Toronto Bioinformatics Hackathon: September 19th 2025 - September 21st 2025

Group 9: Ionizable lipid screening for efficient nucleic acid delivery

0. Abstract

Lipid nanoparticles are promising delivery mechanisms for a variety of drugs into the cytoplasm.[1] They have been employed to deliver a wide range of nucleic acids, most notably the mRNA in Pfizer's and Moderna's COVID-19 vaccines.[2] The main component and most important factor in improving delivery efficiency are the ionizable lipids.[3] These lipids are positively charged during the lipid nanoparticle formulation, but are neutral at physiological pH. This positive charge under acidic pH values promotes attraction with the negatively charged nucleic acid backbone, hence improving encapsulation efficiency.[4] Moreover, protonation aids with endosomal escape, since the protonated ionizable lipid interacts with the anionic endosomal membrane.[5] However, it is essential that the lipids remain neutral at physiological levels to avoid toxicity and immunogenicity, and to therefore increase circulation time and cellular uptake.[6] Hence, there is a narrow interval of acceptable pKas, usually between 6.1 and 6.7. [7, 8]

Although lipids are designed rationally, they must still be synthesized to be screened, a process that can take months and lots of pricey reagents. Our project would allow for a more efficient screening of ionizable lipids by creating a model to predict their pKa, encapsulation efficiency, size and polydispersity index. Hence, researchers would only need to synthesize the lipids that the model considers a hit, greatly reducing the time and resources spent on low-quality lipids.

Firstly, a publicly available database was created from patents and publication. Afterwards, numerous supervised learning approaches were explored.

1.Methods

1.1. Database Building Methods

The following intellectual property documents/patents were read: CA2998810A1, WO202006137A1. The reported polydispersity index (PDI), encapsulation efficiency (EE%), size (nm) and pKa were extracted. The skeleton chemical structures were manually converted to the Simplified Molecular Input Line Entry System (SMILES) using OSRA (https://cactus.nci.nih.gov/cgi-bin/osra/index.cgi) and verifying the output.

1.2. Supervised Learning Methods

Models were attempted using a variety of methods, including the following: random forest regression, k-nearest neighbours regression, XGBoost regression, support vector regression.

1.3. Website

The website was built with HTML, CSS and Javascript. It is accessible on Github.

2. Results

2.1. Database

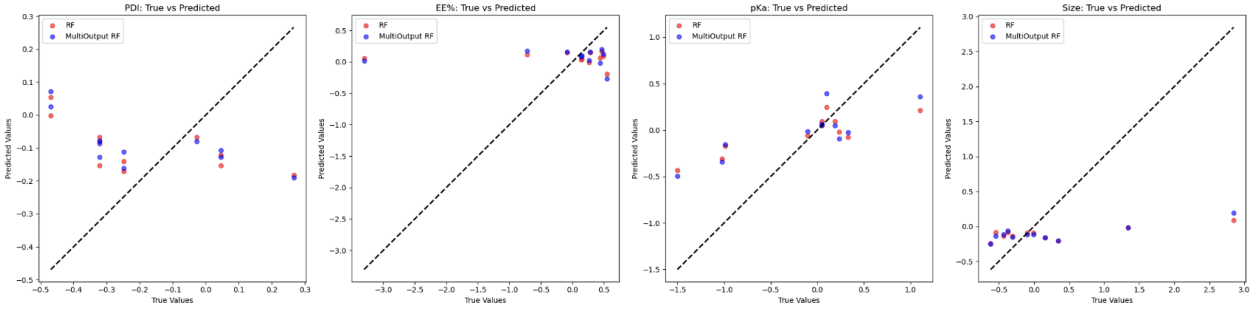
The database was created using patent CA2998810A1. It has 120 entries and can be accessed through github. It contains the SMILEs, Encapsulation Efficiency (%), Polydispersity Index, acid dissociation constant and size (nm). .

2.2. Polydispersity Index, Encapsulation Efficiency, Size and pKa predictor

2.2.1.RandomForestRegressor, with and without MultiOutputRegressor

A random Forest Regressor (regular and MultiOutputRegressor) was trained using RandomizedSearch and then Gridsearch close to the best RandomSearch objectives (see github).

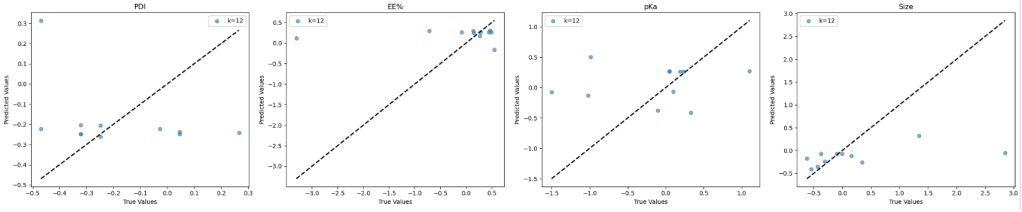
```
Best RF Params: {'max_depth': 40, 'max_features': 'log2', 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimator
s': 750}
Best RF CV R²: -0.0830799859234658

Best MultiOutput RF Params: {'estimator__max_depth': 5, 'estimator__max_features': 'log2', 'estimator__min_samples_lea
f': 1, 'estimator__min_samples_split': 9, 'estimator__n_estimators': 500}
Best MultiOutput RF CV R²: -0.05540769049020842
```



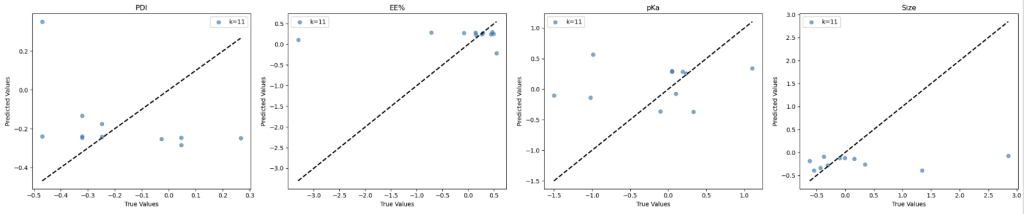
2.2.3.K-Nearest-Neighbors Regressor

K-Nearest-Neighbors Regressor was run with different values for K

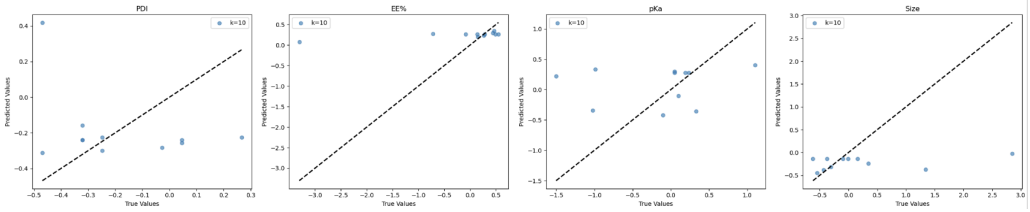
```
KNN with k=12
  PDI: R²=-1.078, MSE=0.106
  EE%: R²=-0.091, MSE=1.232
  pKa: R²=-0.180, MSE=0.590
  Size: R²=0.045, MSE=0.926
```



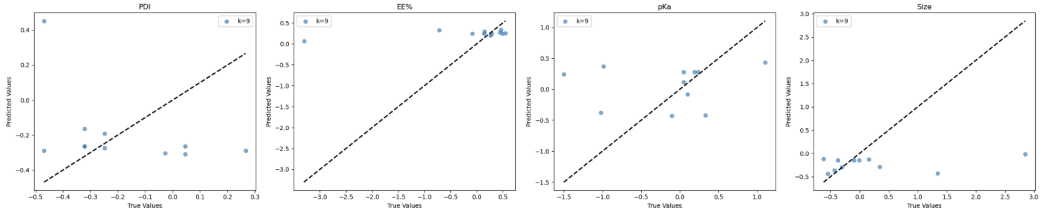

```
KNN with k=11
  PDI: R²=-1.302, MSE=0.117
  EE%: R²=-0.088, MSE=1.228
  pKa: R²=-0.174, MSE=0.587
  Size: R²=-0.152, MSE=1.117
```



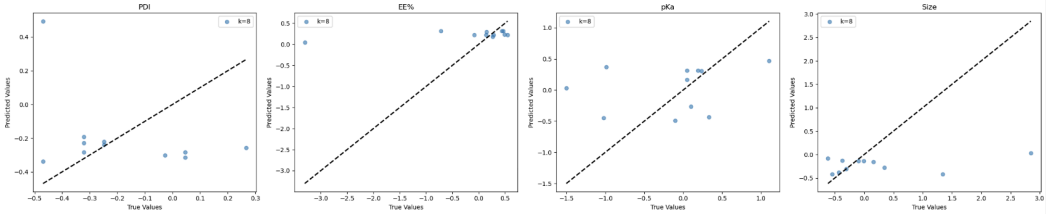
KNN with k=10
  PDI: R²=−1.389, MSE=0.122
  EE%: R²=−0.024, MSE=1.156
  pKa: R²=−0.165, MSE=0.583
  Size: R²=−0.116, MSE=1.082



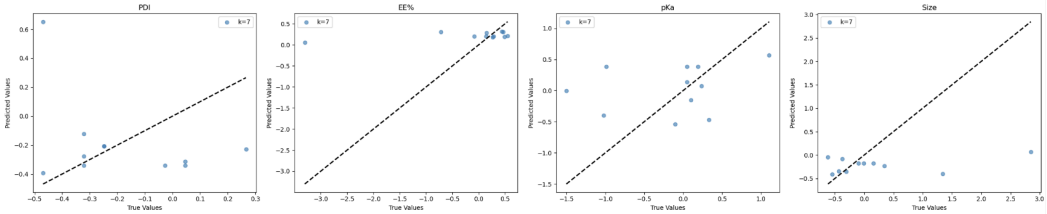
KNN with k=9
  PDI: R²=−1.718, MSE=0.138
  EE%: R²=−0.031, MSE=1.164
  pKa: R²=−0.183, MSE=0.592
  Size: R²=−0.138, MSE=1.103



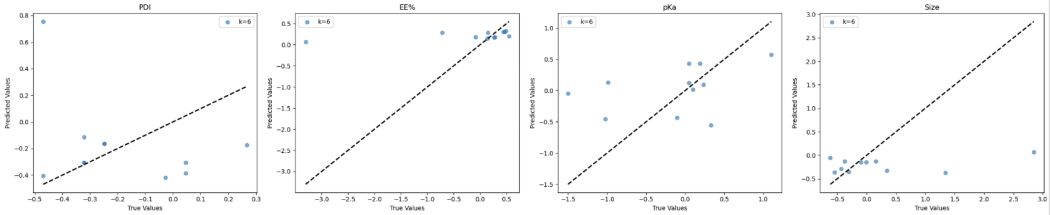
KNN with k=8
  PDI: R²=−1.786, MSE=0.142
  EE%: R²=−0.020, MSE=1.151
  pKa: R²=−0.067, MSE=0.534
  Size: R²=−0.117, MSE=1.082



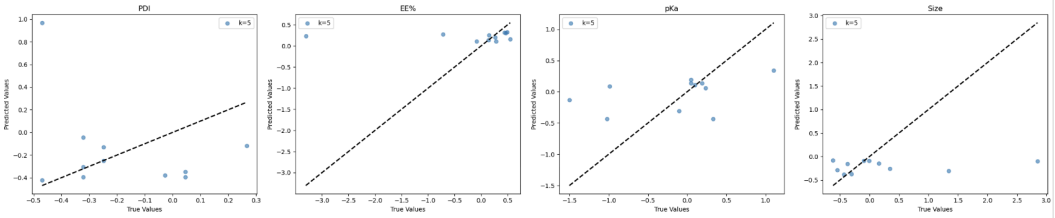
KNN with k=7
  PDI: R²=−2.446, MSE=0.175
  EE%: R²=−0.021, MSE=1.152
  pKa: R²=−0.070, MSE=0.535
  Size: R²=−0.096, MSE=1.063



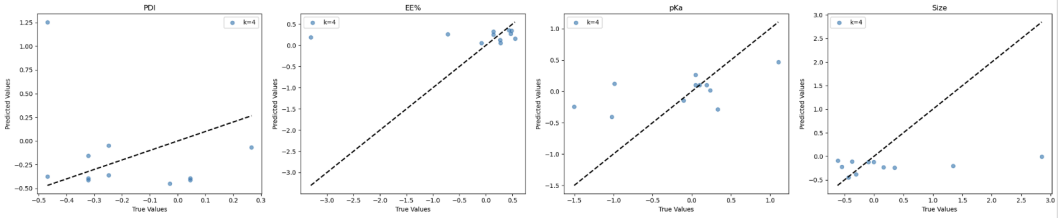
KNN with k=6
  PDI: R²=−2.956, MSE=0.201
  EE%: R²=−0.019, MSE=1.150
  pKa: R²=0.072, MSE=0.464
  Size: R²=−0.091, MSE=1.058



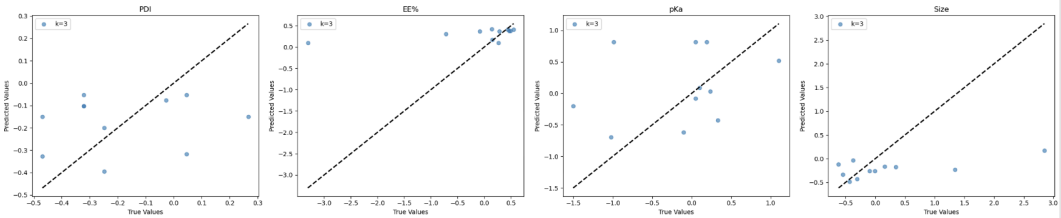
KNN with k=5
  PDI: R²=-3.969, MSE=0.253
  EE%: R²=-0.114, MSE=1.258
  pKa: R²=0.153, MSE=0.424
  Size: R²=-0.152, MSE=1.117



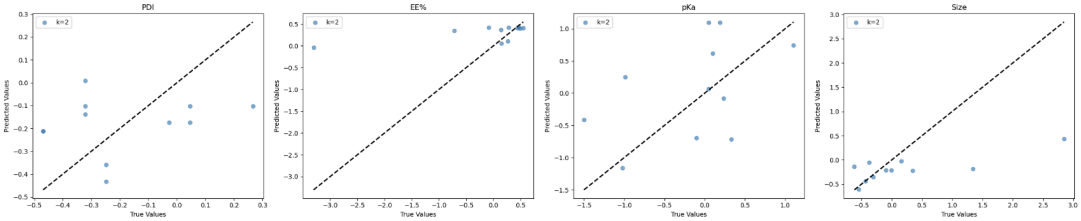
KNN with k=4
  PDI: R²=-5.710, MSE=0.341
  EE%: R²=-0.089, MSE=1.229
  pKa: R²=0.260, MSE=0.370
  Size: R²=-0.079, MSE=1.046



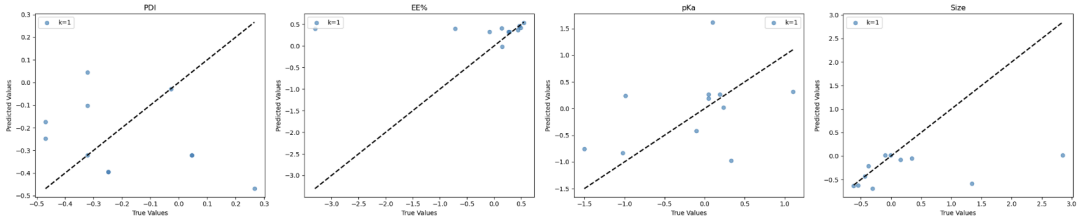
KNN with k=3
  PDI: R²=-0.130, MSE=0.057
  EE%: R²=-0.046, MSE=1.181
  pKa: R²=-0.318, MSE=0.659
  Size: R²=0.013, MSE=0.957



KNN with k=2
  PDI: R²=-0.066, MSE=0.054
  EE%: R²=0.017, MSE=1.109
  pKa: R²=-0.203, MSE=0.602
  Size: R²=0.165, MSE=0.809

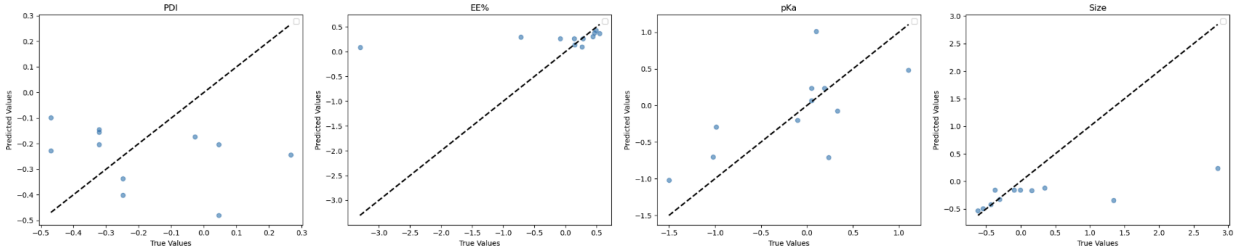


KNN with k=1
  PDI: R²=-1.094, MSE=0.107
  EE%: R²=-0.227, MSE=1.385
  pKa: R²=-0.262, MSE=0.631
  Size: R²=-0.137, MSE=1.102



2.2.4. MultiOutputRegressor Gradient Boosting

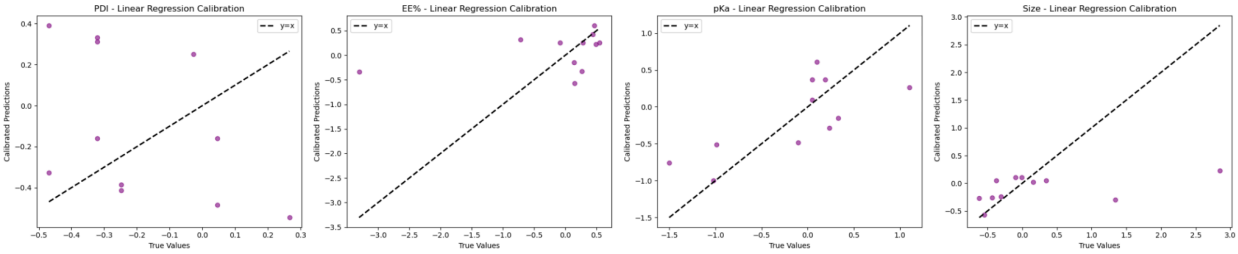
```
PDI: R²=-0.645, MSE=0.084
EE%: R²=-0.029, MSE=1.162
pKa: R²=0.429, MSE=0.286
size: R²=0.060, MSE=0.912
```



2.2.5. Best method: Support Vector Machines + Linear Regression + Linear Transformation

Firstly, Support Vector Machines was used for a regression. Upon noticing a linear trend in the true vs predicted plot for it, we took the linear regression of this line and applied a linear transformation to transform this line into y=x. This was used to make the obtain results, which can be seen below.

```
PDI - Linear regression calibration: R²=-3.856, RMSE=0.497
EE% - Linear regression calibration: R²=0.104, RMSE=1.006
pKa - Linear regression calibration: R²=0.540, RMSE=0.480
Size - Linear regression calibration: R²=0.061, RMSE=0.954
```



3. Conclusion and Website

The above work is a proof of concept for future developments on the screening of ionizable lipids for lipid nanoparticles. We built what is, to our knowledge, the first publicly available dataset for lipid nanoparticles with their PDI, EE%, size and pKa. The results were acceptable given the size of the dataset. Future direction would include exploring and scraping more data from literature and patents, as well as predicting viability of theoretical ionizable lipids in order to discover new potentially viable lipids. These further developments could greatly reduce the time and resources spent on the development of new ionizable lipids, and help advance various therapies that depend on lipid nanoparticles for delivery, including gene therapy and mRNA vaccines.

4. Repository

https://github.com/hackbio-ca/lipid-nanoparticle-screening/tree/main

5. Bibliography

[1] Alfutaimani, A. S., Alharbi, N. K., S Alahmari, A., A Alqabbani, A., & Aldayel, A. M. (2024). Exploring the landscape of Lipid Nanoparticles (LNPs): A comprehensive review

of LNPs types and biological sources of lipids. *International Journal of Pharmaceutics: X*, *8*(100305), 100305. https://doi.org/10.1016/j.ijpx.2024.100305

[2]Hou, X., Zaks, T., Langer, R., & Dong, Y. (2021). Lipid nanoparticles for mRNA delivery. *Nature Reviews. Materials*, *6*(12), 1078–1094. https://doi.org/10.1038/s41578-021-00358-0

[3]Han, X., Zhang, H., Butowska, K., Swingle, K. L., Alameh, M.-G., Weissman, D., & Mitchell, M. J. (2021). An ionizable lipid toolbox for RNA delivery. *Nature Communications*, *12*(1), 7233. https://doi.org/10.1038/s41467-021-27493-0

[4]Schober, G. B., Story, S., & Arya, D. P. (2024). A careful look at lipid nanoparticle characterization: analysis of benchmark formulations for encapsulation of RNA cargo size gradient. *Scientific Reports*, *14*(1), 2403. https://doi.org/10.1038/s41598-024-52685-1

[5]Semple, S. C., Akinc, A., Chen, J., Sandhu, A. P., Mui, B. L., Cho, C. K., Sah, D. W. Y., Stebbing, D., Crosley, E. J., Yaworski, E., Hafez, I. M., Dorkin, J. R., Qin, J., Lam, K., Rajeev, K. G., Wong, K. F., Jeffs, L. B., Nechev, L., Eisenhardt, M. L., … Hope, M. J. (2010). Rational design of cationic lipids for siRNA delivery. *Nature Biotechnology*, *28*(2), 172–176. https://doi.org/10.1038/nbt.1602

[6]Schober, G. B., Story, S., & Arya, D. P. (2024). A careful look at lipid nanoparticle characterization: analysis of benchmark formulations for encapsulation of RNA cargo size gradient. *Scientific Reports*, *14*(1), 2403. https://doi.org/10.1038/s41598-024-52685-1

[7]Jayaraman, M., Ansell, S. M., Mui, B. L., Tam, Y. K., Chen, J., Du, X., Butler, D., Eltepu, L., Matsuda, S., Narayanannair, J. K., Rajeev, K. G., Hafez, I. M., Akinc, A., Maier, M. A.,

Tracy, M. A., Cullis, P. R., Madden, T. D., Manoharan, M., & Hope, M. J. (2012). Maximizing the potency of siRNA lipid nanoparticles for hepatic gene silencing in vivo. *Angewandte Chemie (International Ed. in English)*, *51*(34), 8529–8533. https://doi.org/10.1002/anie.201203263

[8]Simonsen, J. B., & Larsson, P. (2025). A perspective on the apparent pKa of ionizable lipids in mRNA-LNPs. *Journal of Controlled Release: Official Journal of the Controlled Release Society*, *384*(113879), 113879. https://doi.org/10.1016/j.jconrel.2025.113879

6. Further References

Benenato, K. E., Kumarasinghe, E. S., & Cornebise, M. (2017). Compounds and Compositions for Intracellular Delivery of Therapeutic Agents (International Publication Number WO 2017/049245). World Intellectual Property Bureau.