Cloud-Based Real-Time Molecular Screening Platform with MolFormer

Brian Belgodere*, Vijil Chenthamarakshan*, Payel Das*, Pierre Dognin*, Toby Kurien*, Igor Melnyk*, Youssef Mroueh*, Inkit Padhi*, Mattia Rigotti*, Jarret Ross*✉, Yair Schiff*, and Richard A. Young*

IBM Research

Abstract. With the prospect of automating a number of chemical tasks with high fidelity, chemical language processing models are emerging at a rapid speed. Here, we present a cloud-based real-time platform that allows users to virtually screen molecules of interest. For this purpose, molecular embeddings inferred from a recently proposed large chemical language model, named MolFormer, are leveraged. The platform currently supports three tasks: nearest neighbor retrieval, chemical space visualization, and property prediction. Based on the functionalities of this platform and results obtained, we believe that such a platform can play a pivotal role in automating chemistry and chemical engineering research, as well as assist in drug discovery and material design tasks. A demo of our platform is provided at www.ibm.biz/molecular_demo.

Keywords: Molecular screening · Drug discovery · Cloud platform

1 Introduction

Machine learning (ML) offers high throughput material exploration that is more efficient than high-cost quantum chemical/empirical force-field calculations and wet lab evaluations. In this work, we present a cloud-based platform for real-time virtual screening of molecules, which uses a general-purpose deep learning model of large organic small molecule libraries. Specifically, our Molecular Explorer Platform builds on our previous work "MolFormer", a large, masked chemical language model trained on over 1.1 billion molecular string representations known as SMILES (see [11] for details). MolFormer provides representations for molecules that we showcase here in a platform enabling neighbor search, chemical space visualization, and property prediction for molecules of interest.

2 Real-Time Screening Platform

Given a backend dataset, such as PubChem [4] or FlavorDB [1], we start by embedding this database through MolFormer and obtain a latent representation

[*] Equal contribution, ordered alphabetically. Contact author: rossja@us.ibm.com

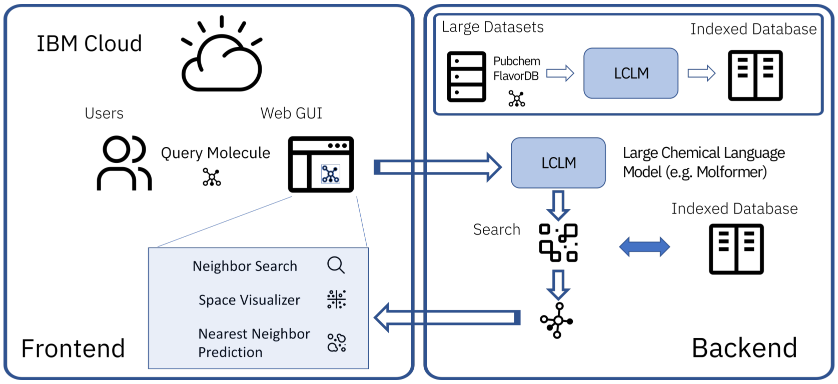
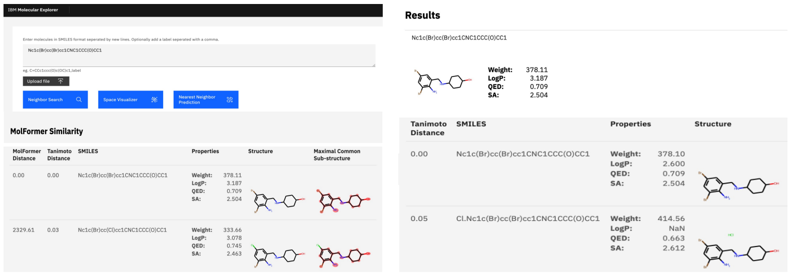


Fig. 1: Diagram of our Molecular Explorer Platform.

of 768 dimensions. To index the database for nearest neighbor search, we start by reducing the dimensionality of MolFormer representations using Discrete Cosine Transform to 128 dimensions. We then leverage the approximate nearest neighbor search library HNSWlib [8]. with hyperparameters calibrated so that retrieval would be faster than 10 milliseconds per query with a recall of 0.99.

Our molecular platform consists of a frontend GUI that enables 3 critical molecule screening functionalities: 1) neighbor search, 2) visualizing latent space of molecules using t-SNE visualization in 2D, and 3) nearest neighbor property prediction using Sklearn [10] for moderately sized datasets and FAISS [3] for large-scale predictions. User queries are provided in the form of a line separated list or `.txt` file of molecule SMILES strings.An implementation of MolFormer running on OpenShift on IBM Cloud enables real-time feedforward embedding of SMILES strings, which are normalized using the RDKit library [5,6]. The obtained MolFormer representation is subsequently used to query the indexed backend database, which returns the user provided N nearest neighbors along with molecular properties, such as logP, QED, and weight, which are computed on-the-fly using RDKit. Optional call to PubChem's similarity search API is provided in our user interface allowing the user to compare it to MolFormer similarity.If a user provides property labels for each SMILES string, such as toxicity or flavor (see use case 2), the molecular platform enables visualization of the embedding space color-coded by labels in t-SNE 2-dimensional space. Finally, nearest neighbor prediction functionalities using known properties of the backend index database are also provided, along with predictions for these properties of query molecules and graphical visualization of the results.

Use Case 1: Similarity search among known drug molecules. A typical task that arises in molecule screening/discovery is to identify similar molecules in existing chemical libraries. This is a frequent use case for medicinal chemists, for example. Our molecule explorer platform allows users to retrieve similar molecules from PubChem using the PubChem API [4] and MolFormer embed-


IBM Molecular Explorer

Enter molecules in SMILES format seperated by new lines. Optionally add a label seperated with a comma.

Nc1c(Br)cc(Br)cc1CNC1CCC(O)CC1

eg. C=CCc1ccc(O)c(OC)c1,label

MolFormer Similarity

MolFormer Distance	Tanimoto Distance	SMILES	Properties	Structure	Maximal Common Sub-structure
0.00	0.00	Nc1c(Br)cc(Br)cc1CNC1CCC(O)CC1	Weight: 378.11 LogP: 3.187 QED: 0.709 SA: 2.504		
2329.61	0.03	Nc1c(Br)cc(Cl)cc1CNC1CCC(O)CC1	Weight: 333.66 LogP: 3.078 QED: 0.745 SA: 2.463		

Results

Nc1c(Br)cc(Br)cc1CNC1CCC(O)CC1

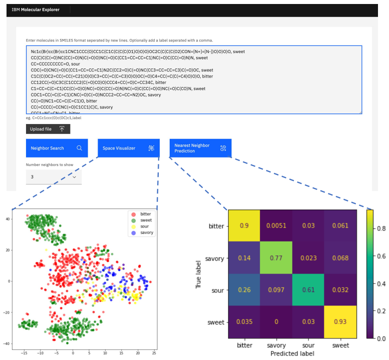
Weight: 378.11
LogP: 3.187
QED: 0.709
SA: 2.504

Tanimoto Distance	SMILES	Properties	Structure
0.00	Nc1c(Br)cc(Br)cc1CNC1CCC(O)CC1	Weight: 378.10 LogP: 2.600 QED: 0.709 SA: 2.504	
0.05	Cl.Nc1c(Br)cc(Br)cc1CNC1CCC(O)CC1	Weight: 414.56 LogP: NaN QED: 0.663 SA: 2.612	

Fig. 2: Use Case 1: Nearest neighbor search in large chemical embedding space.

dings. To achieve this, we index over 100 Million molecules from pubchem embedded in the MolFormer latent space. As an example, we show the neighbor retrieval results for known drug molecules (Table S4 from [2]) obtained using the platform in Figure 2. The maximal common subgraph of the query molecule and closest molecules are also shown allowing a user to understand the key differences between the query molecules and its closest neighbors.

Use Case 2: Flavor molecules screening. The molecule explorer platform also allows a user to upload a set of molecules along with a their corresponding class (property) labels and visualize their chemical space. The user can visually explore the t-SNE [7] representation of those molecules obtained using MolFormer embeddings and check if the resulting chemical space captures the distribution of class labels for a particular application. Alternatively, a k-NN classifier can be trained on the MolFormer embeddings and performance characteristics of the classifier can be visualized as a confusion matrix. We show the application of these techniques to molecules with different flavor descriptions from [1]. The flavor database consists of 25,595 individual flavor molecules with up to 43 different attributes. 4 basic flavors were chosen for evaluation; bitter, sweet, sour, and savory. Figure 3 shows that our chemical space map captures the different flavors and provides excellent predictive performance.


IBM Molecular Explorer

Enter molecules in SMILES format seperated by new lines. Optionally add a label seperated with a comma.

```
Nc1c(Br)cc(Br)cc1CNC1CCC(O)CC1C(C1C(C(C(C(O1)O)O)O)OC2C(C(C(C(O2)CON=[N+]=[N-])O)O)O)O, sweet
CC(C)C(C(=O)NC(CC(=O)N)C(=O)O)NC(=O)C(CC1=CC=CC=C1)NC(=O)C(CC(=O)N)N, sweet
CC=CCCCCCCCC=O, sour
COC(=O)CNC(=O)C(CC1=CC=CC=C1)N2C(CC2=O)C(=O)NC(CC3=CC=CC=C3)C(=O)OC, sweet
C1C(C(OC2=CC(=CC(=C21)O)O)C3=CC(=C(C=C3)O)O)OC(=O)C4=CC(=C(C(=C4)O)O)O, bitter
CC12CC(=O)C3C(C1CCC2(C(=O)CO)O)CCC4=CC(=O)C=CC34C, bitter
C1=CC=C(C=C1)CC(C(=O)O)NC(=O)C(CC(=O)N)NC(=O)C(CC(=O)O)NC(=O)C(CO)N, sweet
COC1=CC(=C(C=C1)CNC(=O)C(=O)NCCC2=CC=CC=N2)OC, savory
CC(=O)NC1=CC=C(C=C1)O, bitter
CC(=CCCC(=CCNC(=O)C1CC1)C)C, savory
CCC1=NC=CN=C1, bitter
```

eg. C=CCc1ccc(O)c(OC)c1,label

Number neighbors to show: 3

	bitter	savory	sour	sweet
bitter	0.9	0.0051	0.03	0.061
savory	0.14	0.77	0.023	0.068
sour	0.26	0.097	0.61	0.032
sweet	0.035	0	0.03	0.93

Fig. 3: Use Case 2: Visualization of unsupervised MolFormer Embeddings in t-SNE space and separation of flavor molecules in that space.

Use Case 3: Drug-like molecules screening. Lastly, we predict the conformity to the RO5 (Lipinski rule of five) of 1.8M molecules out of ∼2M from the CheMBL dataset [9], which presented SMILES representations. A k-NN classifier was trained on 1.44M MolFormer embeddings with the FAISS library [3] and used to predict RO5 violations of 360k held-out molecules based on their neighbors, resulting in a classification accuracy of 90% (see Fig. 4). We then predicted HBA (hydrogen bond acceptor) and HBD (hydrogen bond donor) on the same split by averaging the HBA and HBD values of $k = 3$ nearest neighbors, obtaining high coefficients of determination of $R^2 = 0.926$ (see Fig. 4).

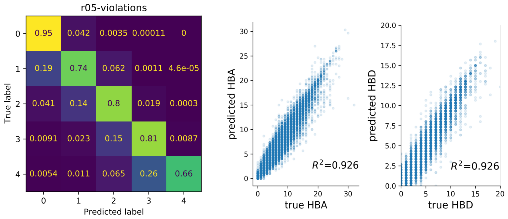


Fig. 4: Use Case 3: Retrieval of r05-violations of 1.8M drug-like molecules with 1-NN gives an average holdout prediction accuracy of 0.90 (left). HBA and HBD are also predicted with high accuracy ($R^2 = 0.926$ for both) by $k = 3$ NN-Regression (right).

References

1. Garg, N., Sethupathy, A., Tuwani, R., NK, R., Dokania, S., Iyer, A., Gupta, A., Agrawal, S., Singh, N., Shukla, S., Kathuria, K., Badhwar, R., Kanji, R., Jain, A., Kaur, A., Nagpal, R., Bagler, G.: FlavorDB: a database of flavor molecules. Nucleic Acids Research **46**(D1), D1210–D1216 (10 2017)
2. Hoffman, S.C., Chenthamarakshan, V., Wadhawan, K., Chen, P.Y., Das, P.: Optimizing molecules using efficient queries from property evaluations. Nature Machine Intelligence **4**(1), 21–31 (Jan 2022). https://doi.org/10.1038/s42256-021-00422-y
3. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Transactions on Big Data **7**(3), 535–547 (2019)
4. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., Zaslavsky, L., Zhang, J., Bolton, E.E.: PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Research **49**(D1), D1388–D1395 (11 2020). https://doi.org/10.1093/nar/gkaa971, https://doi.org/10.1093/nar/gkaa971
5. Landrum, G.: RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling (2013)
6. Landrum, G.: Rdkit: Open-source cheminformatics. https://www.rdkit.org (2013)
7. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(86), 2579–2605 (2008), http://jmlr.org/papers/v9/vandermaaten08a.html
8. Malkov, Y.A., Yashunin, D.A.: Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. IEEE transactions on pattern analysis and machine intelligence **42**(4), 824–836 (2018)

9. Mendez, D., Gaulton, A., Bento, A.P., Chambers, J., De Veij, M., Félix, E., Magariños, M., Mosquera, J., Mutowo, P., Nowotka, M., Gordillo-Marañón, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C., Segura-Cabrera, A., Hersey, A., Leach, A.: ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Research **47**(D1), D930–D940 (11 2018)
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
11. Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., Das, P.: Do large scale molecular language representations capture important structural information? (2021)