Supplemental information

Multi-objective latent space optimization of generative molecular design models

A N M Nafiz Abeer, Nathan M. Urban, M. Ryan Weil, Francis J. Alexander, and Byung-Jun Yoon
### Note S1 Performance of weighted retraining for six pairs of properties

#### Note S1.1 Effect of weight parameter $k$ with complete dataset

| Pair            | Property     | Training data | Initial Model | Model after 10th retraining with $k = 10^{-1}$ | $k = 10^{-2}$ | $k = 10^{-3}$ | $k = 10^{-4}$ | $k = 10^{-5}$ | $k = 10^{-6}$ |
|-----------------|--------------|---------------|---------------|-----------------------------------------------|---------------|---------------|---------------|---------------|---------------|
|                 |              |               |               | $k = 10^{-1}$ | $k = 10^{-2}$ | $k = 10^{-3}$ | $k = 10^{-4}$ | $k = 10^{-5}$ | $k = 10^{-6}$ |
| logP, SAS       | logP         | 2.4583        | 1.982         | 2.9349                                       | 3.6783        | 4.0592        | 4.8422        | 5.2963        | 6.0017        |
|                 | logP (top 10%) | 4.6321        | 4.2461        | 4.9848                                       | 5.6109        | 5.9546        | 7.018          | 8.0196        | 8.9998        |
|                 | SAS          | 3.0535        | 3.2642        | 2.5012                                       | 2.1818        | 2.047         | 1.8387         | 1.4716         | 1.6566        |
|                 | SAS (top 10%) | 1.9545        | 2.0544        | 1.5789                                       | 1.4669        | 1.4112        | 1.3475         | 1.004          | 1.1522        |
|                 | Structural Diversity | 0.8356        | 0.7904        | 0.7512                                       | 0.7309        | 0.6773        | 0.6205         | 0.6126         |               |
| logP, NP score  | logP         | 2.4583        | 1.982         | 2.9872                                       | 3.2708        | 3.6407        | 4.0105         | 6.0847         | 6.3781        |
|                 | logP (top 10%) | 4.6321        | 4.2461        | 5.0467                                       | 5.5148        | 5.992         | 6.4334         | 8.1291         | 7.9213        |
|                 | NP score      | -1.3122       | -1.0196       | -0.7303                                      | -0.4954       | -0.257        | 0.2244         | 1.5905         | 1.6276        |
|                 | NP score (top 10%) | 0.1482        | 0.1253        | 0.4336                                       | 0.8051        | 1.247         | 2.1975         | 2.3925         | 2.5946        |
|                 | Structural Diversity | 0.8356        | 0.8206        | 0.8231                                       | 0.8295        | 0.8347        | 0.7752         | 0.7529         |               |
| NP score, SAS   | NP score      | -1.3122       | -1.0196       | -0.7919                                      | -0.5822       | -0.2151       | 0.0759         | 0.9727         | 1.314         |
|                 | NP score (top 10%) | 0.1482        | 0.1253        | 0.3403                                       | 0.86          | 1.3508        | 1.7946         | 2.7121         | 2.9791        |
|                 | SAS          | 3.0535        | 3.2642        | 2.8367                                       | 2.6373        | 2.621         | 2.2284         | 2.3073         | 2.7489        |
|                 | SAS (top 10%) | 1.9545        | 2.0544        | 1.6902                                       | 1.525         | 1.5213        | 1.3358         | 1.2841         | 1.4864        |
|                 | Structural Diversity | 0.8356        | 0.8166        | 0.8122                                       | 0.8245        | 0.8301        | 0.8042         | 0.8342         |               |
| logP, DRD2      | logP         | 2.4583        | 1.982         | 2.751                                        | 3.3405        | 3.7416        | 4.3834         | 8.3201         | 7.0397        |
|                 | logP (top 10%) | 4.6321        | 4.2461        | 4.9568                                       | 5.5338        | 5.934         | 7.0267         | 10.7406        | 8.1715        |
|                 | DRD2         | 0.0048        | 0.0066        | 0.0183                                       | 0.0374        | 0.0693        | 0.1477         | 0.5464         | 0.3587        |
|                 | DRD2 (top 10%) | 0.0386        | 0.058         | 0.1572                                       | 0.3028        | 0.5151        | 0.8123         | 0.9971         | 0.9284        |
|                 | Structural Diversity | 0.8356        | 0.8203        | 0.8133                                       | 0.7989        | 0.7812        | 0.6088         | 0.5845         |               |
| DRD2, SAS       | DRD2         | 0.0048        | 0.0066        | 0.0085                                       | 0.021         | 0.0217        | 0.1063         | 0.424          | 0.5253        |
|                 | DRD2 (top 10%) | 0.0386        | 0.058         | 0.0734                                       | 0.1937        | 0.1996        | 0.7922         |               | 1             |
|                 | SAS          | 3.0535        | 3.2642        | 2.7897                                       | 2.3898        | 2.2327        | 1.958          | 1.9466         | 2.1307        |
|                 | SAS (top 10%) | 1.9545        | 2.0544        | 1.7315                                       | 1.4765        | 1.3778        | 1.2675         | 1.1678         | 1.2093        |
|                 | Structural Diversity | 0.8356        | 0.8148        | 0.7932                                       | 0.776         | 0.7311        | 0.6509         | 0.4659         |               |
| NP score, DRD2  | NP score     | -1.3122       | -1.0196       | -0.5209                                      | -0.3084       | 0.0339        | 0.7343         | 1.3774         | 1.8645        |
|                 | NP score (top 10%) | 0.1482        | 0.1253        | 0.0179                                       | 0.0384        | 0.0412        | 0.0859         | 0.2032         | 0.0744        |
|                 | DRD2         | 0.0048        | 0.0066        | 0.0179                                       | 0.0384        | 0.0412        | 0.0859         | 0.2032         | 0.0744        |
|                 | DRD2 (top 10%) | 0.0386        | 0.058         | 0.1454                                       | 0.3052        | 0.3116        | 0.6029         | 0.933          | 0.5707        |
|                 | Structural Diversity | 0.8356        | 0.8497        | 0.8518                                       | 0.8526        | 0.8531        | 0.8268         | 0.823          |               |

**Table S1. Impact of weight parameter $k$ in proposed approach with complete dataset.** Average property values of the training data and 1000 molecules randomly selected from the latent space of the initial model and the model learned after 10th iteration of weighted retraining from starting with the complete dataset. For the top 10% average value, molecules are ranked according to the corresponding property. The structural diversity is computed as the average structural distance based on ECFC4 fingerprints over all pairs in the set of 1000 molecules.
**Note S1.2  Effect of weight parameter $k$ with reduced dataset (top 20% molecules removed)**

| Pair     | Property          | Training data | Initial Model | Model after $10^{th}$ retraining with $k = 10^{-1}$ | $k = 10^{-2}$ | $k = 10^{-3}$ | $k = 10^{-4}$ | $k = 10^{-5}$ | $k = 10^{-6}$ |
|----------|-------------------|---------------|---------------|-----------------------------------------------------|---------------|---------------|---------------|---------------|---------------|
|          |                   |               |               | $k = 10^{-1}$                                      | $k = 10^{-2}$ | $k = 10^{-3}$ | $k = 10^{-4}$ | $k = 10^{-5}$ | $k = 10^{-6}$ |
| logP, SAS| logP (top 10%)    | 2.0875        | 1.6463        | 2.4422                                             | 2.7768        | 3.0227        | 3.2961        | 4.6374        | 5.7802        |
| logP, SAS| logP              | 3.8794        | 4.016         | 4.6657                                             | 4.8221        | 5.0974        | 5.4151        | 6.9437        | 8.3022        |
| logP, SAS| SAS               | 3.2455        | 3.5983        | 3.1535                                             | 3.0309        | 2.8112        | 2.7935        | 2.077         | 1.5667        |
| logP, SAS| SAS (top 10%)     | 2.2257        | 2.2641        | 2.0878                                             | 1.9964        | 1.8365        | 1.8263        | 1.252         | 1.0853        |
|          | Structural Diversity | 0.8474   | 0.8319        | 0.822                                              | 0.82          | 0.8098        | 0.7441        | 0.5793        | 0.5793        |
| logP, NP score| logP (top 10%)  | 2.2063        | 1.7481        | 2.305                                              | 2.4386        | 2.6125        | 2.8047        | 4.1676        | 7.7328        |
| logP, NP score| logP            | 4.0605        | 4.003         | 4.6038                                             | 4.7897        | 5.1549        | 5.3801        | 7.1442        | 10.6178       |
| logP, NP score| NP score        | -1.4805       | -1.1313       | -0.9178                                            | -0.8948       | -0.8349       | -0.6123       | -0.5973       | 0.0025        |
| logP, NP score| NP score (top 10%) | -0.4047  | 0.0479         | 0.1571                                             | 0.1556        | 0.2705        | 0.49          | 0.4912        | 1.1428        |
|          | Structural Diversity | 0.8404  | 0.8332        | 0.8312                                             | 0.8302        | 0.8376        | 0.7971        | 0.697         | 0.697         |
| NP score, SAS| NP score        | -1.4945       | -1.1036       | -0.9777                                            | -0.918        | -0.8978       | -0.8154       | -0.4006       | 0.5817        |
| NP score, SAS| NP score (top 10%) | -0.434   | 0.0183         | 0.1405                                             | 0.1754        | 0.2748        | 0.4079        | 1.2818        | 1.4811        |
| NP score, SAS| SAS             | 3.0918        | 3.4395        | 3.2631                                             | 3.164         | 3.0964        | 3.0436        | 2.3313        | 1.9048        |
| NP score, SAS| SAS (top 10%)    | 2.1067        | 2.1432        | 2.0527                                             | 1.9997        | 1.9438        | 1.7133        | 1.294         | 1.1261        |
|          | Structural Diversity | 0.8411  | 0.8364        | 0.8338                                             | 0.8323        | 0.8375        | 0.8144        | 0.7737        | 0.7737        |
| logP, DRD2| logP (top 10%)    | 2.2028        | 1.8422        | 2.2442                                             | 2.6281        | 2.8932        | 3.0949        | 3.7252        | 9.9092        |
| logP, DRD2| logP              | 4.0734        | 4.1474        | 4.606                                              | 4.8051        | 5.196         | 5.3806        | 6.1771        | 21.8808       |
| logP, DRD2| DRD2             | 0.001         | 0.0028        | 0.0075                                             | 0.0159        | 0.0244        | 0.0216        | 0.0278        | 0.181         |
| logP, DRD2| DRD2 (top 10%)    | 0.0053        | 0.0203        | 0.0607                                             | 0.1337        | 0.2061        | 0.1778        | 0.2206        | 0.8569        |
|          | Structural Diversity | 0.8351  | 0.8333        | 0.8319                                             | 0.831         | 0.8288        | 0.7987        | 0.7148        | 0.7148        |
| DRD2, SAS| DRD2             | 0.0013        | 0.0044        | 0.0084                                             | 0.0084        | 0.0081        | 0.009         | 0.0356        | 0.1094        |
| DRD2, SAS| DRD2 (top 10%)    | 0.0071        | 0.0358        | 0.0732                                             | 0.072         | 0.07         | 0.0779        | 0.3287        | 0.7098        |
| DRD2, SAS| SAS               | 3.1799        | 3.4301        | 3.313                                              | 3.1823        | 2.9907        | 2.9342        | 2.2998        | 1.6688        |
| DRD2, SAS| SAS (top 10%)     | 2.1636        | 2.1501        | 2.0841                                             | 2.0245        | 1.9232        | 1.8324        | 1.5063        | 1.1516        |
|          | Structural Diversity | 0.8452  | 0.8364        | 0.8334                                             | 0.8325        | 0.8315        | 0.8012        | 0.7252        | 0.7252        |
| NP score, DRD2| NP score        | -1.5318       | -1.2088       | -0.9943                                            | -0.8575       | -0.7821       | -0.6855       | -0.4658       | 0.7919        |
| NP score, DRD2| NP score (top 10%) | -0.5758  | -0.1227       | 0.0749                                             | 0.2002        | 0.2949        | 0.414         | 0.8877        | 1.932         |
| NP score, DRD2| DRD2            | 0.0008        | 0.0041        | 0.012                                              | 0.0116        | 0.0204        | 0.02         | 0.0562        | 0.2996        |
| NP score, DRD2| DRD2 (top 10%)   | 0.0041        | 0.035         | 0.1061                                             | 0.0966        | 0.1695        | 0.1692        | 0.4658        | 0.9653        |
|          | Structural Diversity | 0.8279  | 0.8367        | 0.837                                              | 0.8377        | 0.8416        | 0.8313        | 0.7327        | 0.7327        |

**Table S2. Impact of weight parameter $k$ in proposed approach with reduced dataset.** Average property values of the training data and 1000 molecules randomly selected from the latent space of the initial model and the model learned after $10^{th}$ iteration of weighted retraining from starting with the reduced dataset (resulting from removal of top 20% samples based on Pareto front rank). For the top 10% average value, molecules are ranked according to the corresponding property. The structural diversity is computed as the average structural distance based on ECFC4 fingerprints over all pairs in the set of 1000 molecules.
Note S2  Comparison with scalarization-based optimization approach

In this section, we compared our approach with the scalarization baseline of\cite{1} to showcase the improvement of our approach, i.e. the Pareto optimality based ranking of training datapoints. For the scalarization baseline, the ranks of the training molecules are computed based on the weighted combination of their properties of interest where each property is standardized using the mean and standard deviation of the property values of complete training dataset. The weighted objective is defined as the summation of the standardized properties, where we take the negative of SAS property so that molecules with lower SAS gets higher weighted objective.

Table S3 shows the hypervolume measure corresponding to the Pareto front for the 500 molecules suggested over 10 iterations of weighted retraining for the six property pairs considered in this work. For each property, we performed weighted retraining based on the weighted objective as the scalarization baseline to empirically demonstrate the effectiveness of our approach where the rank of the molecules are governed by the Pareto optimality. For all property pairs, our approach outperforms the weighted objective method in most of the cases of \( k \). In particular for lower values of \( k \) (when rank is highly influenced by the property values), the improvement by our approach over the scalarization baseline is more pronounced.

| Property Pair      | Rank Method     | Hypervolume          |
|--------------------|-----------------|----------------------|
|                    |                 | \( k = 10^{-1} \)   | \( k = 10^{-2} \) | \( k = 10^{-3} \) | \( k = 10^{-4} \) | \( k = 10^{-5} \) |
| logP, SAS          | weighted objective | 5.5105 | 7.0641 | 7.2546 | 7.9114 | 7.1866 |
|                    | Ours            | 6.0670 | 6.8845 | 7.3066 | 8.0301 | 10.2210 |
| logP, NP score     | weighted objective | 6.7919 | 10.8326 | 17.2120 | 19.1626 | 19.2328 |
|                    | Ours            | 7.6259 | 11.7250 | 15.5011 | 19.9497 | 22.9996 |
| NP score, SAS      | weighted objective | 2.7127 | 3.4984 | 4.1356 | 4.6484 | 4.0521 |
|                    | Ours            | 2.9911 | 3.6366 | 4.2429 | 4.8607 | 5.3492 |
| logP, DRD2         | weighted objective | 2.6766 | 2.9882 | 3.5095 | 3.7510 | 2.9338 |
|                    | Ours            | 2.2635 | 3.0061 | 3.8571 | 5.4800 | 7.3460 |
| DRD2, SAS          | weighted objective | 0.8345 | 1.0876 | 1.2425 | 1.3752 | 1.3909 |
|                    | Ours            | 1.2639 | 1.3795 | 1.3993 | 1.5283 | 1.6890 |
| NP score, DRD2     | weighted objective | 1.3480 | 2.0743 | 2.0023 | 1.8920 | 2.0412 |
|                    | Ours            | 1.6305 | 1.9528 | 2.1722 | 3.5243 | 3.2791 |

Table S3. Hypervolume of property space dominated by the Pareto front for the 500 molecules suggested over 10 weighted retraining iterations for six property pairs. For the “weighted objective” rank method, the ranking of the molecules at each retraining is based on a scalar quantity which is the summation of standardized properties using the mean and standard deviation for the corresponding properties in the training dataset. On the other hand, our approach utilizes the Pareto front rank based on both properties. The reference point for hypervolume computation is set to the average property value of the complete training dataset.
Note S3  Optimization for three properties

To demonstrate the effectiveness of our approach more than two properties, we considered the latent space optimization for logP, SAS and DRD2 where all properties except SAS are expected to be maximized. We followed the same procedure as the bi-objective cases, and ran the weighted retraining for \(k = 10^{-3}, 10^{-4}, 10^{-5}\). Table S4 the hypervolume of the properties for 500 molecules suggested over 10 weighted retraining iterations using two methods for determining the rank of training molecules: one is based on weighted objective and the other one is our approach which uses the Pareto front rank derived from all three properties. The higher hypervolume in all three values of \(k\) indicates the efficacy of our approach over the weighted objective-based ranking.

| Property Triplet   | Method  | Hypervolume |
|--------------------|---------|-------------|
|                    |         | \(k = 10^{-3}\) | \(k = 10^{-4}\) | \(k = 10^{-5}\) |
| logP, SAS, DRD2    | weighted objective | 2.8937 | 3.9735 | 3.6107 |
|                    | Ours    | 3.7959 | 4.3234 | 5.0013 |

Table S4. Hypervolume of property space dominated by the Pareto front for the 500 molecules suggested over 10 weighted retraining iterations for logP, SAS and DRD2. For the “weighted objective” rank method, the ranking of the molecules at each retraining is based on a scalar quantity which is the summation of properties standardized using the mean and standard deviation for the corresponding properties in the training dataset. On the other hand, our approach utilizes the Pareto front rank based on all three properties. The reference point for hypervolume computation is set to the average property value of the molecules in complete training dataset.

Note S4  Impact of retraining on molecule reconstruction accuracy

Our goal is to make the generative model, JT-VAE more biased to produce molecules with better properties and we achieve this goal by optimizing the JT-VAE parameters in such a way that the model primarily learns to reconstruct molecules that have better properties than others in the training dataset. While the weighted retraining makes the latent space more biased toward the molecules with desired properties, it also means that the retrained model has poor reconstruction performance for molecules with low-scoring properties. We empirically investigated this effect by comparing the reconstruction performance of the pre-trained JT-VAE model as well as the retrained models corresponding to \(k\in{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}}\) for the property pair: logP and SAS. Specifically, we compute the negative log-likelihood (nLL) over the validation dataset for pre-trained JT-VAE and the retrained models corresponding to 5 cases of \(k\). The model with lower nLL has better performance in constructing the molecules of the validation dataset. For each model, we repeat the nLL computation 5 times, and their average and standard deviation are reported in Table S5.

| JT-VAE Model | nLL (5 repetitions) |
|--------------|----------------------|
| Pre-trained  | 1.4816 (0.0040)      |
| \(k = 10^{-1}\) | 3.4983 (0.0037)      |
| \(k = 10^{-2}\) | 4.6442 (0.0033)      |
| \(k = 10^{-3}\) | 6.1943 (0.0051)      |
| \(k = 10^{-4}\) | 11.0966 (0.0031)     |
| \(k = 10^{-5}\) | 20.1343 (0.0051)     |

Table S5. Reconstruction performance on the validation dataset. Negative log-likelihood (nLL) for pre-trained JT-VAE and the retrained models using our proposed approach for \(k\in{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}}\) for property pair: logP and SAS. The metric is computed over the same validation split 5 times, and their average and standard deviation are reported.

The pre-trained model shows the best reconstruction performance (lowest nLL) since it was trained to reconstruct wide range of molecules, unlike the retrained models. As we retrain the generative model with the smaller value of \(k\) which means more weights on the molecules with better properties, the retrained model becomes biased towards high-scoring molecules. Since validation split contains molecules with a broad range of properties, for the retrained
models with smaller $k$, we are seeing higher nLL which corresponds to the poor reconstruction accuracy for the molecules within the validations dataset.

**Note S5  Comparison with MARS$^2$**

In this section, we compare our approach with Xie et al.$^2$’s proposed sampling-based technique – MARS. Since this approach trains the molecule generator unit, i.e., molecular graph editing model in an online fashion, it has a certain advantage over the deep generative models (such as the VAE-based models that we have focused on in the study) which are trained in self-supervised fashion without considering the properties of the training molecules. We have used the implementation of$^2$ to generate 500 molecules for the same combination of properties considered in our work. For the scoring function of logP and NP score, the property value is standardized using the same mean and standard deviation that we used for weighted objective in Sections **Note S2** and **Note S3**. We also have followed their implementation to normalize the SAS into $[0, 1]$ interval where the higher score corresponds to the lower SAS.

| Properties | Hypervolume | MARS$^2$ | \(k = 10^{-5}\) | \(k = 10^{-4}\) |
|------------|-------------|---------|----------------|----------------|
| logP, SAS  | 12.6780     | 10.2210 | 8.0301         |                |
| logP, NP score | 45.0480   | 22.9996 | 19.9497        |                |
| NP score, SAS | 3.3847    | 5.3492  | 4.8607         |                |
| logP, DRD2 | 6.6405      | 7.3460  | 5.4800         |                |
| DRD2, SAS  | 1.5571      | 1.6890  | 1.5283         |                |
| NP score, DRD2 | 3.3672    | 3.2791  | 3.5243         |                |
| logP, SAS, DRD2 | 5.1056   | 5.0013  | 4.3234         |                |

Table S6. Comparison between MARS and our approach for different combinations of properties. Hypervolume of property space dominated by the Pareto front for the 500 molecules suggested using MARS$^2$ and our proposed approach of 10 weighted retraining iterations with \(k \in \{10^{-5}, 10^{-4}\}\) for different combination of properties. The reference point for hypervolume computation is set to the average property value of the molecules in complete training dataset.

Table S6 shows the hypervolume metric for 500 molecules suggested by MARS and our Pareto front rank approach with \(k \in \{10^{-5}, 10^{-4}\}\). For property pairs – (logP, DRD2), (DRD2, SAS), (NP score, SAS) and (NP score, DRD2) our proposed approach produced more Pareto-optimal molecules than MARS. On the other hand, the MARS approach tends to show superior performance over ours when logP is one of the properties. As mentioned in$^2$, it is easier for MARS to optimize for logP by generating larger molecules whereas the generative models like JT-VAE in our work struggle to do so. However, our proposed approach enables the JT-VAE model to outperform MARS for the property pair logP and DRD2. We speculate that, for challenging objectives like DRD2, MARS loses its advantage in logP because simply adding more molecular units does not translate to increased DRD2 inhibition probability. This is also reflected in the case of the property triplet where MARS and our approach perform similarly.

**References**

1. Tripp, A., Daxberger, E., and Hernández-Lobato, J. M. (2020). Sample-efficient optimization in the latent space of deep generative models via weighted retraining. Advances in Neural Information Processing Systems 33, 11259–11272.

2. Xie, Y., Shi, C., Zhou, H., Yang, Y., Zhang, W., Yu, Y., and Li, L. (2021). Mars: Markov molecular sampling for multi-objective drug discovery. In: International Conference on Learning Representations.