# Quantum Chemical Properties-Enhanced Hybrid Machine Learning Approach for Explainable CRISPR-Cas9 sgRNA Efficiency Prediction

Zhenxuan Zhao, Nguyen Quoc Khanh Le, Matthew Chin Heng Chua

**ABSTRACT**

This study explores the enhancement of sgRNA efficiency prediction in CRISPR-Cas9 systems by incorporating quantum chemical properties into machine learning models. The research develops a novel feature matrix that integrates conventional features such as GC content, melting temperature, and minimum free energy with quantum chemical properties like HOMO-LUMO gap, hydrogen bonding energy, and stacking interaction. An iterative random forest (iRF) model with model selection features and a feedforward neural network model that has better predictive ability were combined to analyze the effectiveness of the novel feature matrix. Explainable Artificial Intelligence (XAI) approaches are also employed to improve the interpretability of the models. The study highlights the importance of quantum chemical properties, proving their critical roles in understanding sgRNA efficiency. The iRF model, trained on E. coli dataset, demonstrated strong predictive capabilities in forecasting sgRNA efficiency, offering valuable insights into the biological mechanisms of CRISPR-Cas9 system. The neural network model further enhanced predictive performance by capturing more complex non-linear relations between variables. Despite these successes, both models showed poor performance when applied to the H. sapiens dataset, indicating a limitation in their ability to generalize across species.

**Keywords:** CRISPR-Cas9, sgRNA efficiency, quantum chemical properties, machine learning

## 1 Introduction

### 1.1 Backgrounds & Project Objectives

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) is a groundbreaking tool in gene editing technology that allows scientists to perform site-specific modifications to the genome [1]. The process was originally discovered as a part of bacteria adaptive immune system that protects them against bacteriophages [23]. Scientists harnessed this natural mechanism to create a powerful tool for editing the target genomes of various organisms with precision. Guided by specially designed sgRNA (single guide RNA) sequences, the Cas9 (CRISPR associated enzyme) nuclease found in bacteria can generate breaks in double-stranded DNA and cleave each strand of DNA at the target site [1, 24].

However, the success of CRISPR-Cas9-based gene editing is highly dependent on the design of the sgRNA. While CRISPR-Cas9 can tolerate minor nucleotide mismatches, its flexibility can lead to off-target effects, where the sgRNA guides Cas9 enzyme to unintended genome locations, potentially causing undesirable gene modifications [25, 26]. Therefore, designing sgRNAs that maximize on-target activity while minimizing off-target effects is critical for successful and precise gene editing [2, 3]. And understanding the features that affect sgRNA efficiency is essential for efficient sgRNA design.

Several factors are commonly known to impact sgRNA efficiency, including overall nucleotide usage, melting temperature, GC content, RNA secondary structure, and position-related features within the target sequence [4, 5]. These conventional features are combined with machine learning models to train models that predict sgRNA efficiency. Recent advances have introduced the integration of quantum chemical properties into machine-learning models to further enhance predictive accuracy [6, 27]. Quantum chemical properties, such as HOMO-LUMO gap (Highest Occupied Molecular Orbital - Lowest Unoccupied Molecular Orbital energy gap), hydrogen bonding energy, stacking interactions offer insights into the molecular interactions and electron energetics at play in the CRISPR-Cas9 system. By incorporating these novel features, models can highlight the influence of more subtle molecular interactions of the binding of target DNA and sgRNA sequence [6].

Machine-learning models are often utilized to offer guidance to efficient in silico design of sgRNA. They include white-box models like regressions and tree-based models, or black-box models like different variants of neural networks. White box models highlight the idea of explainable AI (XAI) that improve the

interpretability of the model and measure the contributions of specific features. Black box models can capture more complex interactions of features by introducing non-linearity influence of features on the predicted variable. These machine learning approaches allow for more precise sgRNA design by effectively measuring the intricate factors and their interplay that affect CRISPR-Cas9 performance. [2-7]

Despite these insights, predicting sgRNA efficiency remains a complex challenge. The limitation of conventional features is a major challenge in sgRNA efficiency prediction. Most current models only differ in modelling techniques and have not much difference in features. Regression or classification models triumph in their transparency but often lack predictability [3]. Convolutional neural network models are widely used to predict sgRNA efficiency on nucleotide sequences by processing them as an image after one-hot encoding. Despite their accuracy, it also makes it challenging to understand how specific features contribute to sgRNA efficiency [6, 7]. Current models often have poor cross-species generalizability due to the distinct nature of each species. Models tailored for a specific species are getting more common to avoid this issue [6].

This research project aims to develop a novel feature matrix and leverage different machine learning models with XAI approach to predict sgRNA efficiency. By incorporating both conventional features and quantum chemical properties, the new feature matrix seeks to provide a deeper understanding of the determinants of sgRNA efficiency. A fusion framework is employed to combine the advantages of different machine learning models. Iterative random forest (iRF) model is built to account for the explainability of the model, which provides clear insights into the individual contributions of the features to the cutting efficiency of sgRNA. The iRF model also excels at selecting important features across iterations. Based on the top features selected by iRF, a refined feature set is created to train a feedforward neural network model which can capture more complex relations of features and sgRNA efficiency. This hybrid approach would enhance predictive accuracy and reveal the underlying biological mechanisms that affect sgRNA activity, and eventually contribute to more effective and reliable design of sgRNA in CRISPR-Cas9 system. The model trained on a E. coli sgRNA library is also validated on a H. sapiens dataset to examine the model predictive ability across different species. [6, 8, 9]

Section 2 introduces the methodology used in this study, including the data source, feature extraction process, and modelling techniques, including integrating quantum chemical properties into the feature matrix and details of both iterative random forest and feedforward neural network model. Section 3 presents a detailed analysis of the results, emphasizing model performance metrics and interpretable feature importance. The discussion also addresses the limitations of the study and suggests potential avenues for future improvements.

# 2 Research Methodology

## 2.1 Modelling Methods

### 2.1.1 Overview


The data collection starts with acquiring publicly available sgRNA library for E. coli and H. sapiens which includes sgRNA sequences and associated cutting efficiency scores. Feature extraction is performed by calculating conventional features and quantum chemical properties to create the novel feature matrix. The iterative random forest (iRF) model is trained on the feature matrix to identify and rank the most critical features influencing sgRNA efficiency. The feature set is refined by retaining top features based on the iRF feature importance. A feedforward neural network model is subsequently trained on the refined dataset to capture complex, non-linear relationships within the data. SHAP value is also applied to interpret feature contribution to the neural network model. Finally, model performance is validated on E. coli data to examine cross-species generalizability by applying both models to H. sapiens data.

### 2.1.2 Iterative Random Forest

Iterative random forest (iRF) [6, 21, 22] is a machine learning technique that builds upon the traditional random forest algorithm. It is designed to enhance the interpretability and robustness of feature selection, particularly in high-dimensional datasets where many features might be irrelevant or redundant. In a standard random forest model, the importance of features is determined based on how much they reduce the impurity when splitting nodes in decision trees. However, this process is done only once, and the feature importances are calculated based on that single model. Compared with traditional Random Forest models, iRF involves multiple iterations of the random forest process. After each iteration, feature importances are updated, and the dataset is modified by re-weighting or pruning less important features. This iterative approach helps refine and stabilize the set of important features. As a result, iRF is particularly useful in feature selection. iRF can refine the set of important features by repeatedly fitting random forests and recalculating feature importances. This helps identify a more stable set of features that are truly important for prediction.

In this analysis, a fraction of 20% of the dataset is randomly sampled in each iteration to ensure variability and robustness in the model. A 20% sampling can also save memory especially when handling a large dataset. The sampled data is split into 80% training data and 20% testing data. The random forest model is trained using the Ranger package in R with hyperparameters tuned through a grid search. The number of trees is set to 400 and iterated 10 times to balance the model complexity and simplicity, and capture the relationships between features and sgRNA efficiency without overfitting. A 5-fold cross-validation process is also used to prevent overfitting and provide a more reliable estimation of parameters. The iRF model can also generate a list of top features which can be utilized to create dataset for more sophisticated modelling with selected features.

### 2.1.3 Feedforward Neural Network

A feedforward neural network model [20] is a fundamental type of neural network where information flows in one direction through layers of neurons. It is the simplest form of neural network architecture and is commonly used for classification and regression. It has an input layer that receives the input data. The model also has one or multiple hidden layers where the input features are transformed through weighted connections and an activation function. At each node in the hidden layers, the weighted sum of inputs is calculated, and the activation function like ReLU is applied to introduce non-linearity, allowing the model to capture more complex patterns. The output layer, also the final layer, produces the final output of the model. In a feedforward neural network model, information flows in one direction from the input layer, through the hidden layers, and finally to the output layer. Models are trained using backpropagation and can adjust the weights of connections between neurons to minimize the error between the predicted output and the actual target. Loss function is used to quantify the difference between the predicted output and the true output. The goal of the training is to minimize this loss, so the model fits well on the training data.

The top 100 features are already identified from the iRF models. And a new feature matrix is selected based on the top 100 features to continue the training of the neural network model. The selected features are normalized, ensuring all features contribute equally to the model in the training process. 80% of the data are randomly selected as the training data. The rest of the 20% are used as the testing data. The neural network model is defined using the keras package in R with three hidden layers. The first hidden layer has 64 neurons and is followed by batch normalization and 50% random dropout of neurons. Batch normalization normalizes the inputs to a layer and helps stabilize the learning process. Dropout is a regularization technique that prevents overfitting by randomly dropping out neurons during training. The second hidden layer has 32 neurons and is also followed by batch normalization and 50% random dropout of neurons. The third hidden layer has 16 neurons followed by batch normalization and 20% random dropout of neurons. ReLU activation is used in each layer to accommodate the dataset. L2 regularization is also applied in each layer to reduce overfitting by penalizing large weights. Mean Square Error (MSE) is used as the loss function because it is a regression problem. The learning rate is set as 0.001 with Adam optimizer. A 5-fold cross validation is also used to examine the overfitting of the final model. The final model is trained on the 80% training data and validated on 20% testing data. The epoch is initially set as 100 alone with an early stopping to stop training early if the validation loss does not improve for 10 consecutive epochs.

In compliance with XAI approach, the SHAP value is introduced to rank the feature importance of the neural network model. The SHAP value explains the marginal contributes of each feature to move from the average prediction to the prediction of a specific instance. It helps provide insights into how each feature contributes to the model prediction.

## 2.2 Datasets

### 2.2.1 E. coli dataset

A publicly available dataset including the Cas9-based genome-wide sgRNA libraries of E. coli targeting promoters, ribosome binding sites (RBS), and gene coding regions for CRISPRi are utilized in the analysis. A total of 56437 sgRNAs designed for E. coli are included. The cutting efficiency scores of sgRNAs are determined by taking the logarithm (logx) of the selected read count to the control read count. [10]

### 2.2.2 H. sapiens dataset

A publicly accessible sgRNA library designed for H. sapiens was utilized. This dataset contains 1278 unique sgRNAs. The cutting efficiency of sgRNAs in was quantified by taking the binary logarithm (log2) of the knockout efficacy to the control condition. The H. sapiens dataset will be used to validate whether the model trained using E. coli data can explain across different species. [11]

## 2.3 Feature Matrix

### 2.3.1 Quantum Chemical Properties

Quantum chemical properties [6] provide a deeper understanding of the molecular interactions that affect sgRNA efficiency, allowing machine-learning models to predict sgRNA performance with greater accuracy by incorporating an expanded feature matrix. The analysis included several quantum chemical properties of base pairs, monomers, dimers, trimers, and tetramers. A sliding-window approach was used to determine the position of each base pairs or k-mers on the 20-bp sgRNA sequence to capture the position related information of the quantum chemical tensors.

**The HOMO–LUMO gap** is a quantum chemical property that describes the energy difference between the highest energy of molecular orbital that can donate electrons (HOMO) and the lowest energy state that can accept electrons (LUMO). This gap is a critical factor in determining a molecule's chemical reactivity, stability, and interactions with other molecules. A large HOMO-LUMO gap means that there is a significant energy barrier to receiving or removing electrons, which makes the molecules less able to combine and form an activated complex. A smaller HOMO-LUMO gap means the complex molecules formed is more stable. HOMO-LUMO gap can provide insights into sgRNA interactions with the DNA target and the stability of sgRNA-DNA complex. By integrating the HOMO-LUMO gap and other quantum chemical properties into machine-learning models, researchers can achieve more accurate predictions of sgRNA efficiency, accounting for the subtle molecular interactions that influence CRISPR-Cas9's performance. [12, 13]

**Hydrogen bonding** energy refers to the energy involved in the formation or breaking of hydrogen bonds. It is a weak interaction between a hydrogen atom attached to a more electronegative atom like oxygen or nitrogen. Hydrogen bond is the link between the complementary pair of bases. It plays a crucial role in the stability of the structures of DNA-RNA complex. Higher hydrogen bonding energy means stronger hydrogen bonds between the sgRNA and its target DNA. By including hydrogen-bonding energy as one of the features, models can make better predictions with the stability of hydrogen bonds between paired bases. [14-16]

**Stacking interaction** is a non-covalent interaction between aromatic rings, such as those found in nucleotide bases in DNA and RNA. These interactions are dominated by a type of van der Waals force. They help stabilize the binding of the sgRNA to its target DNA sequence. The strength and nature of these interactions can affect the overall stability of the sgRNA-DNA complex, influencing how effectively the Cas9 enzyme can recognize and cleave the target site. [17, 18]

### 2.3.2 Conventional Features

In this study, several conventional features are also included in the feature matrix. GC content [3] is defined as the percentage of GC in a sgRNA sequence. Temperature of melting (Tm) [6] is the temperature at which a double-stranded DNA or RNA molecule breaks into two single strands. GC content and Tm can be

calculated using Biopython in python. Single strand of RNA often folds into a secondary structure by pairing bases on the same RNA strand. And this property can be represented by the minimum free energy (MFE) of the RNA molecule. MFE can be calculated with ViennaRNA in python [19]. Position independent nucleotide composition features [4] are also critical in conventional prediction of sgRNA efficiency. The counts of monomers and dimers in each RNA sequence are included in this study. The position dependent features [4, 11, 20], which is the k-mer information in a specific position on a 20-bp sgRNA sequence is calculated with one-hot encoding. Position dependent monomers and dimers are included for further analysis. To get the one-hot encoding of position dependent monomers, each nucleotide in the sgRNA sequences is processed respectively by identifying the exact nucleotide in the designated position and record it in an empty dictionary of all possible monomers in a specific position. The one-hot encoding of dimers is calculated with a similar approach that run through all 19 positions of dimers on a 20-bp sgRNA sequence.

## 3 Results and Discussion

### 3.1 Results Summary

| Model | $R^2$ | RMSE | Dataset | Number of Features |
|---|---|---|---|---|
| Benchmark | 0.249 | 9.113 | Guo et al. 2018 | 6232 |
| iRF (Full Dataset) | 0.217 | 2.566 | Guo et al. 2018 | 721 |
| iRF (Conventional Dataset) | 0.205 | 2.588 | Guo et al. 2018 | 205 |
| Neural Network | 0.290 | 2.456 | Doench et al. 2014 | 100 |

**Table 1:** Results Summary

### 3.2 Discussion

#### 3.2.1 iRF model

A total of 56437 sgRNAs designed for E. coli are used to train the iRF models. The novel feature matrix with quantum chemical features has 721 predictive variables. And the conventional feature matrix without quantum chemical features has 405 predictive variables. An 80% and 20% split is applied to create the training data and testing data.

The model trained on feature matrix that included quantum chemical properties achieved a root mean squared error (RMSE) of 2.566, indicating that the average deviation of the predicted sgRNA efficiencies from the actual values is about 2.566 units. This relatively low RMSE suggests that the model is highly accurate in its predictions, minimizing errors in predicting sgRNA efficiency. The adjusted $R^2$ of 0.217 indicates that the model explains approximately 21.7% of the variance in sgRNA efficiency, accounting for the number of features used in the model. The adjusted $R^2$ of 0.217 shows quite ideal results compared with similar models.

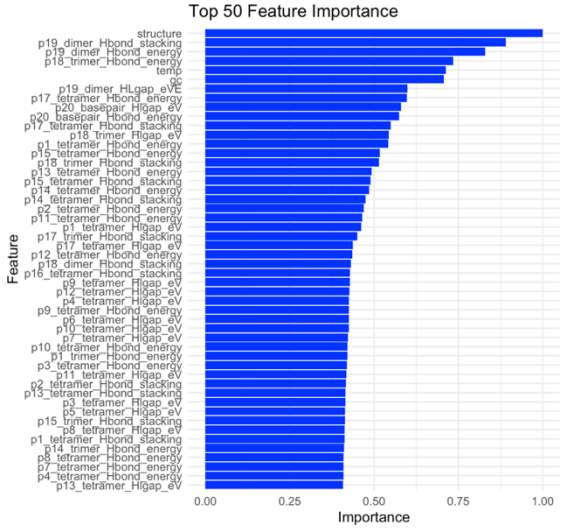


**Figure 1:** Top 50 Important Features of Quantum Chemical Model

Feature importance (Figure 1) analysis further highlights the strength of quantum chemical features. Different quantum chemical features, such as hydrogen bonding energy, stacking interactions, of different k-mers and positions on the genome rank on the top of the list. Conventional features like minimum free energy, temperature of melting and GC content also score high on the list which means they can provide unique and important insights into sgRNA efficiency that quantum chemical properties cannot replace.

The model trained on the conventional dataset, on the other hand, has a slightly higher RMSE of 2.588 compared to the quantum chemical model RMSE of 2.566. This indicates that the conventional model is slightly less accurate, with predictions deviating more from the actual sgRNA efficiencies. The slight drop in RMSE of quantum chemical model highlights its superior accuracy in prediction. The adjusted $R^2$ for the conventional model is 0.205, which is lower than the 0.217 achieved by the quantum chemical model. The adjusted $R^2$ also excluded the effect of the increase of predictive variables on the model fitting. This suggests that the quantum chemical model has a better overall model fitting than conventional model. The integration of both traditional and quantum features allows the model to provide a comprehensive prediction framework that considers both broad and specific factors influencing sgRNA efficiency.

The conventional model relies primarily on sequence-based position dependent and independent features without incorporating the more nuanced quantum chemical properties (Figure 2). Minimum free energy, temperature of melting and GC content rank top among the top features in conventional model. Position dependent and independent features dominate the rest of the list, which would be replaced when quantum chemical features are incorporated.

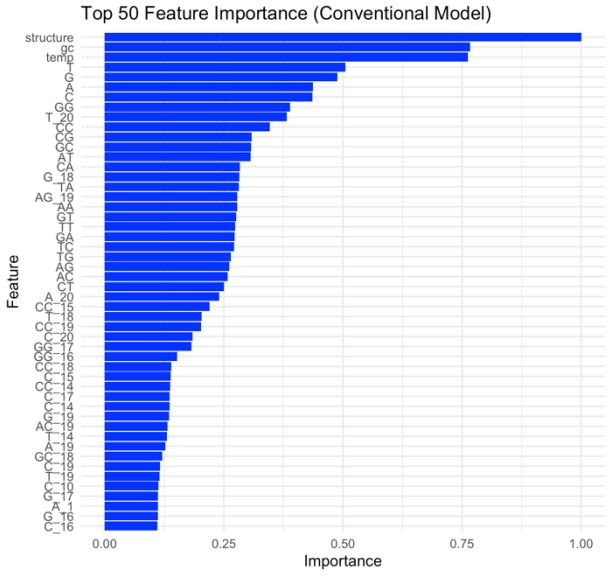


**Figure 2:** Top 50 Important Features of Conventional Model

Overall, the results demonstrate that integrating quantum chemical properties into the machine learning feature matrix not only improves the accuracy of sgRNA efficiency prediction but also provides interpretable insights into the biological mechanism of CRISPR-Cas9 system.

### 3.2.2 Feedforward Neural Network model

A similar dataset of 56437 sgRNAs designed for E. coli and 100 selected features are used to train the neural network models. An 80% and 20% split is applied to create the training data and testing data. The neural network model has been evaluated using both 5-fold cross-validation and final evaluation metrics. These metrics provide a detailed understanding of the model predictive performance, both during the cross-validation process and after training on the entire dataset.

Evaluation metrics of the cross-validation process are calculated by taking the average values of the models trained on all 5 folds. The average mean square error (MSE) of 5-fold cross-validation indicates the average squared difference between the predicted and actual values of the 5-fold cross-validation. The model has an average MSE of 6.488, suggesting that there is some deviation between predictions and actual outcomes. The average mean absolute error (MAE) of 5-fold cross-validation measures the average errors in the predictions, without considering their direction. An MAE of 2.043 indicates that, on average, the model's

predictions are about 2.043 units off from the actual values. The average adjusted $R^2$ of 0.232 during cross-validation shows that about an average of 23.2% of the variance in the data is explained by the model. This suggests that the neural network model captures more variability and has better general model fitting than the iRF model.

In the final neural network model trained on the entire dataset, the training automatically stopped at the 92nd epoch. The model MSE is 6.031, which is slightly lower than the cross-validated MSE. This suggests that the model performs slightly better when evaluated on the full training data. The 1.972 MAE of the final model further confirms the model accuracy with an average error of 1.972 units which is slightly lower than the cross-validated MAE.

In the final neural network model, an RMSE of 2.456 indicates that the final neural network model has predictions deviate from the actual values by approximately 2.456 units, which performs better than the iRF model trained on both novel feature matrix with quantum chemical features included, and the conventional feature matrix.

The final neural network model has an adjusted $R^2$ of 0.290, suggesting that the model explains around 29% of the variance in the sgRNA efficiency data. This is a significant improvement achieved on a highly trimmed feature matrix compared with iRF models, indicating that the neural network model has much better model fitting and ability to predict sgRNA efficiency.

According to the SHAP value (Figure 3) which measures the individual contribution of each feature, most of the top features are related to quantum chemical properties such as hydrogen bonding energy and stacking interactions across various position on the sgRNA sequence.

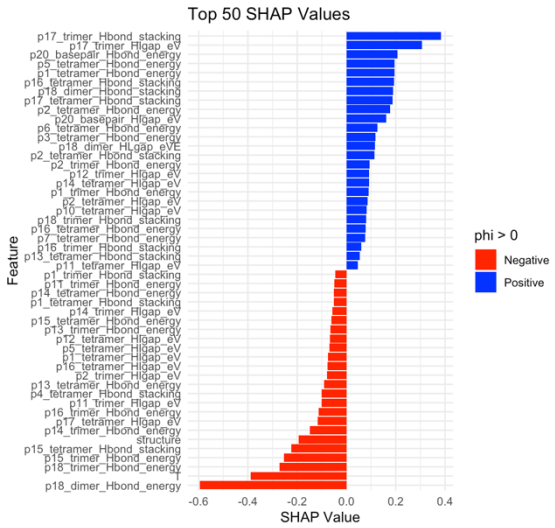


**Figure 3:** Top 50 SHAP Values of Neural Network Model

As in the iRF models, these properties are also crucial for understanding the stability and interaction of the sgRNA with its target in the neural network models. However, the ranking of the important features are quite different from the iRF one. It indicates the non-linear relation of the quantum chemical properties and sgRNA efficiency.

### 3.2.3 Validation on H. sapiens

The H. sapiens sgRNA library with 1278 sgRNA are utilised. The iRF model trained with the extended dataset and the final neural network model are saved for validation on the H. sapiens dataset of the cross-species explainability of the model trained on E. coli data. Both models have reported terrible performance on H. sapiens data with negative R-squared. Negative R-squared usually indicates model overfitting on the training data. In this context, it means both models trained capture too many unique characteristics of E. coli that they lose predictive power on the efficiency of sgRNAs designed for human genome. Even though the two models have good performance in E. coli scenarios, it should be aware that the models might have problems predicting the efficiency of sgRNA designed for other species.

### 3.2.4 Limitations

The model has poor cross species generalisation. The iRF model and neural network models have very different emphasis on the features but this study can't reveal the reason for that.

### 3.2.5 Future Work

In conclusion, this study highlights the significance of incorporating quantum chemical properties into predictive models for sgRNA efficiency. The integration of these advanced features into the iRF model not only improves the accuracy of predictions, but also provides interpretable insights into the mechanisms of CRISPR-Cas9 system. The neural network model also underscored the value of quantum chemical properties, achieving better predictive performance than iRF model. The SHAP value analysis further revealed that the quantum chemical features are highly important to the model prediction, although their importance differs in a non-linear manner compared to the iRF model. Despite these success, the models demonstrated poor performance when applied to H. sapiens data, indicating a limitation in the cross-species applicability of the models. It suggests that while the models are highly effective within the specific context of E. coli, they may overfit to the unique characteristics of this species, limiting their generalizability to other organisms.

## 4 References


[1] Doudna, Jennifer A., and Emmanuelle Charpentier. "The New Frontier of Genome Engineering with CRISPR-Cas9." Science 346, no. 6213 (2014): 1077–1077. http://www.jstor.org/stable/24745404.

[2] Chuai, G.-H., Wang, Q.-L., & Liu, Q. (2017). In Silico Meets In Vivo: Towards Computational CRISPR-Based sgRNA Design. Trends in Biotechnology, 35(1), 12-21. https://doi.org/10.1016/j.tibtech.2016.06.008.

[3] Konstantakos, V., Nentidis, A., Krithara, A., & Paliouras, G. (2022). CRISPR–Cas9 gRNA efficiency prediction: An overview of predictive tools and the role of deep learning. Nucleic Acids Research, 50(7), 3616-3637. https://doi.org/10.1093/nar/gkac192.

[4] Peng, H., Zheng, Y., Blumenstein, M., Tao, D., & Li, J. (2018). CRISPR/Cas9 cleavage efficiency regression through boosting algorithms and Markov sequence profiling. Bioinformatics, 34(18), 3069-3077. https://doi.org/10.1093/bioinformatics/bty298.

[5] Jifang Yan, Guohui Chuai, Chi Zhou, Chenyu Zhu, Jing Yang, Chao Zhang, Feng Gu, Han Xu, Jia Wei, Qi Liu, Benchmarking CRISPR on-target sgRNA design, Briefings in Bioinformatics, Volume 19, Issue 4, July 2018, Pages 721–724, https://doi-org.libproxy1.nus.edu.sg/10.1093/bib/bbx001.

[6] Noshay, J. M., Walker, T., Alexander, W. G., Klingeman, D. M., Romero, J., Walker, A. M., Prates, E., Eckert, C., Irle, S., Kainer, D., & Jacobson, D. A. (2023). Quantum biological insights into CRISPR-Cas9 sgRNA efficiency from explainable-AI driven feature engineering. Nucleic Acids Research, 51(19), 10147–10161. https://doi.org/10.1093/nar/gkad736.

[7] Dimauro, G., Colagrande, P., Carlucci, R., Ventura, M., Bevilacqua, V., & Caivano, D. (2019). CRISPRLearner: A deep learning-based system to predict CRISPR/Cas9 sgRNA on-target cleavage efficiency. Electronics, 8(12), 1478. https://doi.org/10.3390/electronics8121478.

[8] Zhu, W., Xie, H., Chen, Y., & Zhang, G. (2024). CrnnCrispr: An interpretable deep learning method for CRISPR/Cas9 sgRNA on-target activity prediction. International Journal of Molecular Sciences, 25(8), 4429. https://doi.org/10.3390/ijms25084429.

[9] Liu, Y., Fan, R., Yi, J., Cui, Q., & Cui, C. (2023). A fusion framework of deep learning and machine learning for predicting sgRNA cleavage efficiency. Computers in Biology and Medicine, 165, 107476. https://doi.org/10.1016/j.compbiomed.2023.107476.

[10] Guo, J., Wang, T., Guan, C., Liu, B., Luo, C., Xie, Z., Zhang, C., & Xing, X.-H. (2018). Improved sgRNA design in bacteria via genome-wide activity profiling. Nucleic Acids Research, 46(14), 7052–7069. https://doi.org/10.1093/nar/gky572.

[11] Doench, J. G., Hartenian, E., Graham, D. B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B. L., Xavier, R. J., & Root, D. E. (2014). Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. Nature Biotechnology, 32(12), 1262-1267. https://doi.org/10.1038/nbt.3026.

[12] Aihara, J. (1999). Reduced HOMO-LUMO Gap as an Index of Kinetic Stability for Polycyclic Aromatic Hydrocarbons. The Journal of Physical Chemistry A, 103(37), 7487-7495. https://doi.org/10.1021/jp990092i.

[13] Huang, Y., Rong, C., Zhang, R., & Liu, S. (2017). Evaluating frontier orbital energy and HOMO/LUMO gap with descriptors from density functional reactivity theory. Journal of Molecular Modeling, 23(3). https://doi.org/10.1007/s00894-016-3175-x.

[14] Ghiandoni, G.M., Caldeweyher, E. Fast calculation of hydrogen-bond strengths and free energy of hydration of small molecules. Sci Rep 13, 4143 (2023). https://doi.org/10.1038/s41598-023-30089-x.

[15] Halder A, Data D, Seelam PP, Bhattacharyya D, Mitra A. Estimating Strengths of Individual Hydrogen Bonds in RNA Base Pairs: Toward a Consensus between Different Computational Approaches. ACS Omega. 2019 Apr 23;4(4):7354-7368. doi: 10.1021/acsomega.8b03689.

[16] Santos, C. S. dos, Drigo Filho, E., & Ricotta, R. M. (2015). Quantum confinement in hydrogen bond of DNA and RNA. Journal of Physics: Conference Series, 597(1), 012033. https://doi.org/10.1088/1742-6596/597/1/012033.

[17] Sponer, J., Sponer, J. E., Mládek, A., Jurečka, P., Banáš, P., & Otyepka, M. (2013). Nature and Magnitude of Aromatic Base Stacking in DNA and RNA: Quantum Chemistry, Molecular Mechanics, and Experiment. Biopolymers, 99(12), 978-988. https://doi.org/10.1002/bip.22322.

[18] Sponer, J., Riley, K. E., & Hobza, P. (2008). Nature and magnitude of aromatic stacking of nucleic acid bases. Physical Chemistry Chemical Physics, 10(19), 2595-2610. https://doi.org/10.1039/b719370j.

[19] Lorenz,R., Bernhart,S.H., Honer Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA package 2.0. AlgorithmsMol.Biol., 6, 26.

[20] Jeremy Charlier, Robert Nadon, Vladimir Makarenkov, Accurate deep learning off-target prediction with novel sgRNA-DNA sequence encoding in CRISPR-Cas9 gene editing, Bioinformatics, Volume 37, Issue 16, August 2021, Pages 2299–2307, https://doi.org/10.1093/bioinformatics/btab112

[21] Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software, 77(1). https://doi.org/10.18637/jss.v077.i01.

[22] Walker, A. M., Cliff, A., Romero, J., Shah, M. B., Jones, P., Gazolla, J. G. F. M., Jacobson, D. A., & Kainer, D. (2022). Evaluating the performance of random forest and iterative random forest based methods when applied to gene expression data. Computational and Structural Biotechnology Journal, 20, 3372–3386. https://doi.org/10.1016/j.csbj.2022.06.037.

[23] Heler, R., Marraffini, L.A. and Bikard, D. (2014), Acquisition of new spacers by CRISPR-Cas immune systems. Molecular Microbiology, 93: 1-9. https://doi-org.libproxy1.nus.edu.sg/10.1111/mmi.12640

[24] Jiang, Fuguo, Doudna, Jennifer. A. (2017), CRISPR–Cas9 Structures and Mechanisms. Annual Review of Biology, 46: 505-529. https://doi.org/10.1146/annurev-biophys-062215-010822

[25] Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ, Wu X, Shalem O, Cradick TJ, Marraffini LA, Bao G, Zhang F. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol. 2013 Sep;31(9):827-32. doi: 10.1038/nbt.2647. Epub 2013 Jul 21. PMID: 23873081; PMCID: PMC3969858.

[26] Doench, J G, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, Tothova Z, Wilen C, Orchard R, Virgin HW, Listgarten J, Root DE. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol. 2016 Feb;34(2):184-191. doi: 10.1038/nbt.3437. Epub 2016 Jan 18. PMID: 26780180; PMCID: PMC4744125.

[27] McFadden, J. and Al-Khalili, J. (2018) The origins of quantum biology. Proc. Math. Phys. Eng. Sci., 474, 20180674. http://0-doi-org.pugwash.lib.warwick.ac.uk/10.1098/rspa.2018.0674.