Electronic Supplementary Material (ESI) for Digital Discovery.
This journal is © The Royal Society of Chemistry 2023

Supporting Information

Automated patent extraction powers generative modeling in focused chemical spaces

Akshay Subramanian[1,*], Kevin Greenman[2,*], Alexis Gervaix[3],

Tzuhsiung Yang[4], Rafael Gómez-Bombarelli[1]

1 Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

2 Department of Chemical Engineering, Massachussets Institute of Technology, Cambridge, MA, USA

3 Swiss Airtainer SA, Yverdons-les-Bains, Vaud, Switzerland

4 Department of Chemistry, National Tsing Hua University, Hsinchu City, Taiwan

***** Contributed equally to this work

S1 Patent Format Inconsistency

As described in Section 2.2, the USPTO makes available machine-readable patents from 2001 to the present. However, these files are not consistent in their format and directory structure. As a result of these inconsistencies, our original extraction pipeline omitted years 2001-2004 because these years used SGML 2.4 or XML 2.5, whereas years 2005-present used XML 4.0-4.7, as described at `https://bulkdata.uspto.gov/`. Additionally, patents from late 2008 to early 2010 were omitted by our original pipeline because of a different directory structure than other patent releases. The initial training dataset for our generative models omitted some or all patents from the aforementioned years. Since our goal is to incorporate structural priors from a general region of domain-relevant chemical space rather than to extract a comprehensive set of domain-relevant molecules, this omission does not invalidate the approach. As we demonstrate, our approach is helpful for focusing chemical space even while omitting all patent years prior to 2001 (since they are not machine readable). For

the same reason, our approach still works while omitting a subset of years after 2001. However, for the sake of completion and to maximize training dataset coverage of relevant structures, we have resolved these issues in the latest version of our PatentChem code (`https://github.com/learningmatter-mit/PatentChem`). Going forward, users who do their own keyword queries with our code will be unaffected by the problems we initially encountered with certain years.

S2 Processing of patent-extracted data before model training

The goal of our pipeline is to generate structures with limited domain knowledge beyond keywords, so we kept processing/filtering to a minimum except for constraints that allowed for better computational tractability and basic filters on molecular mass. For example, we applied a 1000 g/mol maximum molecular mass cutoff on the OPD dataset primarily because JT-VAE has a sequential decoding process that enumerates combinations of fragment pairs, which scales with the size of fragments and is thus very slow for large molecules. This has the added benefit of eliminating polymers and large candidates (non-ideal for deposition techniques such as chemical vapor deposition). Similarly on the TKI dataset, we imposed maximum and minimum cutoffs of 700 g/mol and 250 g/mol respectively to eliminate candidates that are not "drug-like". We apply the minimum molecular mass constraint in the TKI case since our property optimization objective was similarity to held-out FDA approved drugs whose molecular masses typically fall above 250g/mol.

Our minimal filtering means there are some structures in our training datasets that are not domain-relevant (such as reagents or intermediates). However, the "false positives" (molecules that the model generates because it thinks they are relevant, when in reality they are not relevant) that come from this can be easily filtered out by the property labeling step. Just as a user can choose their own property-labeling method appropriate for their design task when using our code, they could also insert additional domain-knowledge-based preprocessing of the training dataset. Our current work demonstrates that the approach can still be useful even without this preprocessing, but additional filtering may improve results in some domains. We have provided some options for possible filters in our PatentChem code, such as minimum and maximum molecular weight and charged/neutral molecules.

S3 REINVENT+SELFIES

S3.1 TKI

Figure S1 shows the similarity to query structure as a function of training iterations, for each of the 27 held-out FDA-approved TKI molecules. In most cases, we observed an increasing trend in the reward. There were however some instances (ex. Nilotinib and Cabozantinib) where training was unstable and did not converge. Reinforcement Learning algorithms are often highly sensitive to hyperparameters, so it is possible that these cases might require further tuning.

S3.2 OPD

Unlike the TKI dataset case where we had access to the oracle reward, training on the OPD dataset required a proxy neural network reward estimator. Figure S3 shows the test performance of the proxy reward predictor on DFT-calculated optical gaps. We observed that while the reward had an increasing trend during training of the agent (Figure S2(a)), the sampled molecules (Figure S2(b)) did not match the training data well structurally. We hypothesised that this behavior arose from agent identifying and targeting high-uncertainty regions of the property predictor. To investigate this, we also attempted running agent training with a new reward that penalized high uncertainty as estimated by ensemble variance on the property predictor. To achieve this, the reward was modified to include a multiplicative masking term that evaluated whether the ensemble uncertainty was smaller than the 99th percentile of training data uncertainties. Hence molecules for which the property predictor was more uncertain than 99% of the training data would have a reward of zero. We were however unable to achieve model convergence with this modified reward function, i.e., rewards did not display an increasing trend. This was because a majority of molecules generated during training were high-uncertainty points and resulted in a reward of zero. This resulted in the agent having access to very sparse information since poor candidates were sampled at a much higher fraction than good ones. It is also possible that the ensemble uncertainty was not an accurate estimator of model confidence at points that are highly Out of Distribution (OOD), as was observed by Scalia et al. in their work. [1]

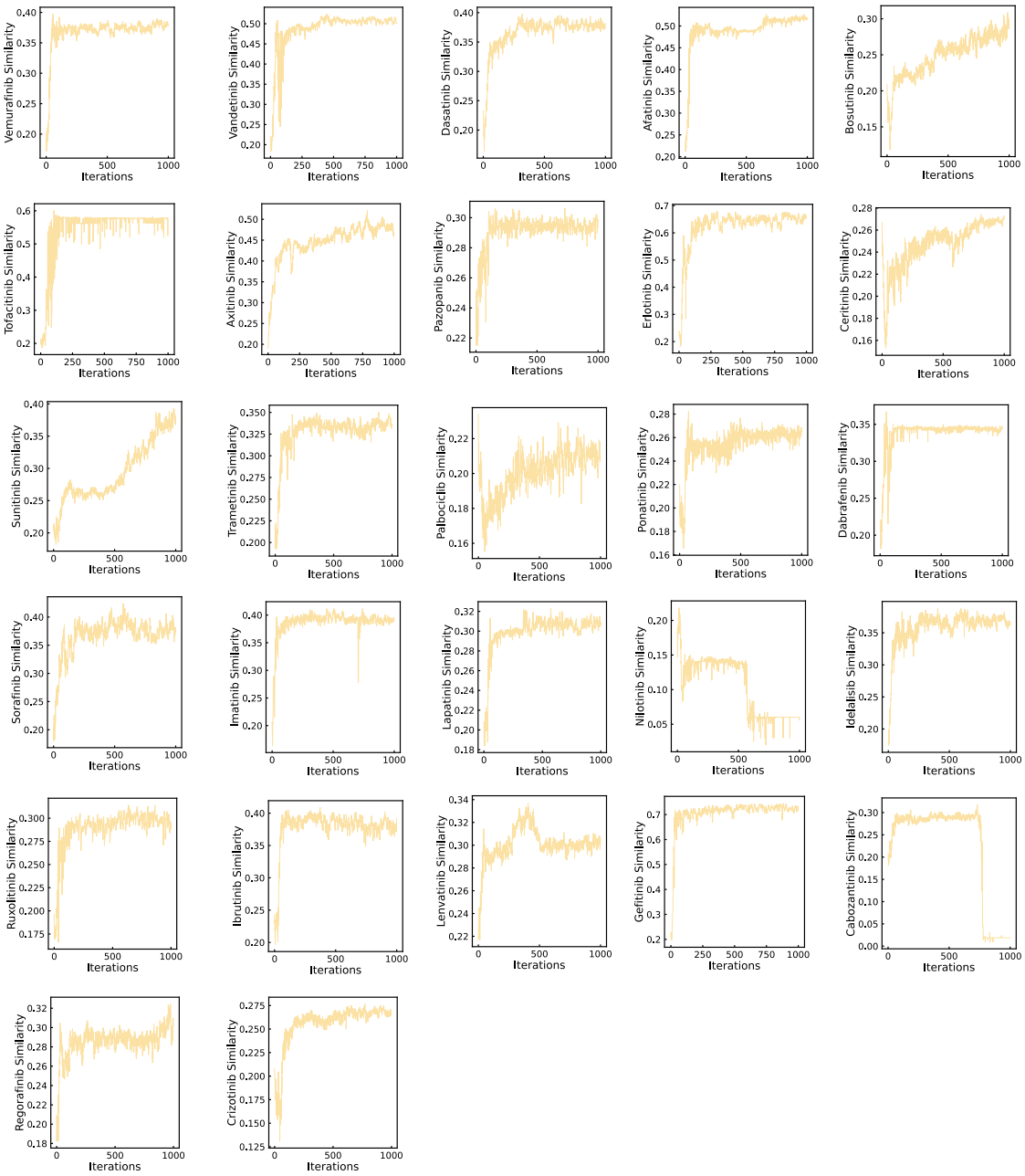

Figure S1: **Tanimoto similarity score computed between generated candidates and FDA approved TKI molecules, as a function of training iteration**

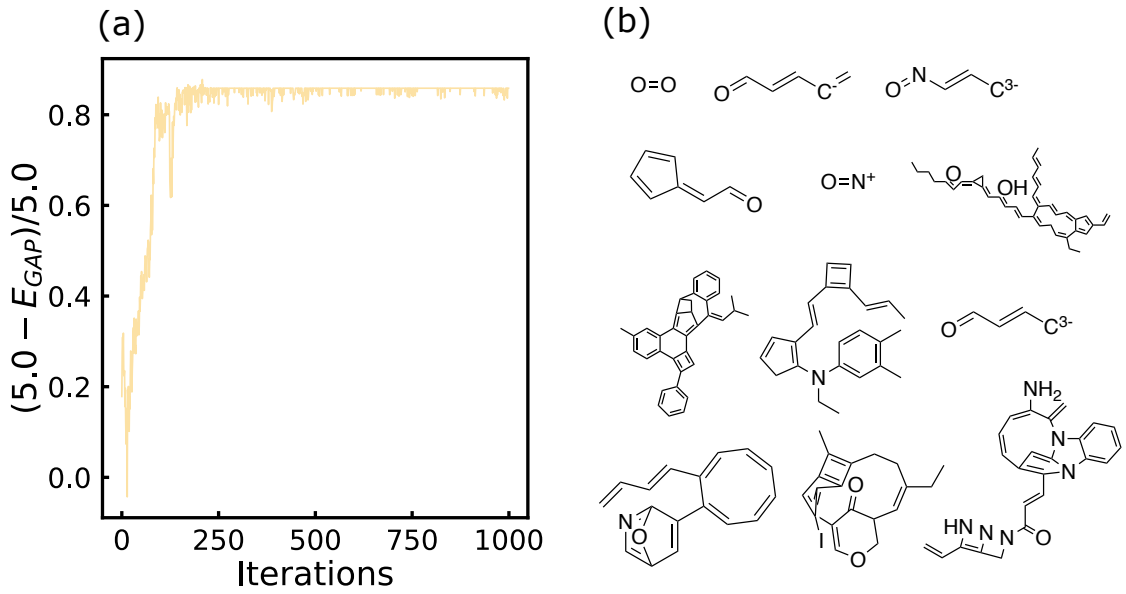

Figure S2: **Results of REINVENT+SELFIES on OPD dataset.** a) Reward score as a function of training iterations. b) Molecules sampled during later stages of training.

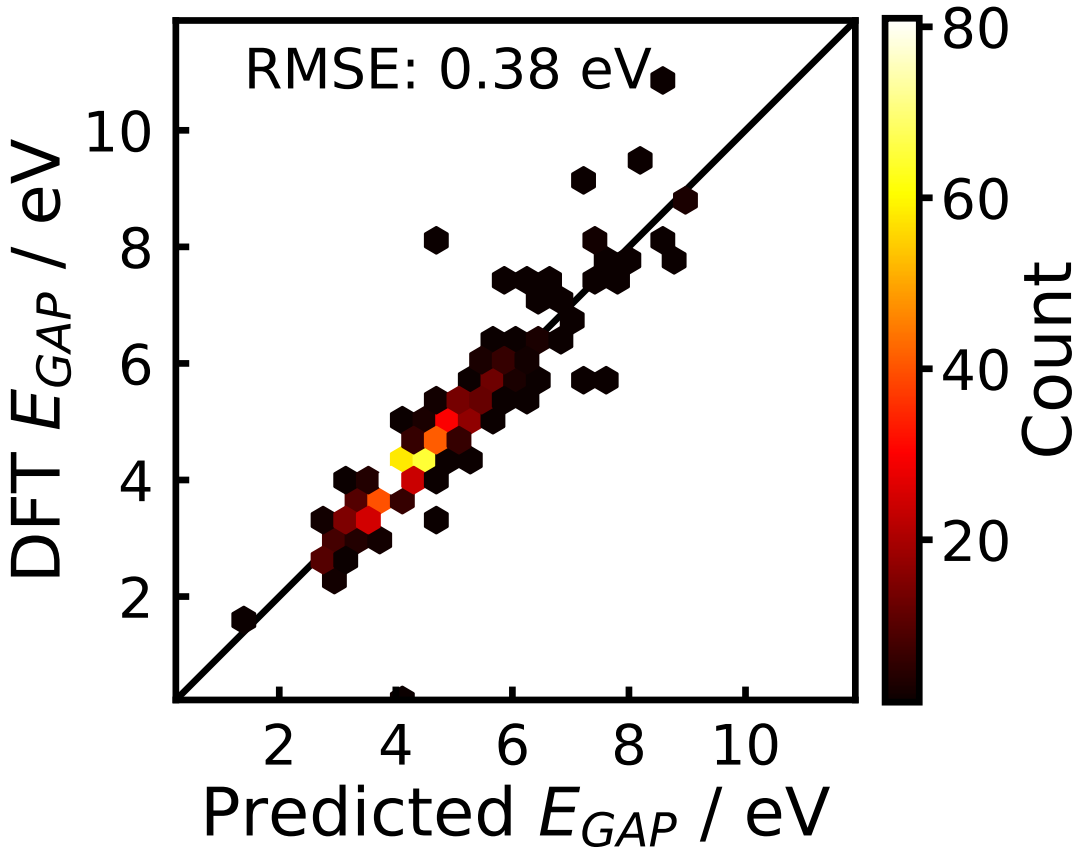

Figure S3: **Comparison between DFT-calculated optical gaps and chemprop-predicted optical gaps, as calculated on the test set.** RMSE on test set was 0.38 eV.

S4 JTVAE training

Since DFT calculations were expensive to perform on the entire patent-mined OPD set, we only labeled a subset of 5568 molecules out of a total of 112436 molecules. To effectively use labeled and unlabeled data during JTVAE training, we utilized all molecules for encoder and decoder training, but only utilized the labeled subset while training the property predictor. The training of encoder, decoder and property predictor were all performed jointly with a multitask loss function. In addition, the property predictor training was

performed on 5 different properties: HOMO, LUMO, optical gap, Synthetic Complexity Score (SCScore) [2], and molecular mass. This was done for two purposes: 1)To aid with latent space regularization, and 2) Multiple tasks could potentially have shared information and thus compound the amount of effective training data seen by the model.

S5 FDA approved TKI candidates

Table S1 lists the names and SMILES representations of the 27 FDA-approved TKI molecules that were held out during all TKI experiments carried out in this paper.

Name	SMILES
Afatinib	CN(C)CC=CC(=O)NC1=C(C=C2C(=C1)C(=NC=N2)NC3=CC(=C(C=C3)F)Cl)OC4CCOC4
Ibrutinib	C=CC(=O)N1CCC[C@H](C1)N2C3=NC=NC(=C3C(=N2)C4=CC=C(C=C4)OC5=CC=CC=C5)N
Pazopanib	CC1=C(C=C(C=C1)NC2=NC=CC(=N2)N(C)C3=CC4=NN(C(=C4C=C3)C)C)S(=O)(=O)N
Axitinib	CNC(=O)C1=CC=CC=C1SC2=CC3=C(C=C2)C(=NN3)/C=C/C4=CC=CC=N4
Idelalisib	CC[C@@H](C1=NC2=C(C(=CC=C2)F)C(=O)N1C3=CC=CC=C3)NC4=NC=NC5=C4NC=N5
Ponatinib	CC1=C(C=C(C=C1)C(=O)NC2=CC(=C(C=C2)CN3CCN(CC3)C)C(F)(F)F)C#CC4=CN=C5N4N=CC=C5
Bosutinib	CN1CCN(CC1)CCCOC2=C(C=C3C(=C2)N=CC(=C3NC4=CC(=C(C=C4Cl)Cl)OC)C#N)OC
Imatinib	CC1=C(C=C(C=C1)NC(=O)C2=CC=C(C=C2)CN3CCN(CC3)C)NC4=NC=CC(=N4)C5=CN=CC=C5
Regorafinib	CNC(=O)C1=NC=CC(=C1)OC2=CC(=C(C=C2)NC(=O)NC3=CC(=C(C=C3)Cl)C(F)(F)F)F
Cabozantinib	COC1=CC2=C(C=CN=C2C=C1OC)OC3=CC=C(C=C3)NC(=O)C4(CC4)C(=O)NC5=CC=C(C=C5)F
Lapatinib	CS(=O)(=O)CCNCC1=CC=C(O1)C2=CC3=C(C=C2)N=CN=C3NC4=CC(=C(C=C4)OCC5=CC(=CC=C5)F)Cl
Ruxolitinib	C1CCC(C1)[C@@H](CC#N)N2C=C(C=N2)C3=C4C=CNC4=NC=N3
Ceritinib	CC1=CC(=C(C=C1C2CCNCC2)OC(C)C)NC3=NC=C(C(=N3)NC4=CC=CC=C4S(=O)(=O)C(C)C)Cl
Sorafinib	CNC(=O)C1=NC=CC(=C1)OC2=CC=C(C=C2)NC(=O)NC3=CC(=C(C=C3)Cl)C(F)(F)F
Crizotinib	C[C@H](C1=C(C=CC(=C1Cl)F)Cl)OC2=C(N=CC(=C2)C3=CN(N=C3)C4CCNCC4)N
Sunitinib	CCN(CC)CCNC(=O)C1=C(NC(=C1C)/C=C\2/C3=C(C=CC(=C)F)NC2=O)C
Dabrafenib	CC(C)(C)C1=NC(=C(S1)C2=NC(=NC=C2)N)C3=C(C(=CC=C3)NS(=O) (=O)C4=C(C=CC=C4F)F)F
Tofacitinib	C[C@@H]1CCN(C[C@@H]1N(C)C2=NC=NC3=C2C=CN3)C(=O)CC#N
Dasatinib	CC1=C(C(=CC=C1)Cl)NC(=O)C2=CN=C(S2)NC3=CC(=NC(=N3)C)N4CCN(CC4)CCO
Lenvatinib	COC1=CC2=NC=CC(=C2C=C1C(=O)N)OC3=CC(=C(C=C3)NC(=O)NC4CC4)Cl
Trametinib	CC1=C2C(=C(N(C1=O)C)NC3=C(C=C(C=C3)I)F)C(=O)N(C(=O)N2C4=CC=CC(=C4)NC(=O)C)C5CC5
Erlotinib	COCCOC1=C(C=C2C(=C1)C(=NC=N2)NC3=CC=CC(=C3)C#C)OCCOC
Nilotinib	CC1=C(C=C(C=C1)C(=O)NC2=CC(=CC(=C2)C(F)(F)F)N3C=C(N=C3)C)NC4=NC=CC(=N4)C5=CN=CC=C5
Vandetinib	CN1CCC(CC1)COC2=C(C=C3C(=C2)N=CN=C3NC4=C(C=C(C=C4)Br)F)OC
Gefitinib	COC1=C(C=C2C(=C1)N=CN=C2NC3=CC(=C(C=C3)F)Cl)OCCCN4CCOCC4
Palbociclib	CC1=C(C(=O)N(C2=NC(=NC=C12)NC3=NC=C(C=C3)N4CCNCC4)C5CCCC5)C(=O)C
Vemurafinib	CCCS(=O)(=O)NC1=C(C(=C(C=C1)F)C(=O)C2=CNC3=C2C=C(C=N3)C4=CC=C(C=C4)Cl)F

Table S1: **Names and SMILES strings of held-out FDA approved TKI molecules.**

S6 Visualizing structural resemblance to training data

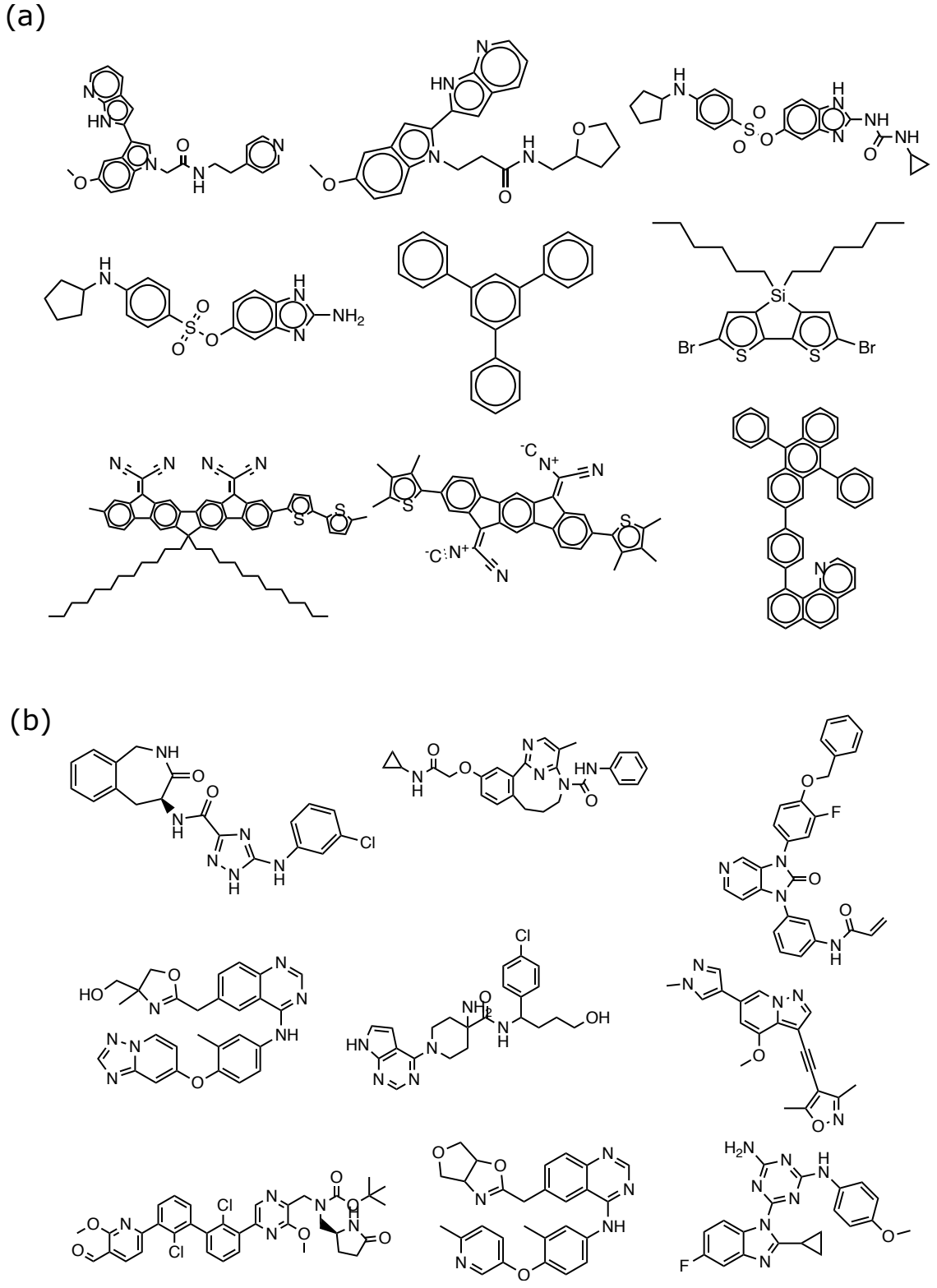

Figure S4: **Sample molecular structures obtained from random sampling of trained RNN+SELFIES model on a) OPD b) TKI dataset**

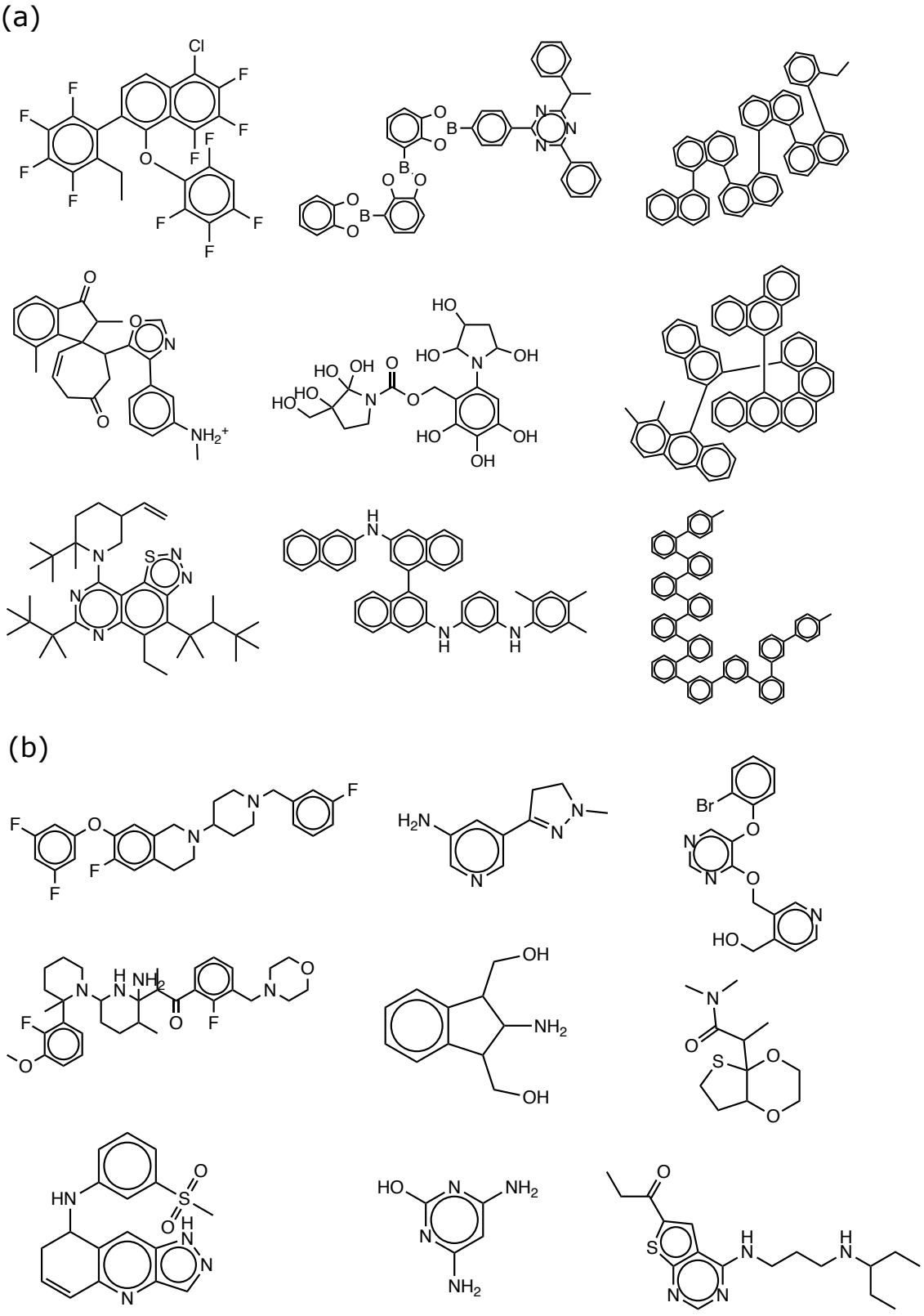

Figure S5: **Sample molecular structures obtained from random sampling of trained JTVAE model on a) OPD b) TKI dataset**

S7 References

[1] G. Scalia, C. A. Grambow, B. Pernici, Y.-P. Li, and W. H. Green, "Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction," *Journal of chemical information and modeling*, vol. 60, no. 6, pp. 2697–2717, 2020.

[2] C. W. Coley, L. Rogers, W. H. Green, and K. F. Jensen, "SCScore: synthetic complexity learned from a reaction corpus," *Journal of chemical information and modeling*, vol. 58, no. 2, pp. 252–261, 2018.