Supplementary Information

GraphXForm: Graph transformer for computer-aided molecular design

Jonathan Pirnay,[a,b,‡] Jan G. Rittig,[c,‡] Alexander B. Wolf,[d,‡]
Martin Grohe,[e] Jakob Burger[d], Alexander Mitsos[f,c,g] and Dominik G. Grimm[a,b,*]

[a] *Technical University of Munich, TUM Campus Straubing for Biotechnology and Sustainability, Bioinformatics, Petersgasse 18, 94315 Straubing, Germany*
[b] *University of Applied Sciences Weihenstephan-Triesdorf, Bioinformatics, Petersgasse 18, 94315 Straubing, Germany*
[c] *Process Systems Engineering (AVT.SVT), RWTH Aachen University, Forckenbeckstraße 51, 52074 Aachen, Germany*
[d] *Technical University of Munich, TUM Campus Straubing for Biotechnology and Sustainability, Laboratory of Chemical Process Engineering, Uferstraße 53, 94315 Straubing, Germany*
[e] *Chair of Computer Science 7, RWTH Aachen University, Ahornstraße 55, 52074 Aachen, Germany*
[f] *JARA Center for Simulation and Data Science (CSD), Aachen, Germany*
[g] *Forschungszentrum Jülich GmbH, Institute of Climate and Energy Systems ICE-1: Energy Systems Engineering, Jülich 52425, Germany*
‡ *These authors contributed equally to this work.*
∗ *Corresponding author: dominik.grimm@hswt.de*

1 Atom alphabet

We list the atom alphabet for the different design tasks as a Python dictionary. We also provide the valence of an atom, as well as its `formal_charge` and `chiral_tag` – for atom specification in RDKit – if needed.

1.1 Drug design tasks

```
{
    "C": {"atomic_number": 6, "valence": 4},
    "C-": {"atomic_number": 6, "valence": 3, "formal_charge": -1},
    "C+": {"atomic_number": 6, "valence": 5, "formal_charge": 1},
    "C@": {"atomic_number": 6, "valence": 4, "chiral_tag": CHI_TETRAHEDRAL_CW},
    "C@@": {"atomic_number": 6, "valence": 4, "chiral_tag": CHI_TETRAHEDRAL_CCW},
    "N": {"atomic_number": 7, "valence": 3},
    "N-": {"atomic_number": 7, "valence": 2, "formal_charge": -1},
    "N+": {"atomic_number": 7, "valence": 4, "formal_charge": 1},
    "O": {"atomic_number": 8, "valence": 2},
    "O-": {"atomic_number": 8, "valence": 1, "formal_charge": -1},
    "O+": {"atomic_number": 8, "valence": 3, "formal_charge": 1},
    "F": {"atomic_number": 9, "valence": 1},
    "P": {"atomic_number": 15, "valence": 7},
    "P-": {"atomic_number": 15, "valence": 6, "formal_charge": -1},
    "P+": {"atomic_number": 15, "valence": 8, "formal_charge": 1},
    "S": {"atomic_number": 16, "valence": 6},
    "S-": {"atomic_number": 16, "valence": 5, "formal_charge": -1},
```

```
    "S+": {"atomic_number": 16, "valence": 7, "formal_charge": 1},
    "S@": {"atomic_number": 16, "valence": 6, "chiral_tag": CHI_TETRAHEDRAL_CW},
    "S@@": {"atomic_number": 16, "valence": 6, "chiral_tag": CHI_TETRAHEDRAL_CCW},
    "Cl": {"atomic_number": 17, "valence": 1},
    "Br": {"atomic_number": 35, "valence": 1},
    "I": {"atomic_number": 53, "valence": 1}
}
```

1.2 Solvent design tasks

```
{
    "C": {"atomic_number": 6, "valence": 4},
    "N": {"atomic_number": 7, "valence": 3},
    "O": {"atomic_number": 8, "valence": 2}
}
```

2 Drug design results

Here, we present the complete results for the 20 goal-directed tasks of the GuacaMol benchmark. To put our results into even further perspective, we included additional baseline methods not mentioned in the main paper.

Supplementary Table 1: GuacaMol goal-directed benchmarks

Benchmark	**SMILES GA‡** [1]	**SMILES LSTM‡** [2]	**Graph GA‡** [3]	**Reinvent¶** [4]	**GEGL¶** [5]	**CReM** [6]	**MolRL-MGPT** [7]	**GraphXForm** (ours)
Celecoxib rediscovery	0.732	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Troglitazone rediscovery	0.515	**1.000**	**1.000**	**1.000**	0.552	**1.000**	**1.000**	**1.000**
Thiothixene rediscovery	0.598	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Aripiprazole similarity	0.834	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Albuterol similarity	0.907	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Mestranol similarity	0.790	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
C11H24 *	0.829	0.993	0.971	0.999	**1.000**	0.966	**1.000**	**1.000**
C9H10N2O2PF2Cl †	0.889	0.879	0.982	0.877	**1.000**	0.940	0.939	**1.000**
Median molecules 1	0.334	0.438	0.406	0.434	0.455	0.371	0.449	**0.459**
Median molecules 2	0.380	0.422	0.432	0.395	**0.437**	0.434	0.422	0.403
Osimertinib MPO	0.886	0.907	0.953	0.889	**1.000**	0.995	0.977	0.972
Fexofenadine MPO	0.931	0.959	0.998	**1.000**	**1.000**	**1.000**	**1.000**	0.999
Ranolazine MPO	0.881	0.855	0.920	0.895	0.933	**0.969**	0.939	0.944
Perindopril MPO	0.661	0.808	0.792	0.764	0.833	0.815	0.810	**0.835**
Amlodipine MPO	0.722	0.894	0.894	0.888	0.905	0.902	0.906	**0.923**
Sitagliptin MPO	0.689	0.545	0.891	0.539	0.749	0.763	0.823	**0.965**
Zaleplon MPO	0.413	0.669	0.754	0.590	0.763	0.770	**0.790**	0.727
Valsartan SMARTS	0.552	0.978	0.990	0.095	**1.000**	0.994	0.997	**1.000**
Deco hop	0.970	0.996	**1.000**	0.994	**1.000**	**1.000**	**1.000**	**1.000**
Scaffold hop	0.885	0.998	**1.000**	0.990	**1.000**	**1.000**	**1.000**	**1.000**
Total score	14.398	17.341	17.983	16.349	17.627	17.919	18.052	**18.227**

* Top-159, no stereochemistry. † Top-250, no stereochemistry. ‡ Results as reported in [8]. ¶ Results as reported in [7].

[1] Naruki Yoshikawa, Kei Terayama, Masato Sumita, Teruki Homma, Kenta Oono, and Koji Tsuda. Population-based de novo molecule generation, using grammatical evolution. *Chemistry Letters*, 47(11):1431–1434, 2018.

[2] Marwin HS Segler, Thierry Kogej, Christian Tyrchan, and Mark P Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science*, 4(1):120–131, 2018.

[3] Jan H Jensen. A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space. *Chemical science*, 10(12):3567–3572, 2019.

[4] Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9, 2017.

[5] Sungsoo Ahn, Junsu Kim, Hankook Lee, and Jinwoo Shin. Guiding deep molecular optimization with genetic exploration. *Advances in Neural Information Processing Systems*, 33, 12008–12021, 2020.

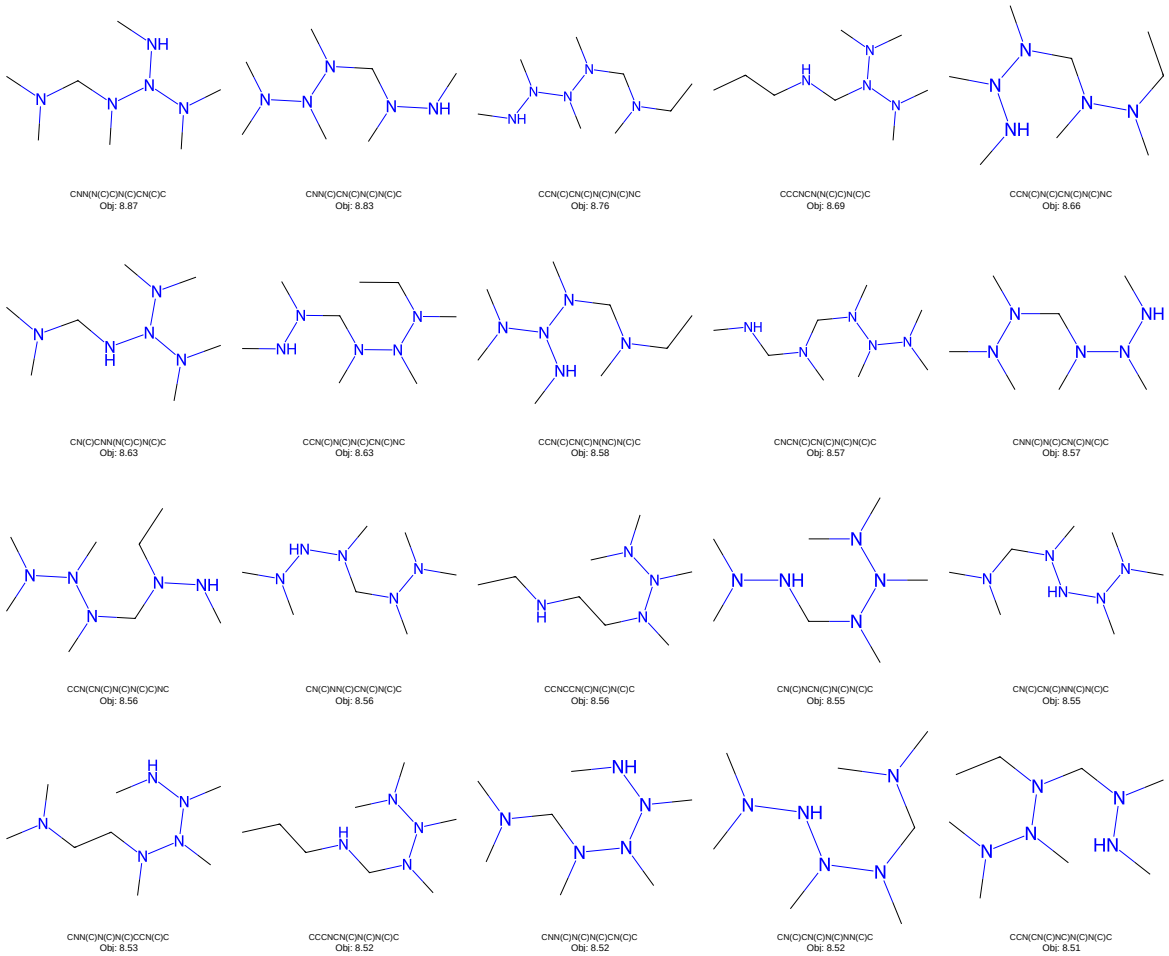
[6] Pavel Polishchuk. CReM: chemically reasonable mutations framework for structure generation. *Journal of Cheminformatics*, 12, 2020.

[7] Xiuyuan Hu, Guoqing Liu, Yang Zhao, and Hao Zhang. De novo drug design using reinforcement learning

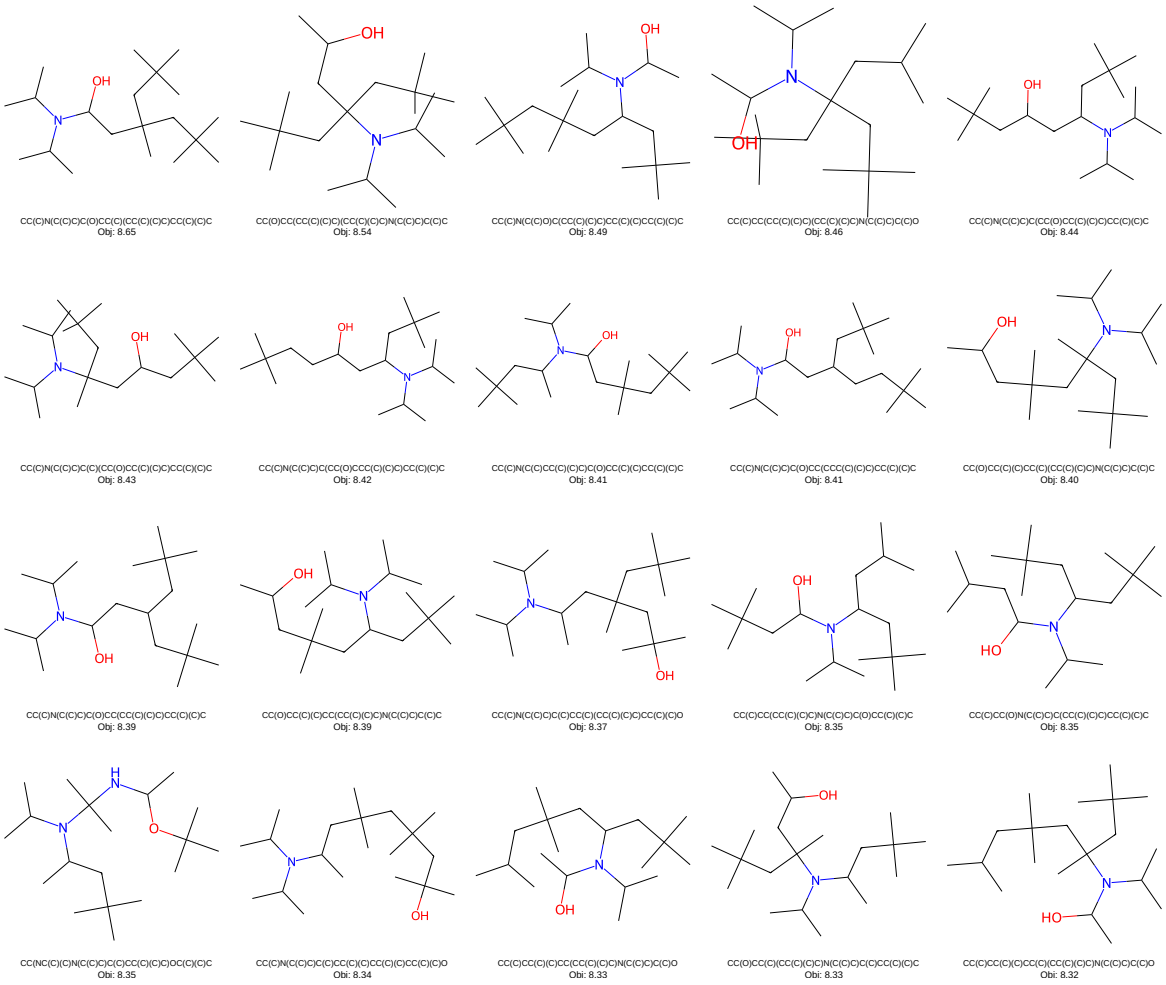
with multiple gpt agents. *Advances in Neural Information Processing Systems*, 36, 7405-7418, 2023.
[8] Brown, N.; Fiscato, M.; Segler, M. H. S.; Vaucher, A. C. GuacaMol: Benchmarking Models for de Novo Molecular Design. *Journal of Chemical Information and Modeling*, 59(3):1096–1108, 2019.

3 Solvent design results

3.1 GraphXForm

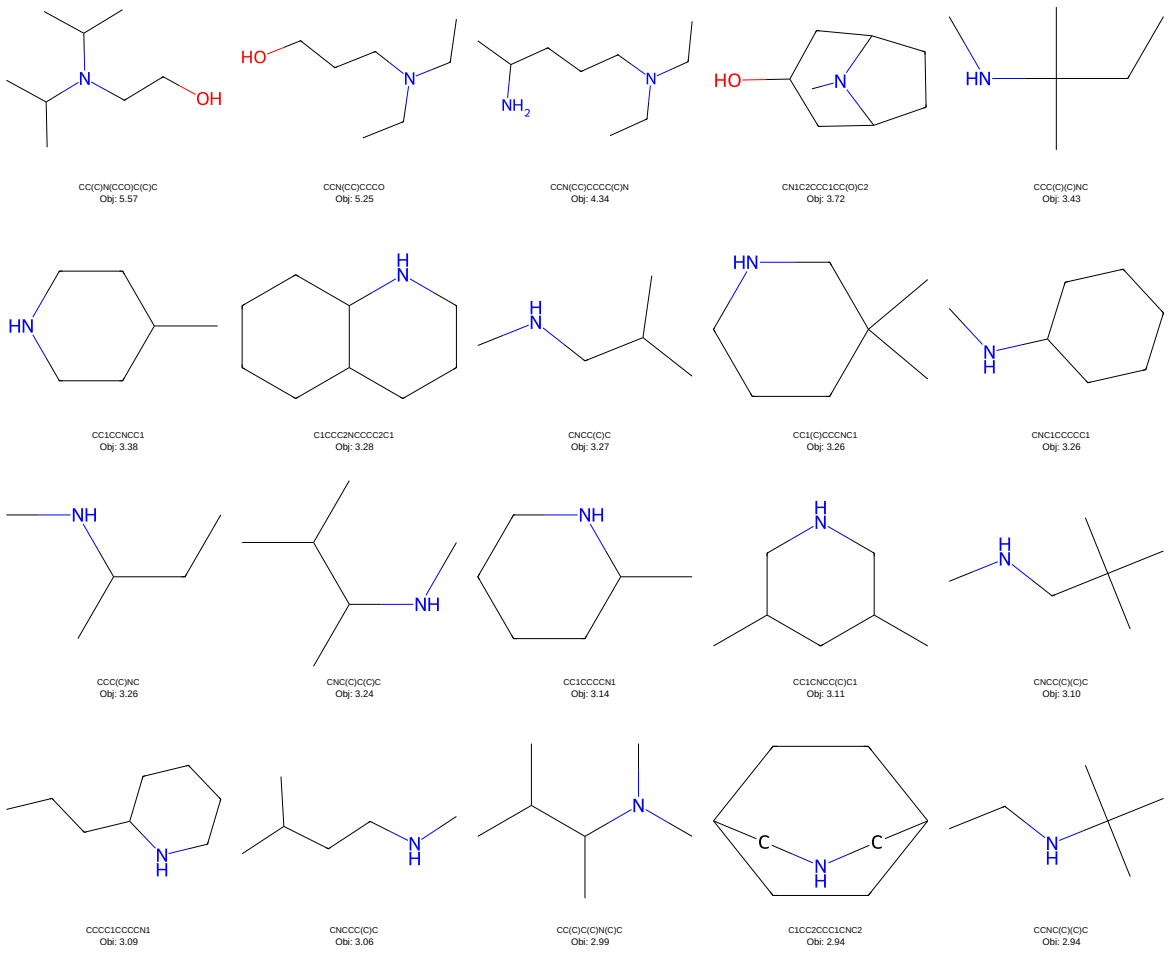


Supplementary Figure 1: Top 20 molecules: GraphXForm, IBA task

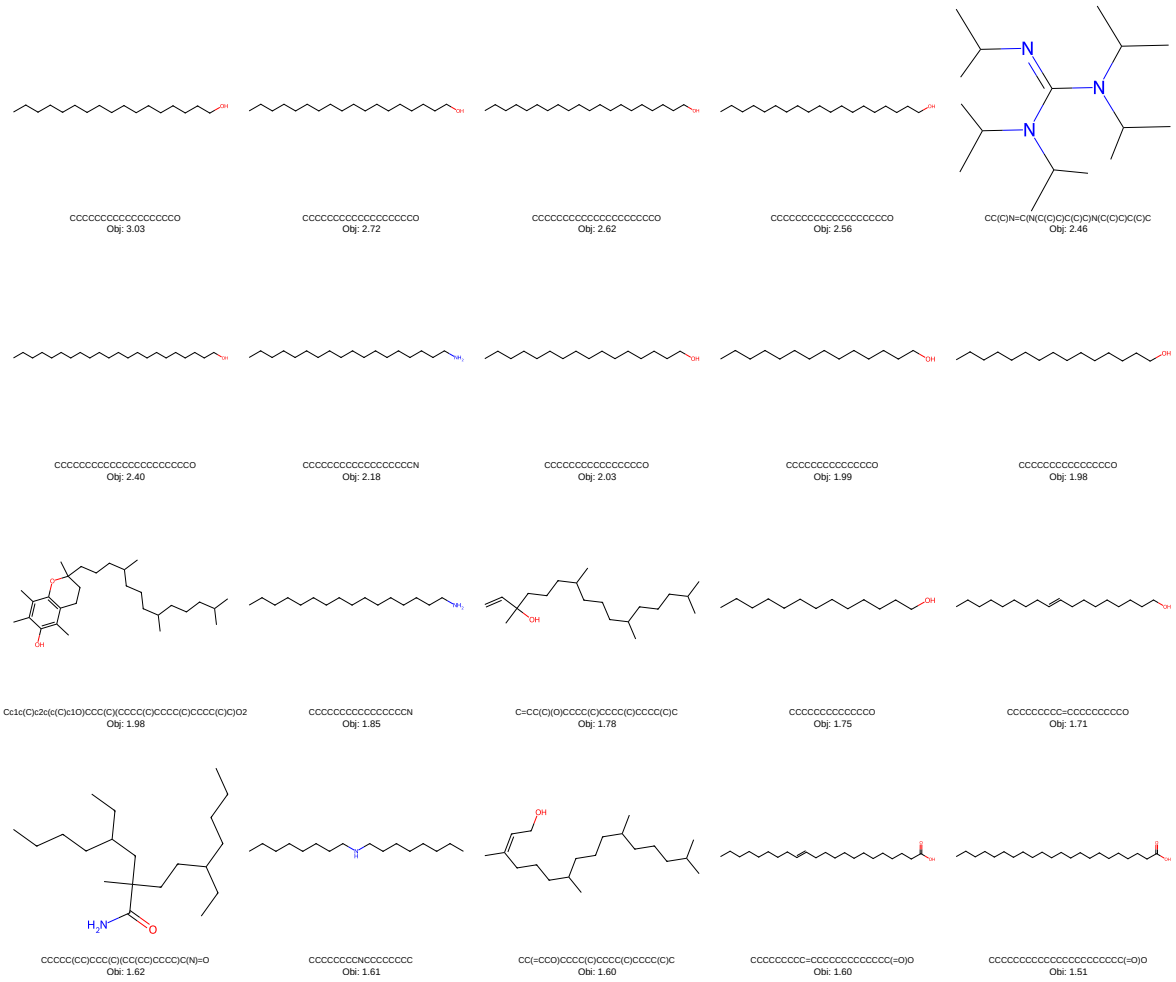


Supplementary Figure 2: Top 20 molecules: GraphXForm, TMB/DMBA task

3.2 List screen

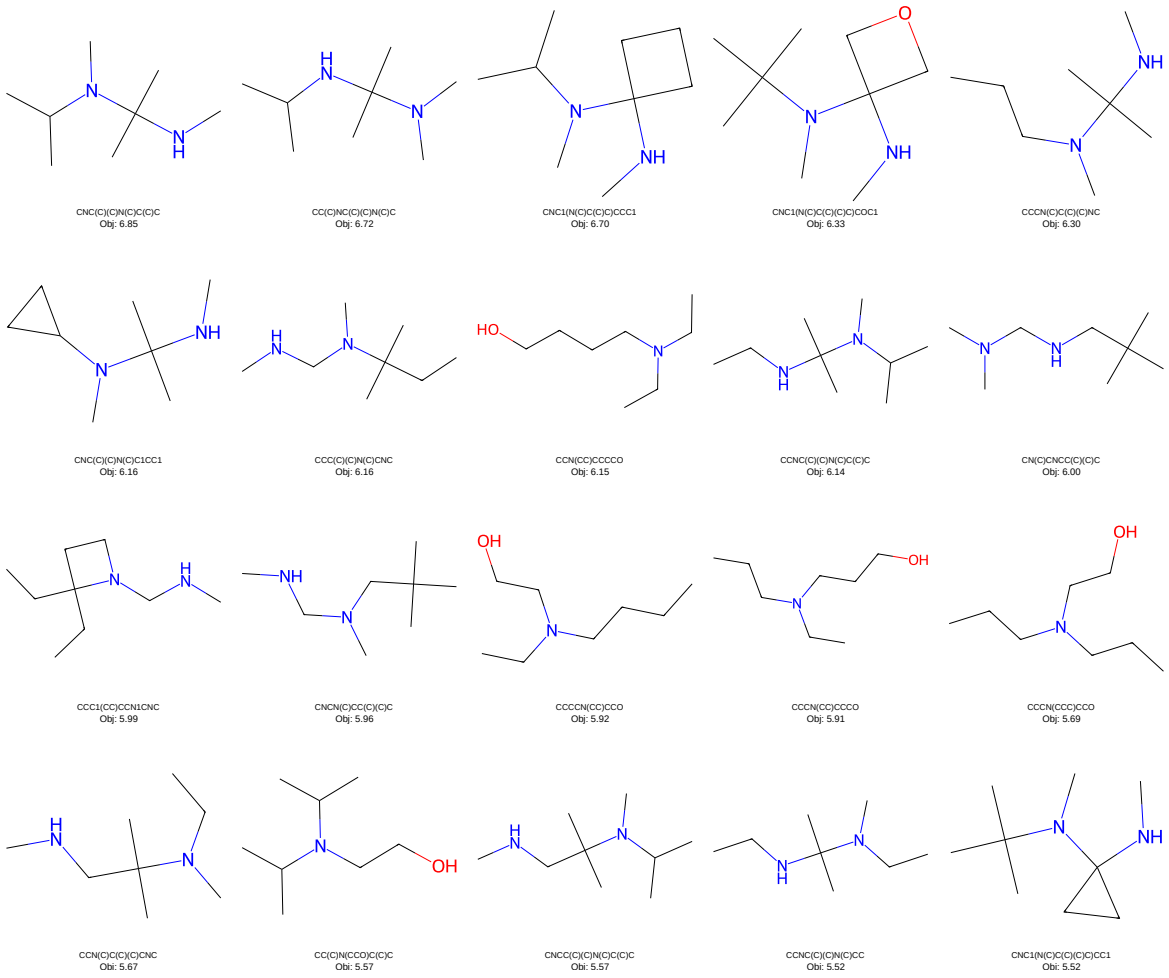


Supplementary Figure 3: Top 20 molecules: List screen, IBA task

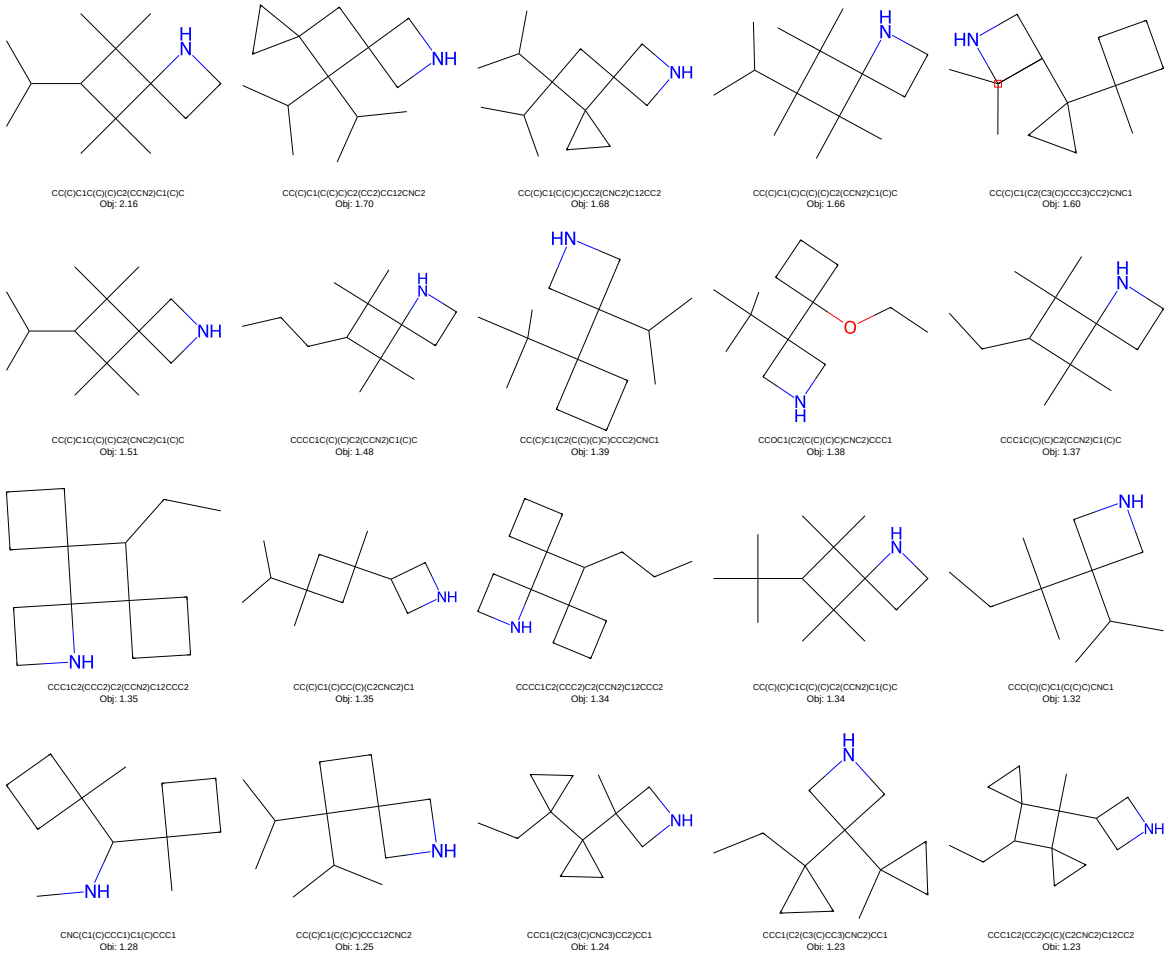


Supplementary Figure 4: Top 20 molecules: List screen, TMB/DMBA task

3.3 JT-VAE

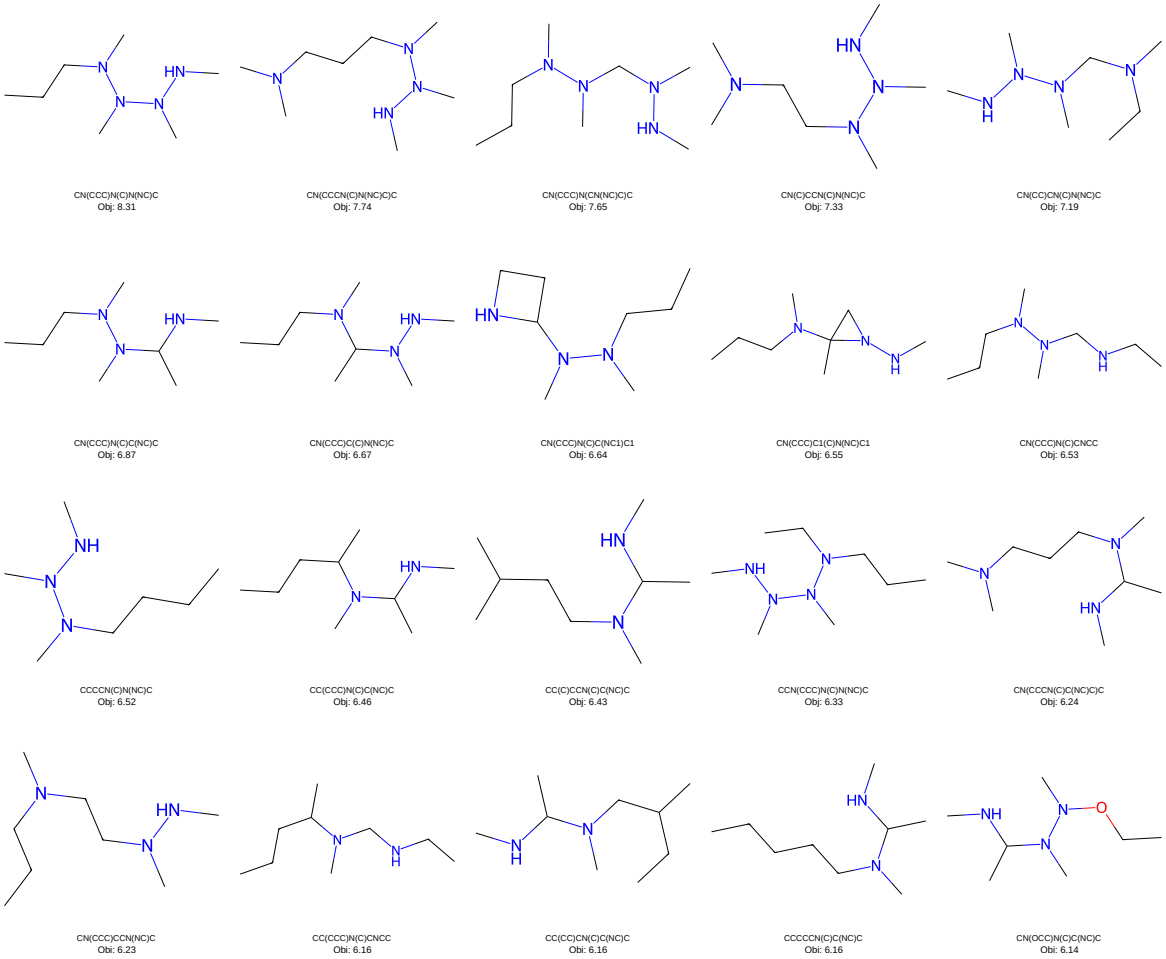


Supplementary Figure 5: Top 20 molecules: JT-VAE, IBA task

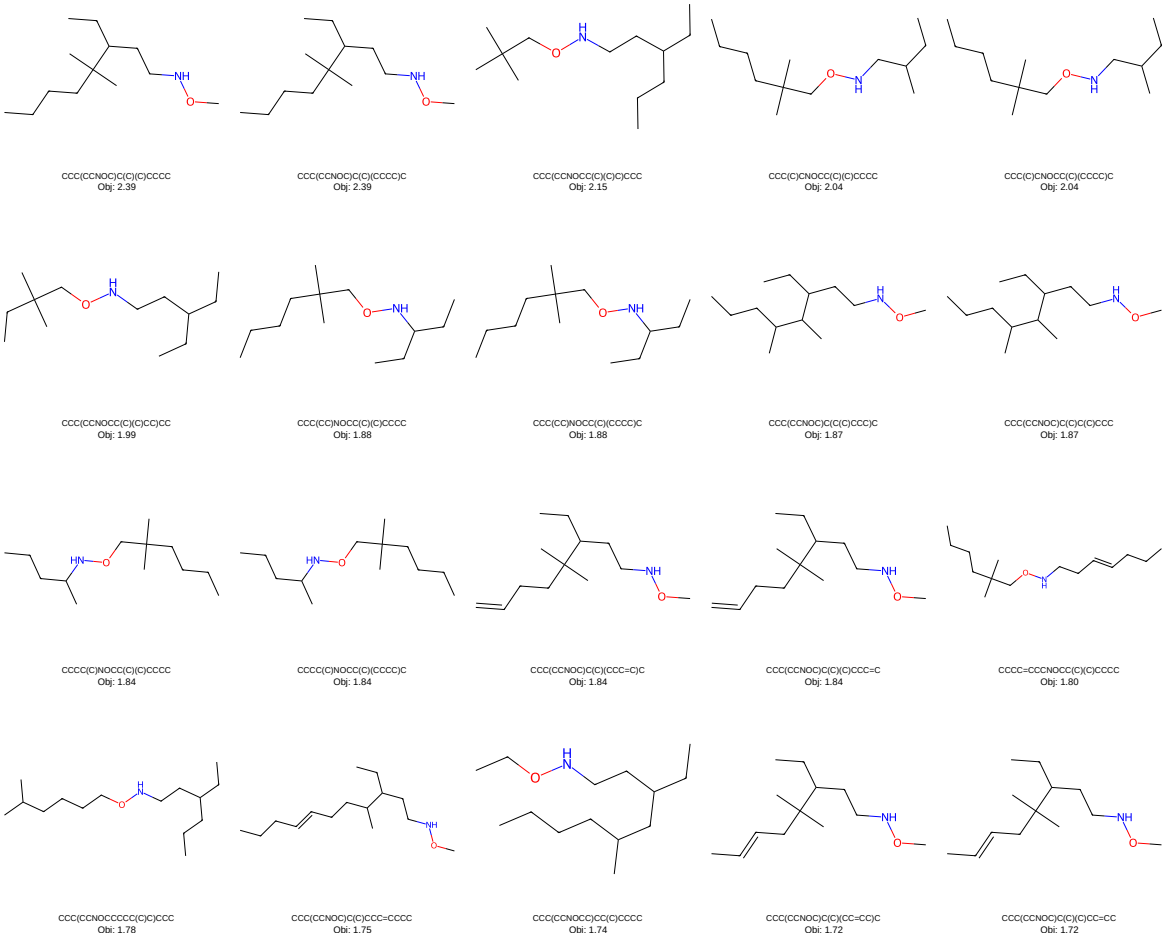


Supplementary Figure 6: Top 20 molecules: JT-VAE, TMB/DMBA task

3.4 STONED

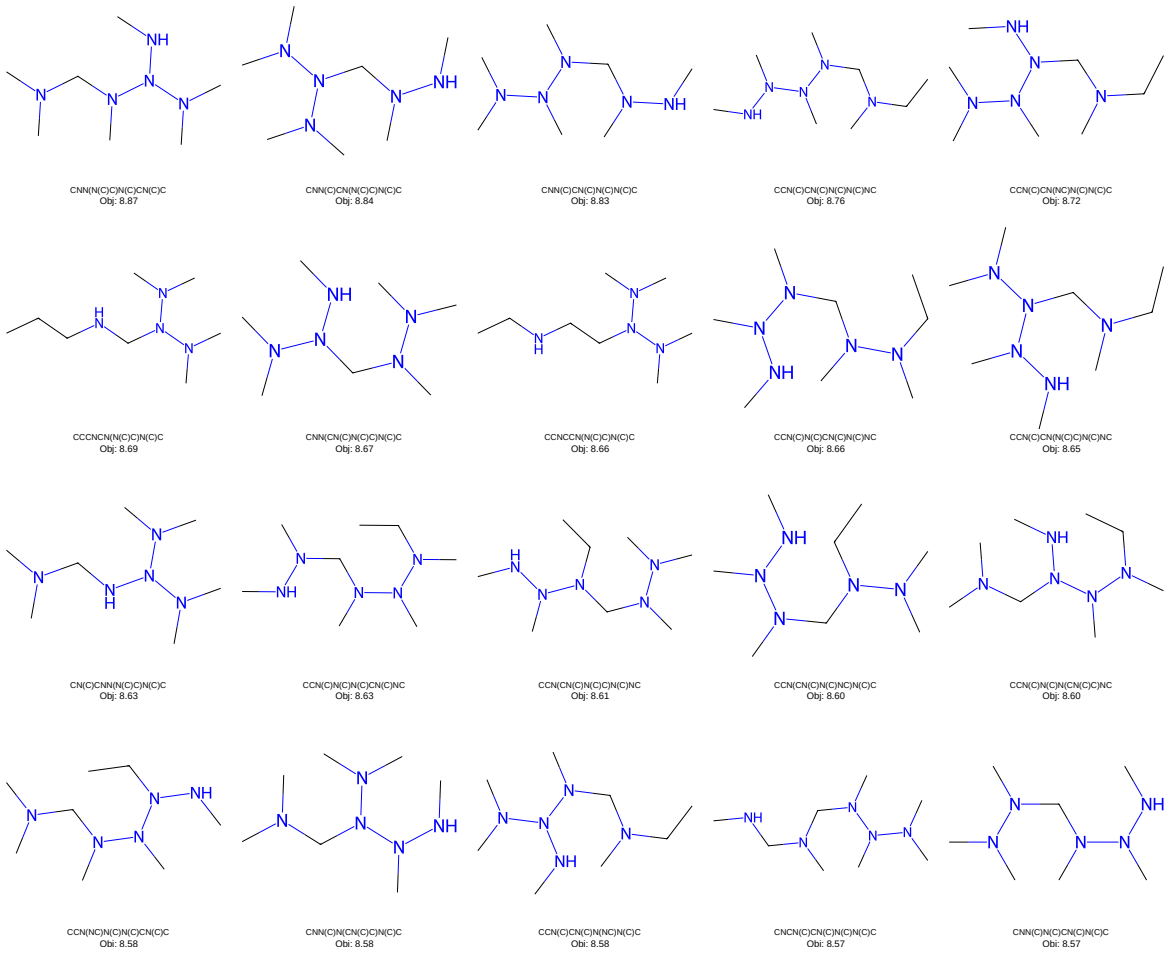


Supplementary Figure 7: Top 20 molecules: STONED, IBA task

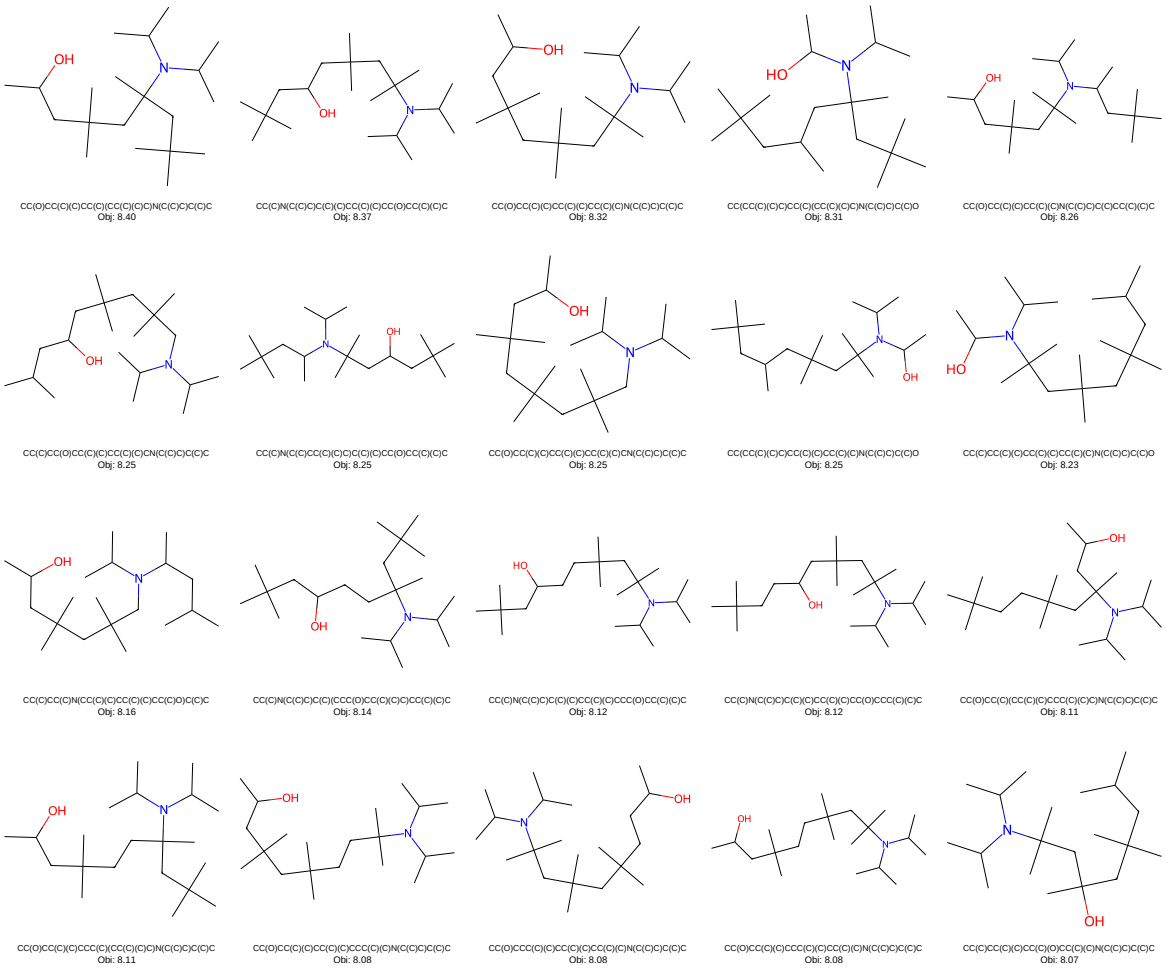


Supplementary Figure 8: Top 20 molecules: STONED, TMB/DMBA task

3.5 Graph GA

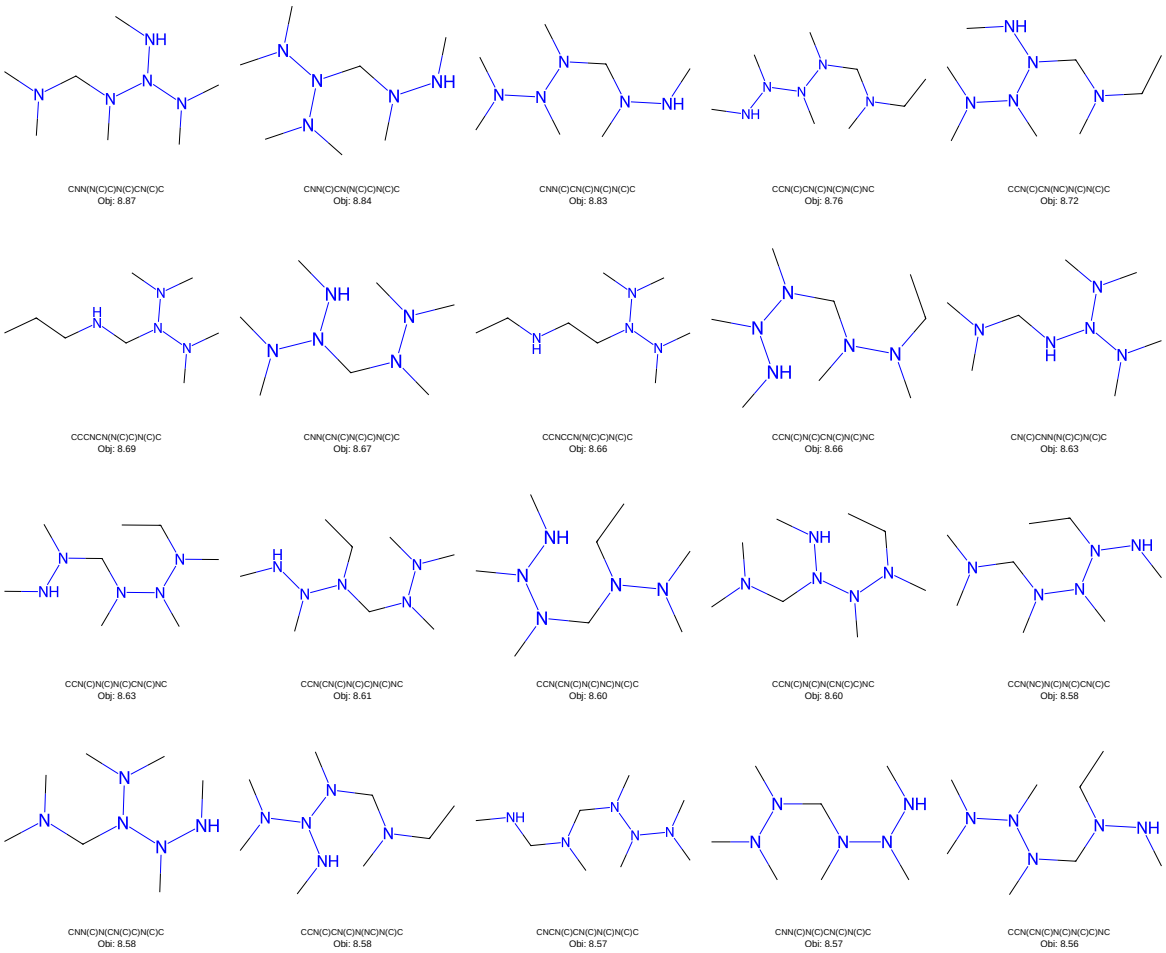


Supplementary Figure 9: Top 20 molecules: Graph GA, IBA task

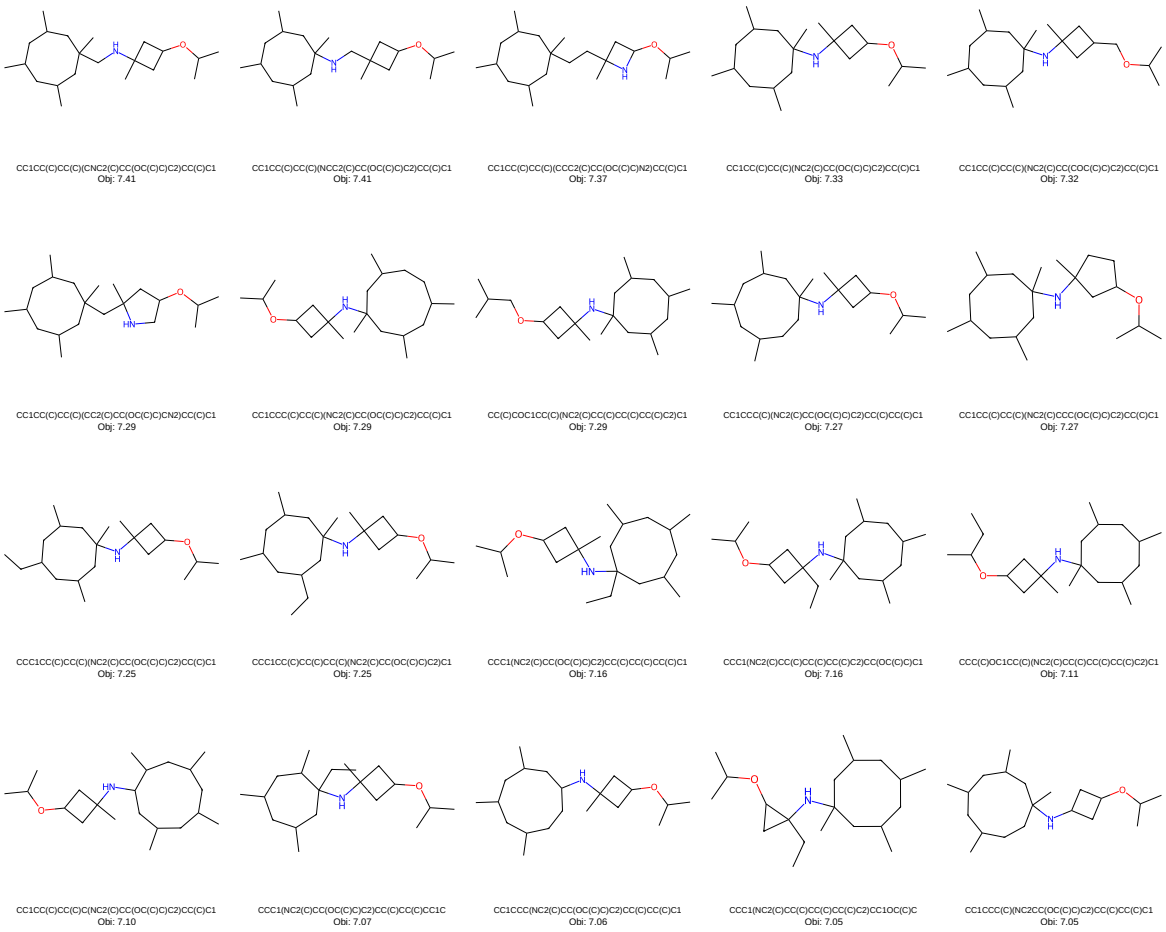


Supplementary Figure 10: Top 20 molecules: Graph GA, TMB/DMBA task

3.6 REINVENT-Transformer

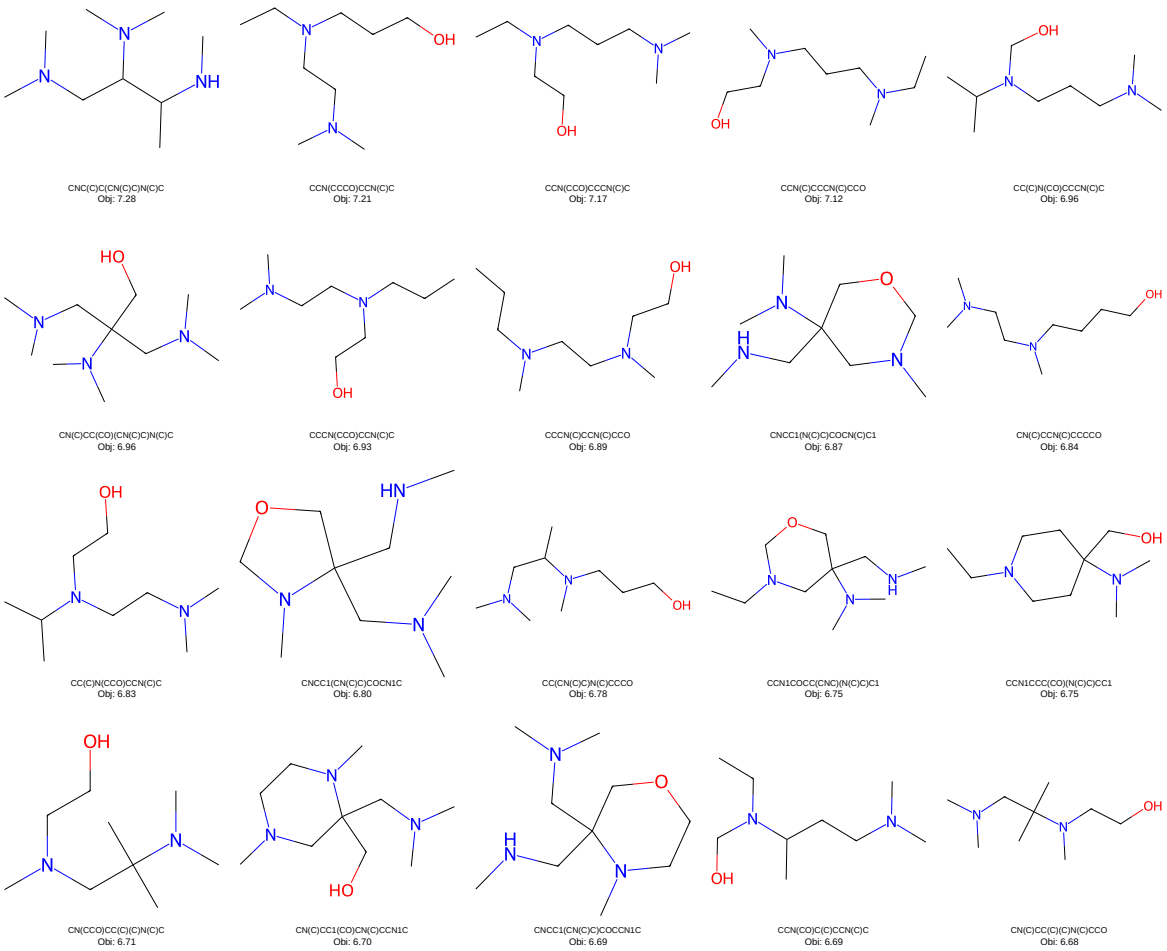


Supplementary Figure 11: Top 20 molecules: REINVENT-Transformer, IBA task

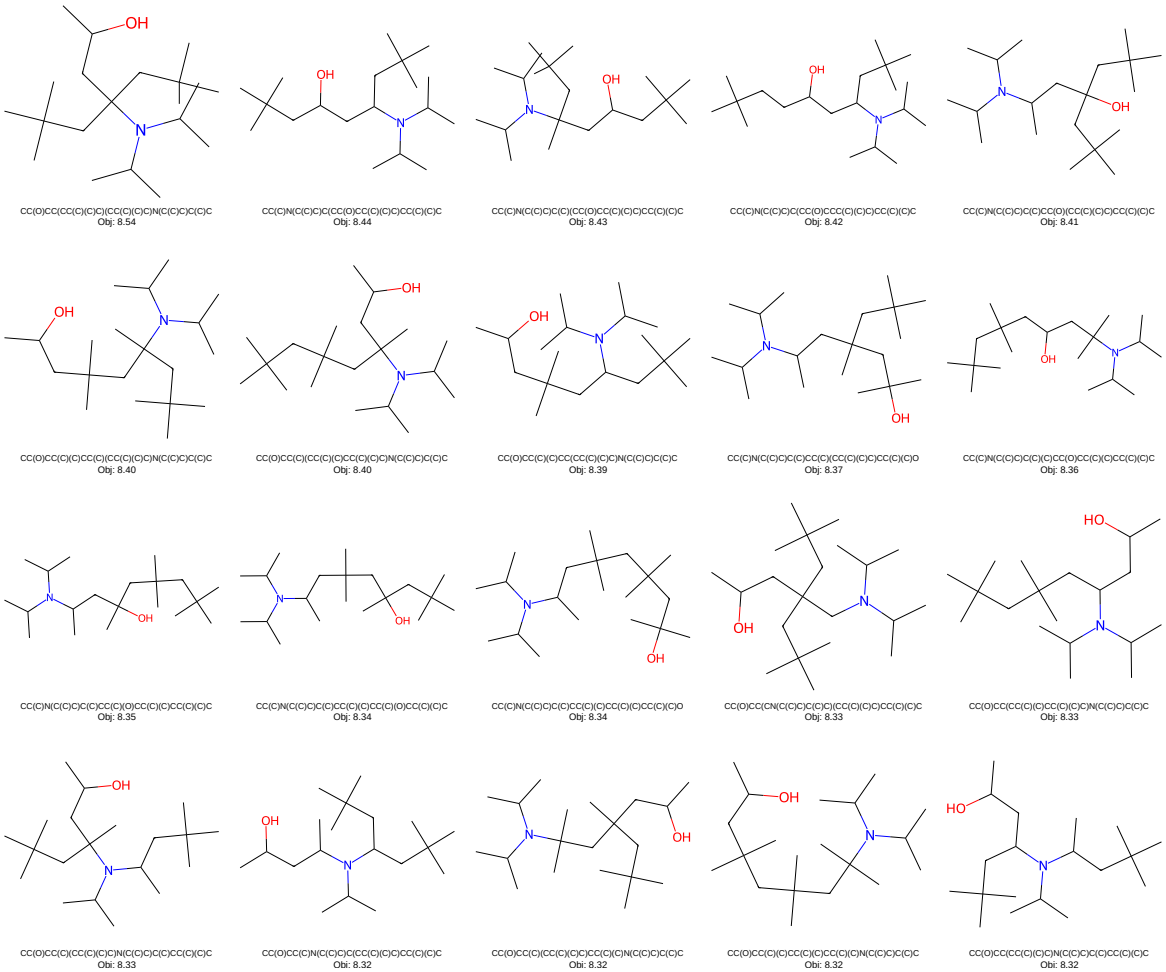


Supplementary Figure 12: Top 20 molecules: REINVENT-Transformer, TMB/DMBA task

3.7 GraphXForm: Structural constraints



Supplementary Figure 13: Top 20 molecules: GraphXForm, IBA task with structural constraints



Supplementary Figure 14: Top 20 molecules: GraphXForm, TMB/DMBA task with structural constraints