TL)
Initial		10000	10000	10000	10000	10000	10000	10000
Chars	$[C, N, S, O, F, Cl, Br, H]$	10000	9999	9469	10000	10000	10000	9999
#atoms	[10, 100]	10000	10000	9995	9972	9953	10000	9914
#heavy atoms	[10, 50]	10000	10000	9930	9945	9933	10000	9717
#hetero atoms	[2, 15]	10000	10000	9742	9984	9917	9996	9910
#Nitrogen atoms	[0, 10]	10000	10000	9998	9870	9997	9990	9979
MW	[100, 1000]	10000	10000	9983	9998	9996	10000	9988
logP	[-2, 10]	10000	10000	9999	9958	9965	10000	9999
ring size	[3, 12]	10000	9980	9972	10000	10000	10000	9953
#rings	[0, 8]	9670	10000	9221	7926	9872	7110	3703
#aroma rings	[0, 5]	9999	10000	9998	5004	9949	9922	9821
#fused aroma rings	[0, 2]	10000	10000	9998	8618	9997	9943	9884
#rigid bonds	[0, 30]	6672	9988	8528	833	8291	988	2198
#RotB	[2, 8]	10000	6367	9744	9562	8136	9793	9079
HBDs	≤ 5	10000	10000	10000	8988	9900	10000	9987
HBAs	[3, 12]	9980	9950	5562	9725	8999	9995	9600
f_{sp^3}	[0.15, 0.8]	9999	9992	8987	3371	7667	9949	9669
TPSA	[40, 140]	9965	9286	3643	8230	8317	9927	9356
QED	[0.3, 1]	10000	9838	9965	1951	8829	9384	7519
Pass Stage 1		6616	5669	1592	195	4089	936	1204

Table 7: Comparison of protein-based molecular generators by descriptor pass rates

Metric	Threshold	DIFFSBDD	DRAGONFLY	DRAGONFLY (B)	DRUGFLOW	POCKET2MOL	PROTOBIND-DIFF	RESGEN	TARGETDIFF
Initial		10000	10000	10000	10000	10000	10000	10000	10000
Chars	$[C, N, S, O, F, Cl, Br, H]$	9618	9993	10000	9308	9926	9991	9753	9707
#atoms	[10, 100]	9961	9937	10000	9929	9999	8023	10000	9952
#heavy atoms	[10, 50]	9952	9947	10000	9997	9977	7913	9812	9939
#hetero atoms	[2, 15]	9868	9894	10000	9992	9722	8537	9610	9725
#Nitrogen atoms	[0, 10]	10000	9989	9998	9999	10000	9817	9991	9998
MW	[100, 1000]	9996	10000	10000	10000	9992	8758	9984	10000
logP	[-2, 10]	9934	9949	10000	9997	9935	9895	9970	9623
ring size	[3, 12]	9872	9995	9989	9914	9610	8645	9485	9178
#rings	[0, 8]	9863	9855	8734	9889	8849	7953	7655	9763
#aroma rings	[0, 5]	10000	9928	9983	9996	9981	9892	9327	10000
#fused aroma rings	[0, 2]	10000	10000	9994	9999	10000	9995	9991	10000
#rigid bonds	[0, 30]	9054	5933	1069	8440	6175	2511	2731	6917
#RotB	[2, 8]	6974	7675	9932	9343	8094	7307	6853	7657
HBDs	≤ 5	9556	9923	10000	9931	9844	9763	9185	8062
HBAs	[3, 12]	9079	9008	9991	9727	7766	8517	9115	9526
f_{sp^3}	[0.15, 0.8]	7308	8465	9991	8390	8659	9672	8081	9048
TPSA	[40, 140]	8108	8670	9981	9343	7557	7689	6588	7006
QED	[0.3, 1]	7806	7779	9946	9564	8492	6444	4958	7249
Pass Stage 1		3665	2779	1022	5464	2657	1466	1080	3444

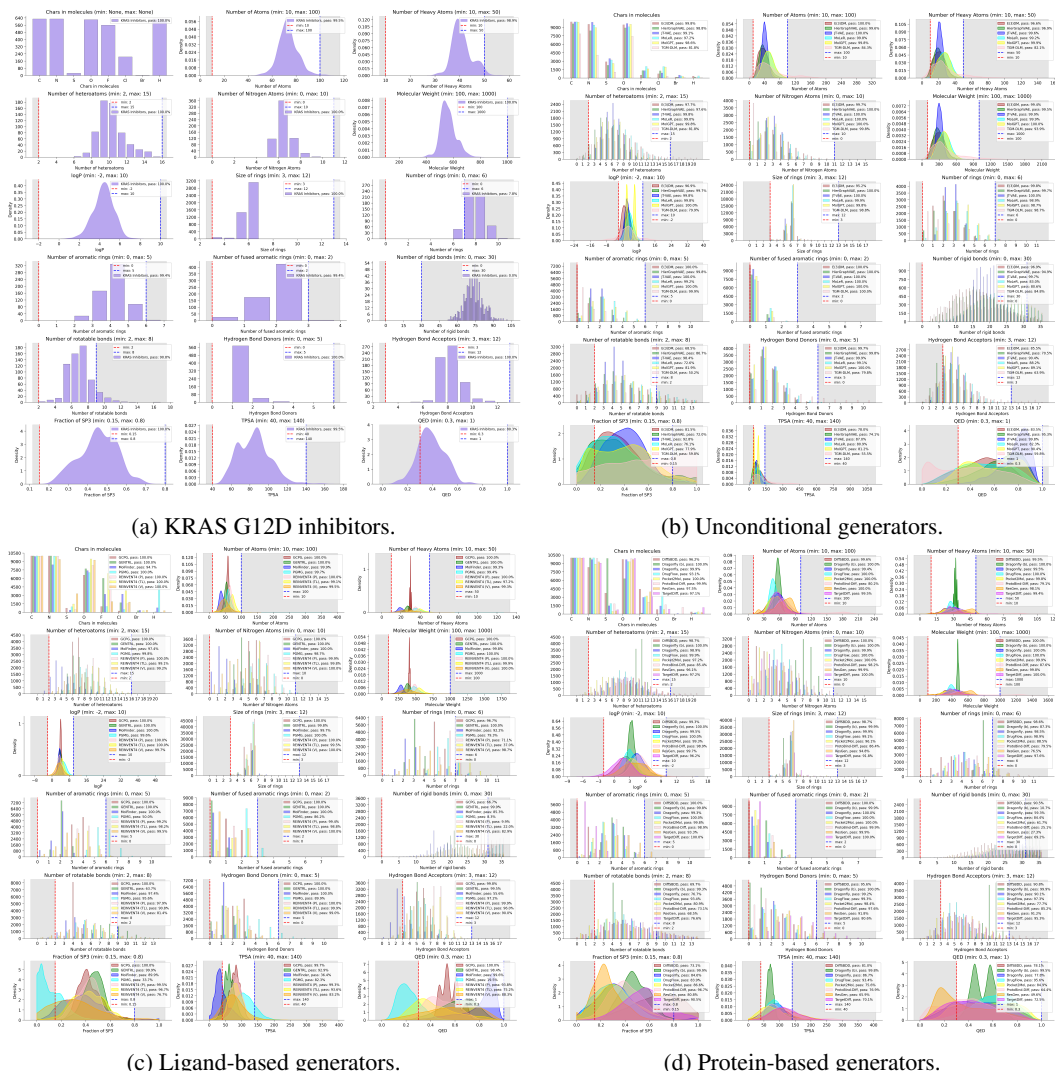


Figure 3: Physicochemical descriptor distributions for KRAS G12D inhibitors and generated molecules. Each panel shows the 18 descriptors used in Stage 1 filtering, comparing (a) known inhibitors, (b) unconditional generators, (c) ligand-based generators, and (d) protein-based generators. Dashed lines mark filtering thresholds, highlighting chemical space coverage and pass rates of different model families.

B.2 STRUCTURE FILTERS DEFINITIONS AND THRESHOLDS

Unless otherwise stated we use MedChem (Emmanuel et al., 2025) 2.0.5. This section extends Structural Filters 3.2 in the main paper. All structural filters are computed after the descriptors stage on molecules that pass all descriptors simultaneously. We compute 5 structural rule sets using MedChem for the sake of removing those molecules that are toxic, reactivate, unstable, or harmful for a human.

- **CommonAlerts:** Aggregates public structural alert sets using MedChem; we use a subset of 15 alert sets to flag motifs linked to assay interference or liabilities; implemented in `medchem.structural.CommonAlertsFilters`. The following structural alert sets are used: University of Dundee NTD Screening Library Filters (Brenk et al., 2008), Bristol-Myers Squibb (BMS) HTS Deck Filters (Pearce et al., 2006), Inpharmatica (Emmanuel et al., 2025), LD50-Oral (Sushko et al., 2012), Glaxo Wellcome Hard Filters (Hann et al., 1999), Pan Assay Interference Compounds (PAINS) Filters (Baell & Hol-

loway, 2010), AlphaScreen frequent Hitters (Schorpp et al., 2014), Frequent-Hitter (Dahlin et al., 2021), Capuzzi et al., 2017, Dahlin et al., 2015), Chelator (Aldrich et al., 2017), SureChEMBL (Papadatos et al., 2016), NIH MLSMR Excluded Functionality Filters (Jadhav et al., 2010), Pfizer LINT filters (Blake, 2005), GST-Hitters (Schorpp et al., 2014), HIS-Hitters (Schorpp et al., 2014), LuciferaseInhibitor (Auld et al., 2008; Thorne et al., 2010).

- **MolGraphStatistics**: Removes graph-theoretic instabilities frequently produced by generative models (e.g., atoms participating in multiple 3–4 member rings); implemented in `medchem.functional.molecular_graph_filter`.
- **MolComplexity**: Excludes outliers by comparing complexity metrics with percentile thresholds computed on ZINC-15 reference set. The following rule sets are used: Bertz (Bertz, 1981), Whitlock (Whitlock, 1998), Barone (Barone & Chanon, 2001), Synthetic & Molecular Complexity (SMCM) (Allu & Oprea, 2005), Total Walk Count complexity (TWC) (Gutman et al., 2001), SAS (Ertl & Schuffenhauer, 2009), QED (Bickerton et al., 2012), cLogP (Wildman & Crippen, 1999); implemented in `medchem.complexity.ComplexityFilter`.
- **NIBR** (Novartis hit-triage substructure filters (Schuffenhauer et al., 2020)): Applies filters which annotates matches with severity and flag potential covalent motifs or special chemotypes (e.g., nitro counts, polyhalogenated aromatics, phenol esters); implemented in `medchem.structural.NIBRFilters`.
- **Bredt** (anti-Bredt bridgehead alkenes (Fawcett, 1950)): Flags alkenes—double bonds at bridgeheads in small bridged rings—indicative of unstable, non-planar π systems; implemented in `medchem.functional.bredt_filter`.

Table 8: Comparison of unconditional molecular generators by structural filters pass rates

Rule set /Model	E(3)DM	HIERGRAPHVAE	JT-VAE	MoLER	MOLGPT	TGM-DLM
Pass Stage 1	3520	3579	7586	3193	3474	1216
Common Alerts	170	1212	3020	748	1114	119
MolGraph statistics	3491	3579	7584	3193	3468	1213
MolComplexity	1130	3512	6955	3084	3212	1009
NIBR	2226	3438	7333	2991	3325	945
Bredt	3329	3574	7585	3113	3465	1198
Pass Stage 2	75	1176	2765	718	1029	100

Table 9: Comparison of ligand-based molecular generators by structural filters pass rates

Rule set /Model	GCPG	GENTRL	MOLFINDER	PGMG	REINVENT4 (V)	REINVENT4 (P)	REINVENT4 (TL)
Pass Stage 1	6616	5669	1592	195	4089	936	1204
Common Alerts	4751	2001	509	49	1406	707	439
MolGraph statistics	6591	5668	1558	195	4088	936	1204
MolComplexity	5832	5432	1046	134	3876	780	1153
NIBR	6596	5448	1363	182	3923	936	1043
Bredt	6616	5662	1529	185	4086	926	1203
Pass Stage 2	4168	1925	366	37	1325	593	413

Table 10: Comparison of protein-based molecular generators by structural filters pass rates

Rule set /Model	DIFFSBDD	DRAGONFLY	DRAGONFLY (B)	DRUGFLOW	POCKET2MOL	PROTOBIND-DIFF	RESGEN	TARGETDIFF
Pass Stage 1	3665	2779	1022	5464	2657	1466	1080	3444
Common Alerts	335	1501	321	1557	800	205	275	445
MolGraph statistics	3530	2779	1020	5442	2657	1459	1080	3444
MolComplexity	1524	2698	960	4977	2273	1423	995	961
NIBR	2871	2766	453	5141	2457	1446	889	3043
Bredt	3403	2778	1019	5449	2638	1463	1063	3227
Pass Stage 2	197	1459	218	1392	682	195	255	136

B.3 SYNTHESIS FEASIBILITY DEFINITIONS AND THRESHOLDS

The Synthetic Accessibility score (SA score) (Ertl & Schuffenhauer, 2009) is a heuristic fragment and complexity-based score that ranges molecules from 1 (easy) to 10 (hard). The original paper

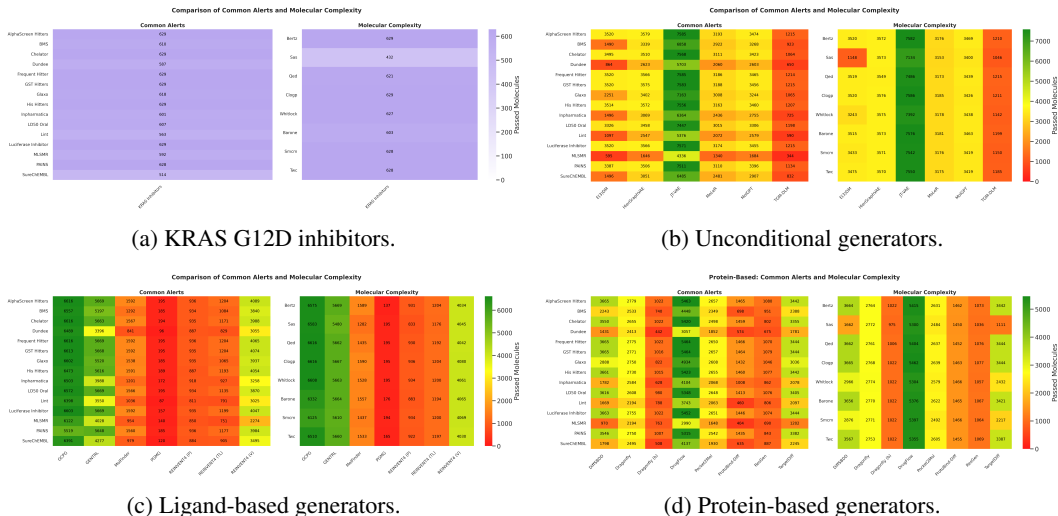


Figure 4: Passing rates of Structural Filters across the three model families and known KRAS G12D inhibitors. Each panel shows rates of molecules passed each rule subset in composition filters Common Alerts and MolComplexity of different generative approaches.

suggested SA score cutoff of 6.0; later comparative studies show the nominal cutoff can be permissive and that stricter boundary better match retrosynthesis-aware classifiers (Voršilák et al., 2020; Chen & Jung, 2024). Consequently, we set a threshold to ≤ 4.5 for SA score.

The Retrosynthetic Accessibility score (RA score) (Thakkar et al., 2021) is a machine-learned classifier that predicts whether an automated retrosynthesis planner is able to find a synthetic route for a molecule; it outputs a probability-like score. We use the natural decision cutoff of > 0.5 , i.e., the model predicts that a retrosynthetic route is likely to be found.

Synthetic Bayesian Accessibility (SYBA) (Voršilák et al., 2020) is a fragment-based Bernoulli Naive Bayes classifier that assigns a continuous SYBA score. By construction, a positive SYBA score indicates the model classifies the compound as easy to synthesize, while a negative score indicates hard to synthesize molecule. With that knowledge we set SYBA threshold to > 0 .

We configured AiZynthFinder (Genheden et al., 2020) (version 4.3.2) for synthesis planning to use the Monte Carlo Tree Search (MCTS) algorithm, the default template-based single-step retrosynthesis model included in the software, and the default ZINC-based set of building blocks. We chose the exploration constant (C) for MCTS to be 1.4, enabled priors with a default value of 0.5, activated cycle pruning to avoid redundant branches, and used the state score as the reward function. We constrained the search to a maximum depth of five transformations, no more than 10,000 algorithm iterations, and no more than 300 seconds for building the retrosynthesis tree. We chose such a small time limit to emulate the setting of high-throughput synthesizability screening in which the filtering of generated molecules should be fast. To prevent trivial solutions, we excluded the target molecule from the stock.

A molecule passes the Synthesis Feasibility stage only if it satisfies all four synthesizability thresholds simultaneously.

Table 11: Comparison of unconditional molecular generators by synthesis feasibility pass rates

Metric /Model	Threshold	E(3)DM	HIERGRAPHVAE	JT-VAE	MoLER	MOLGPT	TGM-DLM
Pass Stage 2		75	1176	2765	718	1029	100
SA score	≤ 4.5	59	1174	2733	718	1026	100
RA score	$(0.5, 1]$	66	1149	2679	701	911	91
SYBA score	$(0, \infty)$	36	1153	2479	704	944	65
AiZynthFinder	<i>routes</i> > 0	0	987	1564	564	693	43
Pass Stage 3		0	975	1549	557	679	35

Table 12: Comparison of unconditional molecular generators by synthesis feasibility pass rates

Metric / Model	Threshold	GCPG	GENTRL	MOLFINDER	PGMG	REINVENT4 (V)	REINVENT4 (P)	REINVENT4 (TL)
Pass Stage 2		4168	1925	366	37	1325	593	413
SA score	≤ 4.5	4116	1912	366	37	1323	511	404
RA score	$(0.5, 1]$	3288	1792	363	24	1194	562	407
SYBA score	$(0, \infty)$	3087	1665	326	37	1293	466	377
AIZynthFinder	routes > 0	1312	318	268	24	940	299	285
Pass Stage 3		1064	303	265	22	918	222	276

Table 13: Comparison of unconditional molecular generators by synthesis feasibility pass rates

Metric / Model	Threshold	DIFFSBDD	DRAGONFLY	DRAGONFLY (B)	DRUGFLOW	POCKET2MOL	PROTOBIND-DIFF	RESGEN	TARGETDIFF
Pass Stage 2		197	1459	218	1392	682	195	255	136
SA score	≤ 4.5	174	1459	217	1375	675	192	249	122
RA score	$(0.5, 1]$	122	1391	198	1188	535	185	211	68
SYBA score	$(0, \infty)$	99	1453	214	1202	508	192	189	42
AIZynthFinder	routes > 0	26	1220	38	466	151	103	67	6
Pass Stage 3		24	1207	38	453	137	102	62	4

Figure 5 shows synthesis paths found with AiZynthFinder for top three models. This figure extends Figure 2 from the main paper and shows the other 6 synthesis paths for unconditional, ligand-based and protein-based models.

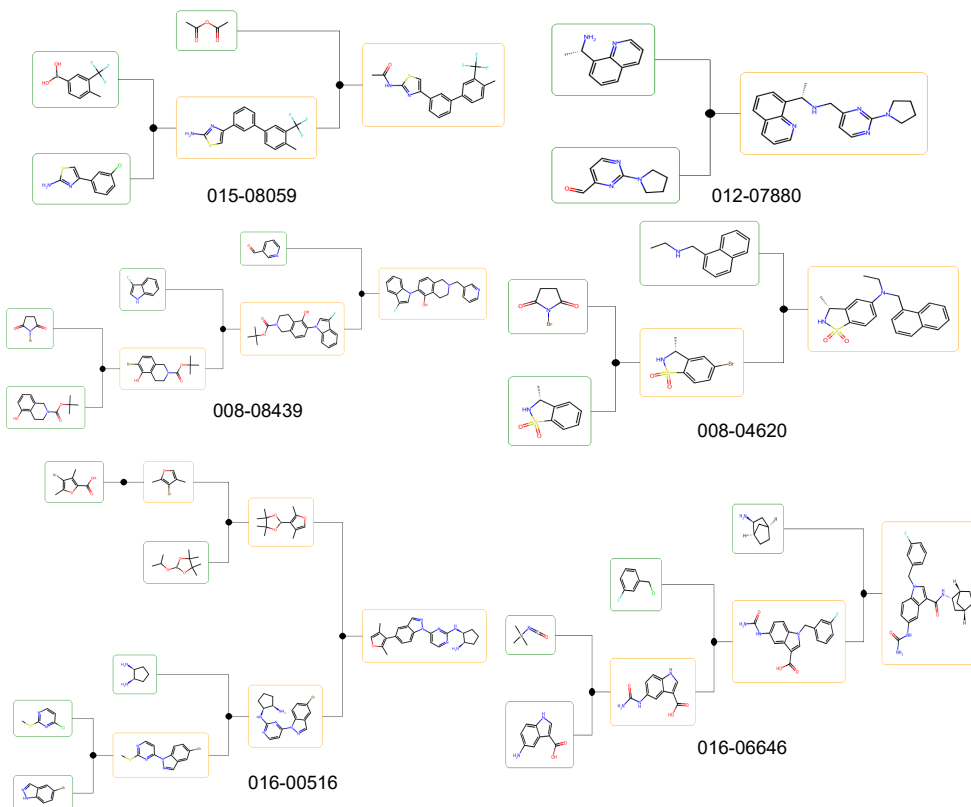


Figure 5: Top generated molecules among three families. Top: unconditional generators (015 - MolGPT, 012 - JT-VAE), middle: protein-based generators (008 - DrugFlow), bottom: ligand-based generators (016 - GCPG).

B.4 DOCKING SCORE AND BINDING AFFINITY DEFINITIONS AND THRESHOLDS

We use *smina* (Koes et al., 2013) with box of size 10, centered at $[-1.258, -12.520, 11.390]$, *exhaustiveness* = 8, *num_modes* = 9, *energy_range* = 3; and GNINA (McNutt et al., 2021) with autobox centered on ligand TH-Z816, *exhaustiveness* = 8, *num_modes* = 9.

We evaluate both *smina* and GNINA as *smina* is an empirical scoring and GNINA is a CNN-based ML scoring, using different scoring philosophies and thus different failure modes. A molecule that scores well in both engines is less likely to be a tool-specific false positive sample. We report intersection and union of the two resulted sets of molecules. Intersection gives high precision, and turns into a smaller set for costly follow-ups, and is good for manual medicinal chemists review and as a final shortlist.

Union with higher recall is useful for exploratory analyses and not discarding potentially good candidates too early.

Table 14: Comparison of unconditional molecular generators by docking and binding affinity estimation pass rates

Tool /Model	E(3)DM	HIERGRAPHVAE	JT-VAE	MoLER	MOLGPT	TGM-DLM
Pass Stage 3	0	975	1549	557	679	35
<i>smina</i>	0	906	1529	510	608	29
GNINA	0	618	1083	380	463	15
Boltz-2	0	742	1170	502	573	24
Pass Stage 4	0	477	816	323	340	10

Table 15: Comparison of ligand-based molecular generators by docking and binding affinity estimation pass rates

Tool /Model	GCPG	GENTRL	MOLFINDER	PGMG	REINVENT4 (V)	REINVENT4 (P)	REINVENT4 (TL)
Pass Stage 3	1064	303	265	22	918	222	276
<i>smina</i>	978	303	265	22	854	207	253
GNINA	692	245	214	19	587	76	178
Boltz-2	1064	293	246	22	844	210	263
Pass Stage 4	648	238	200	19	518	72	164

Table 16: Comparison of protein-based molecular generators by docking and binding affinity estimation pass rates

Tool /Model	DIFFSBDD	DRAGONFLY	DRAGONFLY (B)	DRUGFLOW	POCKET2MOL	PROTOBIND-DIFF	RESGEN	TARGETDIFF
Pass Stage 3	24	1207	38	453	137	102	62	4
<i>smina</i>	24	884	22	448	137	96	55	4
GNINA	20	767	22	395	95	70	42	4
Boltz-2	13	1191	38	399	94	99	46	0
Pass Stage 4	13	575	15	344	69	66	37	0

C MODELS EVALUATION DETAILS

C.1 INPUT DATA

The code is available at <https://anonymous.4open.science/r/MolGenBenchmark-F185/README.md>.

For conditional generation we have executed experiments on KRAS with the G12D oncogenic mutation - glycine-12 mutated to negatively charged aspartate. This is exploited by the design of TH-Z816 (switch-II pocket): the inhibitor has a piperazine moiety that forms a salt-bridge interaction with Asp12. Unlike G12C, which has a cysteine that can undergo covalent modification, G12D is harder to target. However, while generating molecules, the negative charge of Asp12 may be exploited by placing a positively charged group on the ligand. That gives specificity and a handle for binding.

For ligand-based models we have used a set of KRAS inhibitors provided by Ghazi Vakili et al. (2025). The initial set contains 645 molecules. We have checked the validity of molecules using RDKit (Landrum, 2013), and found 13 molecules with invalid structures (valence violation, prohibited symbols, etc). We filtered them out, leaving with 632 unique and valid molecules. Their statistics can be found in Figure 3a and Figure 4a. The structure binded with TH-Z816 is shown in Figure 6a and overall statistics on known inhibitors is shown in Figure 6b.

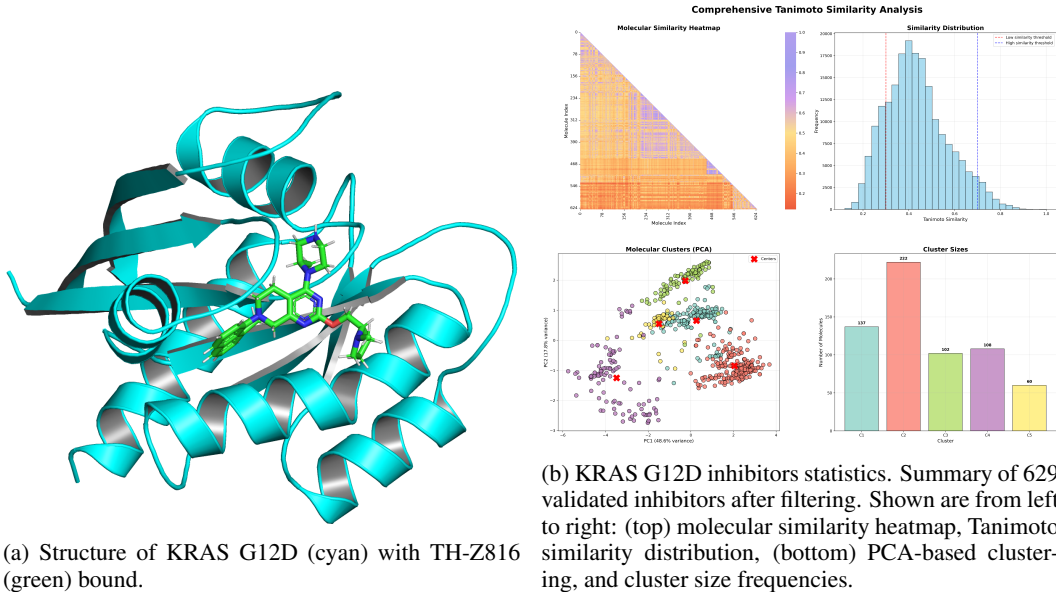


Figure 6: KRAS G12D binding with TH-Z816, and inhibitor set statistics.

C.2 REINVENT4

We intentionally evaluated three REINVENT4 variants (vanilla, author-provided prior, and a KRAS-fine-tuned prior) to isolate how prior choice and transfer-learning within autoregressive SMILES models affect downstream utilities, as diversity, synthesizability, and docking. REINVENT4 was chosen for this ablation for three practical reasons: (i) it is a widely used, modular autoregressive framework with well-documented fine-tuning pipelines and accessible checkpoints, enabling controlled prior and fine-tuning experiments; (ii) prior and fine-tuning effect is known to be especially strong and therefore worth probing; and (iii) extending equivalent, rigorously comparable fine-tuning pipelines to every model family (graph VAEs, 3D diffusion, flow models) would require substantial architecture-specific engineering and hyperparameter sweeps beyond the scope of this benchmark. We therefore treat the three REINVENT4 variants as a sensitivity study of the autoregressive prior — not as an exhaustive evaluation of all possible transfer-learning strategies across architectures. All priors are listed on Löffler (2025), and open-source.

REINVENT4 (V) (Vanilla, checkpoint: `reinvent.prior`). The unmodified REINVENT4 prior provided by the authors. Prior was trained on ChEMBL data (Gaulton et al., 2012); sampling used the following settings: `temperature = 1.0`, `sample_strategy="beamsearch"`, `num_smiles = 500`; no additional constraints applied. We sampled 10,000 SMILES per run; `num_smiles` argument is the number of smiles to be sampled per 1 input SMILES. We input 629 KRAS G12D inhibitors to that prior, and got 49,119 SMILES utilizing one A100 40GB GPU in less than 8 hours, highlighting that though REINVENT4 (V) is fast but results in a small number of valid molecules ($49,119 / (629 \times 500) \times 100\% = 15.618\%$).

REINVENT4 (P) (Provided prior, checkpoint: `mol2mol_medium_similarity.prior`). The REINVENT4 prior checkpoint provided by the authors that was trained with an explicit medium-Tanimoto similarity objective (author config: `similarity = 0.7`). Mol2Mol family priors were trained on PubChem data (Kim et al., 2023). In sake of fair comparison, we filtered known KRAS inhibitors with following strategy: if a molecule is present in Pubchem, we remove it. After filtering we resulted with 583 molecules, that were passed as input to REINVENT4 (P). Sampling used the

following settings: *temperature* = 1.0, *sample_strategy*="beamsearch", *num_smiles* = 500. We sampled 10,000 SMILES per run; as a result, we got 126,001 SMILES less than 8 hours with 43.225% of valid molecules ($126,001 / (583 \times 500) \times 100\% = 43.25\%$).

REINVENT4 (TL) (Transfer-learned). We fine-tuned a REINVENT4 prior (*reinvent.prior*) on 583 known KRAS G12D inhibitors with Mol2Mol training pipeline. We defined a similarity threshold of 0.7, and trained on 1000 epochs for 72 hours utilizing one A100 40GB GPU. For evaluation, we picked 7 checkpoints from the training process: [3, 5, 10, 100, 200, 500, 1000]. We sampled 10,000 SMILES per run for each checkpoint, and calculated basic descriptors. Figure 7 shows the distribution of the descriptors for each checkpoint.

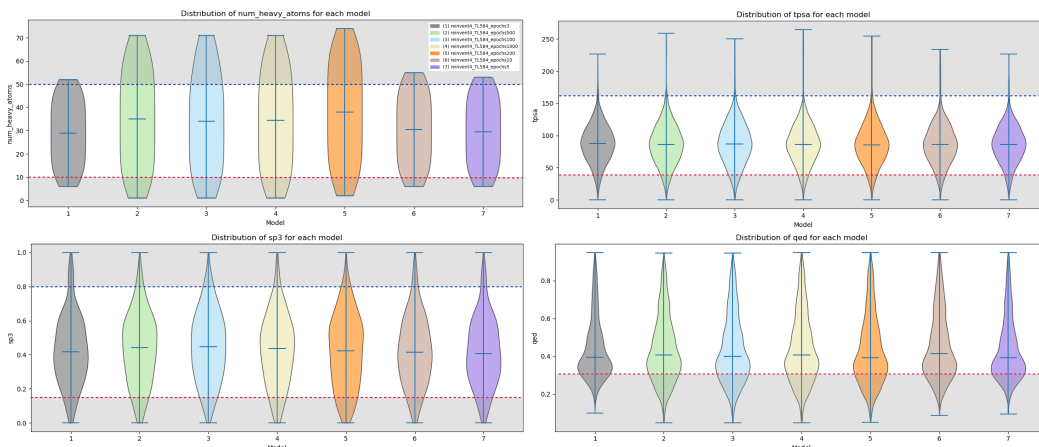


Figure 7: Distribution of basic descriptors for REINVENT4 (TL) checkpoints fine-tuned on 583 known KRAS G12D inhibitors.