SUPPLEMENTARY MATERIAL - LEARNING TO NAVIGATE THE SYNTHETICALLY ACCESSIBLE CHEMICAL SPACE USING REINFORCEMENT LEARNING

Sai Krishna Gottipati *[1], Boris Sattarov *[1], Sufeng Niu[10], Yashaswi Pathak[1,6], Haoran Wei[1,9], Shengchao Liu[2,8], Karam J. Thomas[1], Simon Blackburn[8], Connor W. Coley[7], Jian Tang[5,8,11], Sarath Chandar[3,5,8], and Yoshua Bengio[2,4,5,8]

[1]99andBeyond
[2]University of Montreal
[3]Ecole Polytechnique Montréal
[4]CIFAR Senior Fellow
[5]Canada CIFAR AI Chair
[6]Center for Computational Natural Sciences and Bioinformatics, IIIT Hyderabad
[7]Department of Chemical Engineering, Massachusets Institute of Technology
[8]Mila - Quebec AI Institute
[9]University of Delaware
[10]Clemson University, South Carolina
[11]HEC Montréal

1 Choice of Parameters

During training, for the bootstrapping phase of first 3,000 time steps, the agent randomly chooses any valid reaction template T and any valid reactant $R^{(2)}$. After the bootstrapping phase, a Gaussian noise of mean 0 and standard deviation 0.1 is added to the action outputted by the π network. No noise is added during the inference phase.

While updating the critic network, we multiply the normal random noise vector with policy noise of 0.2 and then clip it in the range -0.2 to 0.2. This clipped policy noise is added to the action at the next time step a' computed by the target actor networks f and π. The actor networks (f and π networks), target critic and target actor networks are updated once every two updates to the critic network.

The representation of molecules plays a very important role in the overall performance of PGFS. We have experimented with three feature representations: ECFP, MACCS and custom features from MolDSet (RLV2), and observed that ECFP as state features and RLV2 as action features are the best representations. We used the following features in RLV2: MaxEStateIndex, MinEStateIndex, MinAbsEStateIndex, QED, MolWt, FpDensityMorgan1, BalabanJ, PEOE-VSA10, PEOE-VSA11, PEOE-VSA6, PEOE-VSA7, PEOE-VSA8, PEOE-VSA9, SMR-VSA7, SlogP-VSA3, SlogP-VSA5, EState-VSA2, EState-VSA3, EState-VSA4, EState-VSA5, EState-VSA6, FractionCSP3, MolLogP, Kappa2, PEOE-VSA2, SMR-VSA5, SMR-VSA6, EState-VSA7, Chi4v, SMR-VSA10, SlogP-VSA4, SlogP-VSA6, EState-VSA8, EState-VSA9, VSA-EState9.

Figures 1,2 show the f network loss, policy loss, value loss and the average (inference) reward on a set of randomly chosen 100 initial reactants as the training progresses; for all possible combinations of state and action feature representations; for shaped-QED and shaped-HIV-CCR5 rewards respectively.

2 Examples of Proposed Compounds And Their Synthetic Paths, Additional Information On Reaction Templates, Building Blocks and Results

Figures 4, 5, 6, 7 depict synthetic routes and structures of the compounds proposed by PGFS with maximum QED, penalized clogP, HIV-integrase and HIV-RT scores respectively. Figure 3 demonstrates the proposed CCR5 structure alongside the most similar one in the training set utilized to build the CCR5 QSAR model. Figure 8 demonstrates

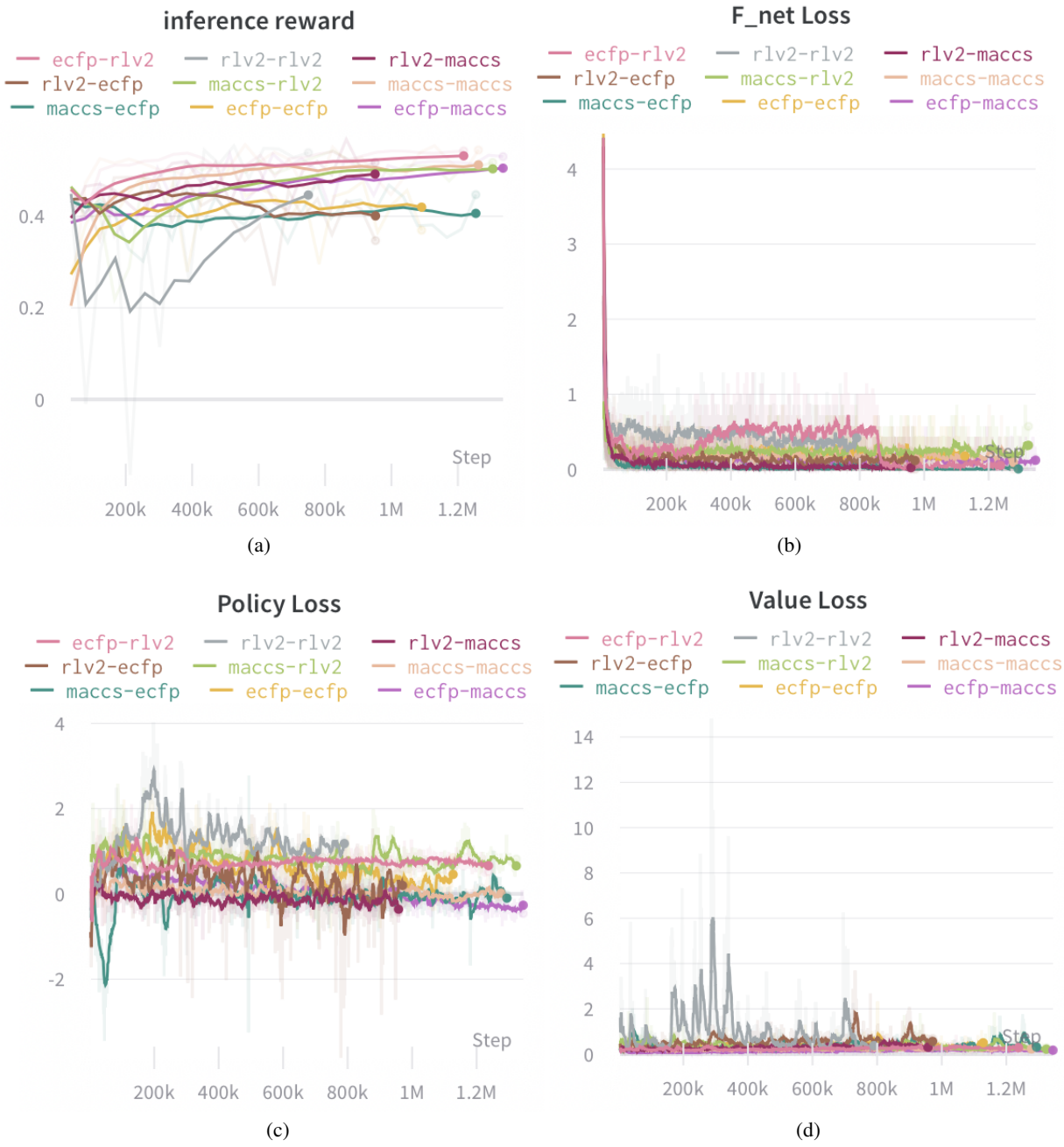


Figure 1: Plots of (a) inference reward; (b) f network loss; (c) policy loss; (d) value loss; for shaped-QED reward. We can observe that ECFP as state features and RLV2 as action features (pink curve: ECFP-RLV2) performed best in terms of inference reward

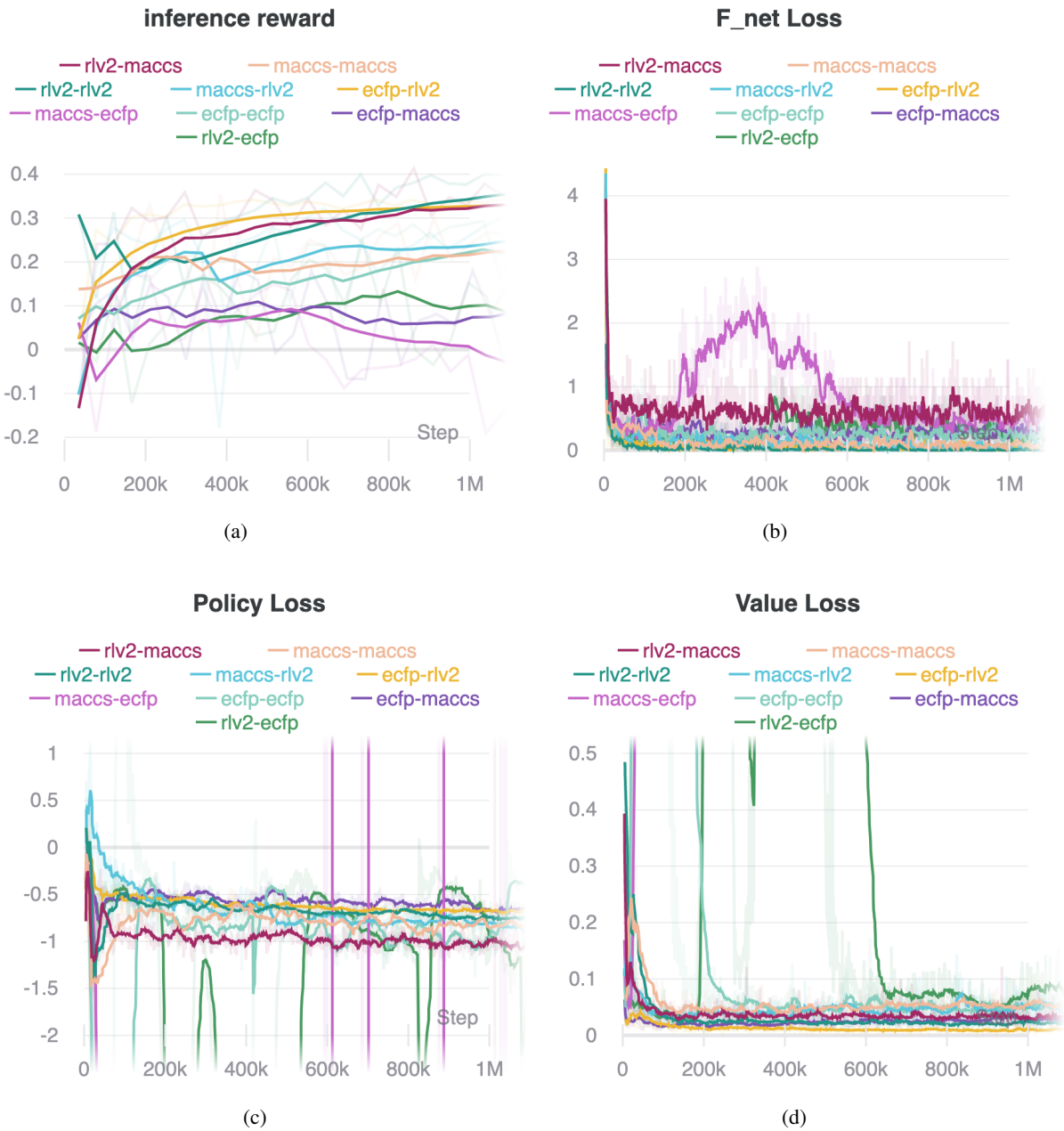


Figure 2: Plots of (a) inference reward; (b) f network loss; (c) policy loss; (d) value loss; for shaped-hiv-ccr5 reward. We can observe that ECFP as state features and RLV2 as action features (yellow curve: ECFP-RLV2) performed best in terms of inference reward

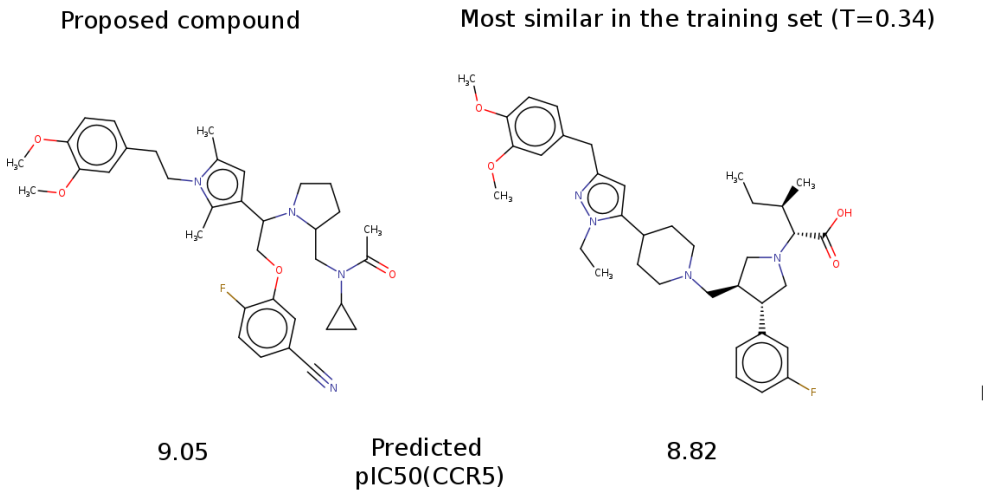


Figure 3: Structure of the compound proposed using PGFS with the highest predicted activity against CCR5 compared to the structure from the corresponding QSAR modeling training set with the highest Tanimoto Similarity ([1]) using Morgan fingerprints with the radius of 2 as implemented in RDKit. The predicted pIC50 value is shown below the proposed structure and experimental pIC50 value is shown below the training set structure.

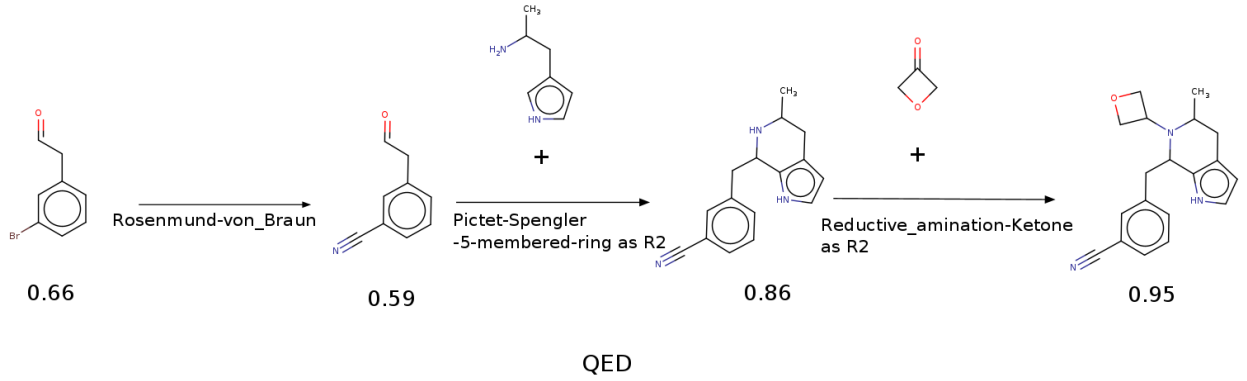


Figure 4: Proposed synthetic route and structure of the compound with the highest score (from the Table 2) obtained during the PGFS training using the QED score as a reward.

performance comparison between Random Search and PGFS using HIV-Int and HIV-RT rewards similar to the Figure 4 of the main text.

Figure 9(a) shows the distribution of available reaction templates for compounds in the Enamine building blocks dataset used in this study. Figure 9(b) shows how many second reactants are available for reaction templates in the set used in this study.

3 Additional information on the QSAR-modeling

3.1 Training and Cross-validation procedure

The QSAR-based validation pipeline used in this study is similar to that reported by [8]. We have used the LightGBM implementation ([4]) of the Gradient Boosting Decision Tree (GBDT) as an algorithm to train a supervised regression model on measured pIC_{50} ($-\log_{10}IC_{50}$, where the IC_{50} is the concentration of a molecule that produces a half-maximum inhibitory response) values associated with 3 HIV-related targets reported in the ChEMBL database (see QSAR data collection and curation). All the parameters of LightGBMRegressor were kept default, except for the 'max-bin', 'learning-rate' and 'min-split-gain' which were set to 15, 0.05 and 0.0001, respectively, to decrease the complexity of the model and training times. Training was done with early stopping on the validation set with 'early-stopping-rounds' = 20, so each model has a different number of estimators. We used 199 molecular descriptors available in RDKit

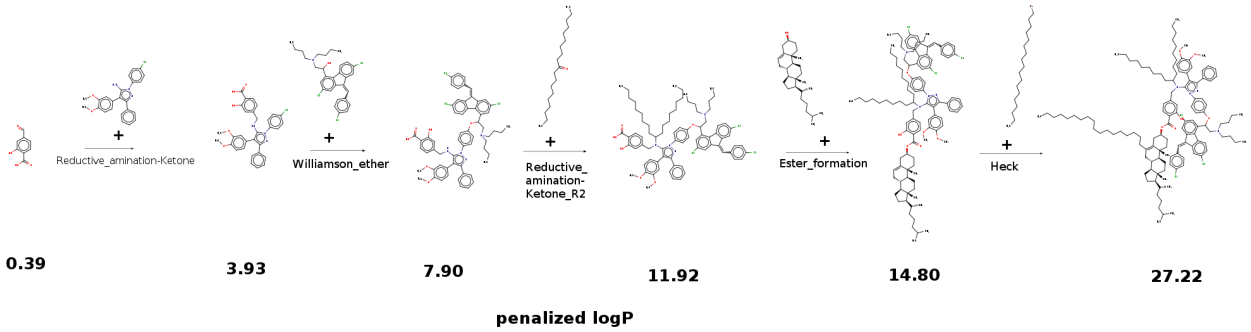


Figure 5: Proposed synthetic route and structure of the compound with the highest score (from the Table 2 of the manuscript) obtained during the PGFS training using the penalized clogP score as a reward.

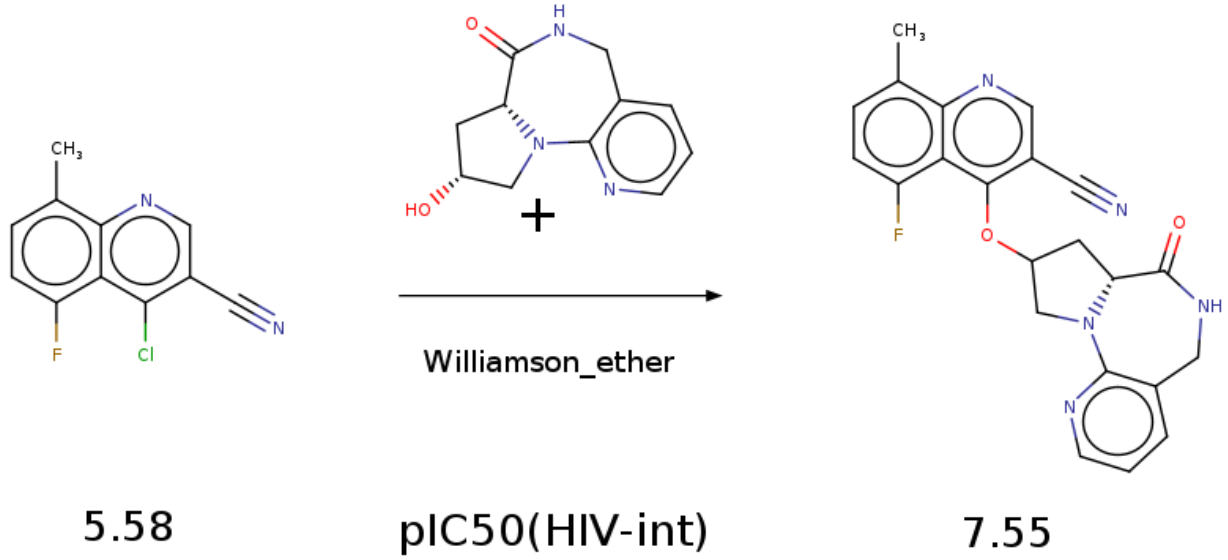


Figure 6: Proposed synthetic route and structure of the compound with the highest score (from the Table 2 of the manuscript) obtained during the PGFS training using predicted activity score (pIC50) against HIV-Integrase as a reward.

Table 1: Cross-validation performance values for the trained QSAR models. Presented metrics and values: R^2 - coefficient of determination, R^2_{adj} - adjusted coefficient of determination ([9]), *MAE* - mean absolute error, *Range* - range of the response values (pIC50) in the dataset. Values in the *avg.* column were calculated by averaging the performance of the 25 models (five-fold cross validation repeated five times) on their unique random validation sets. Values in the *agg.* column were calculated by combining all out-of-fold predictions: Each compound instance was predicted as the average prediction of only five models that were built while the instance was not present in the training set of these models during the five-fold cross validation repeated five times. These prediction values were compared with the ground truth. The formulas for calculating the metrics as well as measured vs. predicted plots are presented in the Appendix Section - 3.2

Dataset	R^2_{adj}		R^2		MAE		Range
	agg	avg.	agg	avg.	agg	avg.	
CCR5	0.72	0.64 ± 0.03	0.72	0.69 ± 0.03	0.51	0.54 ± 0.02	4.04-10.30
HIV-Int	0.68	0.45 ± 0.07	0.69	0.65 ± 0.04	0.45	0.48 ± 0.03	4.00-8.15
HIV-RT	0.53	0.40 ± 0.06	0.55	0.52 ± 0.05	0.51	0.53 ± 0.03	4.00-8.66

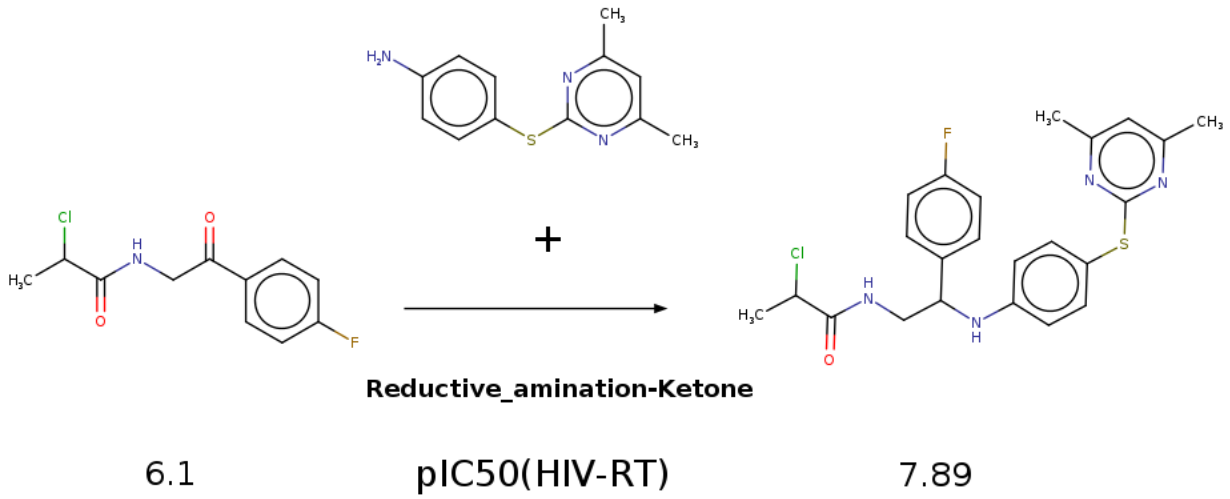


Figure 7: Proposed synthetic route and structure of the compound with the highest score (from the Table 2 of the manuscript) obtained during the PGFS training using predicted activity score (pIC50) against HIV-RT as a reward.

([5]) as an initial set of features prior to feature selection. The training procedure was performed using a five-fold cross-validation (5-fold CV) procedure repeated five times. To reduce the complexity of the produced models during the five-fold cross validation, after the initial set was split into training and validation sets and features were scaled using the training set of that fold random features from every pair of highly correlated features (Pearson's correlation coefficient higher than 0.9) were removed. Then recursive feature elimination procedure was conducted using a five-fold cross-validation on the training set of the current fold and feature importances from LightGBMRegressor (sklearn API). The minimal required set of features with the highest adjusted determination coefficient (measured on 5-fold CV of the current training set) R^2_{adj} were selected to build a final model of this fold which was evaluated on its validation set. The resulting final 25 models (5-fold CV repeated 5 times) for each HIV-target were used as an ensemble to predict pIC_{50} of the produced compounds. Each model is associated with its own unique but overlapping set of selected features. The selected set of features for each model is reported in the github repository in the corresponding "list_of_selected_features_names.txt" files. The development of multiple models for each given target makes it more difficult for the RL framework to exploit adversarial attacks on the predictor and increases the overall robustness of the predictive model. The applicability domain (AD) ([11]) of the ensemble of predictors was estimated using a feature space distance-based ([7, 10]) approach. The feature sets for the AD calculation were selected as features that were used by all the 25 models in the ensemble for each target. These features are the most important for the ensemble prediction because they were selected via an independent feature selection procedure every time at every fold of the CV procedure. According to our definition of AD in this study, a compound is considered to be inside of the AD of the predictive ensemble if its average distance D_i in the normalized feature space to its K closest neighbors from the training set is equal or less than the sum of the mean of that value estimated for each point of the training set $\overline{D_t}$ and corresponding standard deviation S multiplied by Z (1.5 in this study).

$$D_i \leq \overline{D_t} + Z * S_t$$

In this study, K used in K closest neighbors depends on the QSAR dataset and is calculated as the square root of the number of compounds in the corresponding dataset. (For CCR5 - K=41, HIV-Int - K=27, HIV-RT - K=37)

3.2 Formulas used to calculate performance metrics in Table 1 of the manuscript

R^2 - coefficient of determination is calculated as:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum(e_i^2)}{\sum(y_i - \bar{y})^2}$$

, where TSS - Total Sum of Squares, ESS - Explained Sum of Squares, RSS - Residual Sum of Squares. e_i^2 - error term.

R^2_{adj} - adjusted coefficient of determination is calculated as :

$$R^2_{adj} = 1 - (1 - R^2) * \frac{n-1}{n-(p+1)}$$

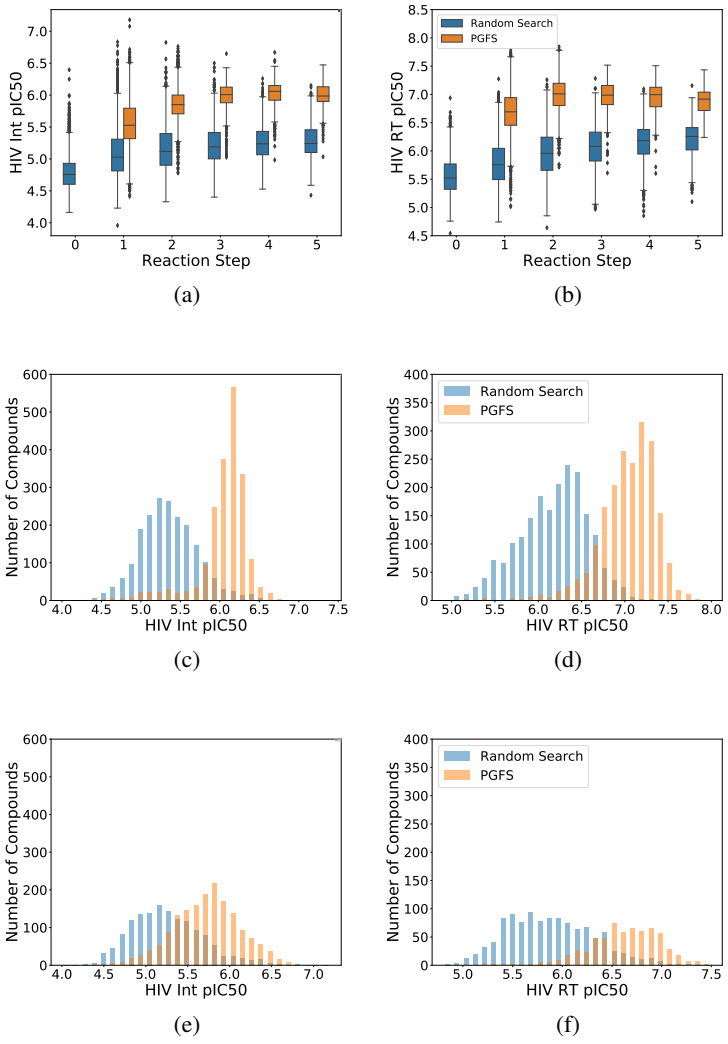
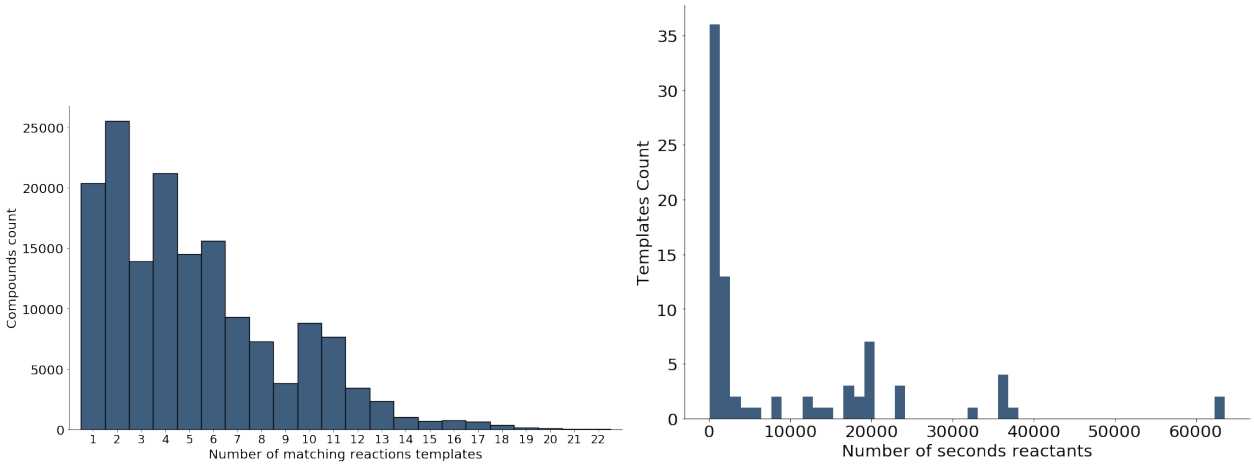
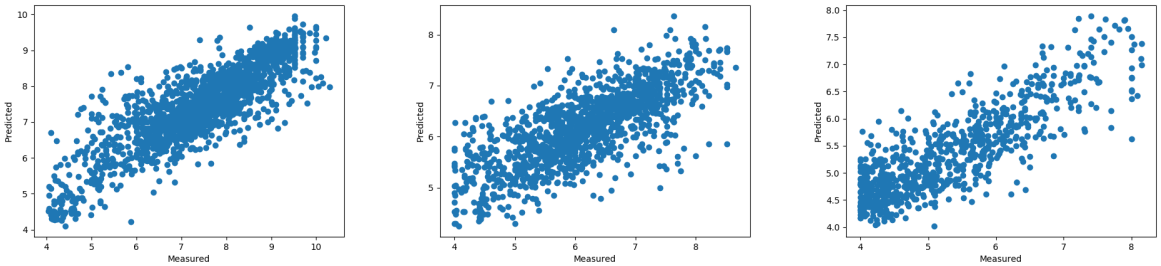


Figure 8: Performance comparison between Random Search and PGFS on two HIV-related QSAR-based scores: HIV-Int and HIV-RT. (a), (b): box plots of the corresponding QSAR-based scores per step of the iterative 5-step virtual synthesis. The first step (Reaction Step =0) in each box plot shows the scores of the initial R1s. (c), (d): distributions of the maximum QSAR-based rewards over 5-step iterations without the AD filtering. (e), (f): distributions of the maximum QSAR-based rewards over 5-step iterations after compounds that do not satisfy AD criteria of the corresponding QSAR model were filtered out from both sets.



(a) (b)

Figure 9: (a) Number of available reaction templates for compounds in the Enamine's building blocks set; (b) Number of available second reactants (R2s) from the Enamine building blocks set for reaction templates that require R2s



(a) (b) (c)

Figure 10: Cross-validation performance of QSAR modeling. Predicted by QSAR models (Y-axis) vs. Measured (X-axis) plots of pIC50 values corresponding to HIV targets, from left to right: CCR5(a), HIV-RT(b), HIV-Integrase(c). Each compound instance was predicted as average prediction of only 5 models that were built while the instance was not present in the training set during the 5-fold cross validation repeated 5 times.

, where R^2 - coefficient of determination, n - number of instances in the dataset, p - number of features used. When calculating cross-validation performance for each separate fold p is number of features used in the final model of that fold after feature selection; in the "aggregated" statistics p is approximated as the average number of features used by 25 models produced by 5-fold CV repeated five times.

3.3 QSAR data collection and curation

As stated in the manuscript the datasets for QSAR modeling were downloaded from ChEMBL25 database. Only structures with available pChEMBL values corresponding to pIC50 were selected. Potential duplicates were removed. Standardization of chemical structures and removal of salts was done using MolVS ([6]). After standardizing entries with the same Canonical SMILES and different pChEMBL (pIC50) values were treated as different measurements of the same compounds, meaning that corresponding pChEMBL(pIC50) values were averaged and only unique compounds were used in further modeling. If a standard deviation of different pIC50 measurements of the same compound exceeded 1.0, the compound was discarded from the dataset. Only compounds with molecular weight (MW) between 100 and 700 were used. Outliers with Z-scores higher than 3.0 calculated using the feature values 'QED', 'MolWt, 'FpDensityMorgan2', 'BalabanJ', 'MolLogP' were discarded as well. In total, 1,719 compounds with unique Canonical SMILES for CCR5, 775 for HIV ingrase and 1,392 for HIV-RT passed all pre-processing and curation procedures. The final curated datasets used for QSAR modeling are reported in the github repository of this project.

4 Experimental setup for GCPN, JTVAE and MSO

The experiments were performed based on the experimental setting detailed in the respective publications: GCPN ([13]), JT-VAE ([3]) and MSO ([12]). The details of which are given below.

1. GCPN: The setup was run on a 32-core machine with a wall-clock time of 30 hrs, with the hyper parameters, training code and the dataset provided by the authors in their publicly available repository.
2. JT-VAE: The setup uses the pre-trained weights provided by the authors. For the bayesian optimization (BO), as suggested by the authors, we train a sparse Gaussian process with 500 inducing points to predict properties (HIV rewards) of molecules. Then, we use five BO iterations along with expected improvement to get the new latent vectors. 50 latent vectors are proposed in each run, we obtain the molecules corresponding to them using the decoder and add them to the training set for the following iteration. Ten such independent runs are performed and the results are combined.
3. MSO: We use the setup corresponding to the GuacaMol ([2]) benchmark as provided by the authors in their work. More precisely, for each HIV reward function, a particle swarm with 200 particles was run for 250 iterations and 40 restarts.

References

[1] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):20, 2015.

[2] Nathan Brown, Marco Fiscato, Marwin HS Segler, and Alain C Vaucher. Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, 59(3):1096–1108, 2019.

[3] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2323–2332, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[4] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pages 3146–3154, 2017.

[5] Greg Landrum. Rdkit: Open-source cheminformatics software. 2016.

[6] Swain Matt. Molvs: Molecule validation and standardization. 2018.

[7] Robert P Sheridan, Bradley P Feuston, Vladimir N Maiorov, and Simon K Kearsley. Similarity to molecules in the training set is a good discriminator for prediction accuracy in qsar. *Journal of chemical information and computer sciences*, 44(6):1912–1928, 2004.

[8] Miha Skalic, Davide Sabbadin, Boris Sattarov, Simone Sciabola, and Gianni De Fabritiis. From target to drug: Generative modeling for the multimodal structure-based ligand design. *Molecular pharmaceutics*, 16(10):4282–4291, 2019.

[9] Anil K Srivastava, Virendra K Srivastava, and Aman Ullah. The coefficient of determination and its adjusted version in linear regression models. *Econometric reviews*, 14(2):229–240, 1995.

[10] Iurii Sushko, Sergii Novotarskyi, Robert Körner, Anil Kumar Pandey, Artem Cherkasov, Jiazhong Li, Paola Gramatica, Katja Hansen, Timon Schroeter, Klaus-Robert Müller, et al. Applicability domains for classification problems: benchmarking of distance to models for ames mutagenicity set. *Journal of chemical information and modeling*, 50(12):2094–2111, 2010.

[11] Alexander Tropsha. Best practices for qsar model development, validation, and exploitation. *Molecular informatics*, 29(6-7):476–488, 2010.

[12] Robin Winter, Floriane Montanari, Andreas Steffen, Hans Briem, Frank Noé, and Djork-Arné Clevert. Efficient multi-objective molecular optimization in a continuous latent space. *Chemical science*, 10(34):8016–8024, 2019.

[13] Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. In *Advances in neural information processing systems*, pages 6410–6421, 2018.