Predicting plant Rubisco kinetics from RbcL sequence data using machine learning

[1]Wasim A Iqbal, [2]Alexei Lisitsa and *[1]Maxim V Kapralov

[1] School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne, NE1 7RU, United Kingdom

[2] Department of Computer Science, University of Liverpool, Liverpool, L69 3BX, United Kingdom

* Maxim.Kapralov@ncl.ac.uk

Date of revised submission: 29/06/2022

Number of figures in manuscript: 5

Number of figures in supplementary: 12

Number of tables in supplementary: 4

Word count (excluding methods): 3433

Short title: Predicting Rubisco kinetics from RbcL sequence data

Highlight

This paper is the first to demonstrate machine learning approaches as a tool for predicting Rubisco kinetics from RbcL sequences.

Abstract

Ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco) is responsible for the conversion of atmospheric CO_2 to organic carbon during photosynthesis and often acts as a rate limiting step in the later process. Screening the natural diversity of Rubisco kinetics is the main strategy used to find better Rubiscos for crop engineering efforts. Here, we demonstrate the use of Gaussian processes (GPs), a family of Bayesian models, coupled with protein encoding schemes for predicting Rubisco kinetics from Rubisco large subunit (RbcL) sequence data. GPs trained on published experimentally obtained Rubisco kinetic datasets were applied to over 9,000 sequences encoding RbcL to predict Rubisco kinetic parameters. Notably, our predicted kinetic values were in agreement with known trends, e.g. higher carboxylase turnover rates (Kcat) for Rubiscos from C_4 or Crassulacean acid metabolism (CAM) species compared to ones found in C_3 species. This is the first study demonstrating machine learning approaches as a tool for screening and predicting Rubisco kinetics, and our approach could be applied to other enzymes.

Key words

Rubisco, Machine learning, Gaussian process, Photosynthesis, Enzyme, Kinetics

Abbreviations

Ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco)

Rubisco large subunit (RbcL)

Rubisco small subunit (RbcS)

Carboxylation turnover rate (Kcat)

Specificity for CO_2 over O_2 (Sc/o)

Michaelis-Menten constant for CO_2 (Kc)

Michaelis-Menten constant for CO_2 at ambient O_2 ($Kc^{21\%O2}$)

Carbon concentrating mechanism (CCM)

Machine learning (ML)

Gaussian process (GP)

Standard error (SE)

Standard deviation (SD)

Length scale (l)

Variance (σ^2)

Coefficient of determination (R^2)

Mean absolute error (MAE)

t-distributed stochastic nearest neighbour (t-SNE)

Crassulacean acid metabolism (CAM)

Introduction

Ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco) is claimed to be the most abundant enzyme on Earth (Bar-On and Milo, 2019). The global conversion of inorganic CO_2 to organic forms is mostly driven by Rubisco making it a gatekeeper of carbon for nearly all life on the planet (Raven, 2013). Form IB Rubiscos found in plants and green algae consists of both large subunits and small subunits, and the large subunits contain the Rubisco active site. Thus, it has long been assumed that the large subunit sequence variation contributes to the diversity of Rubisco kinetics (Kellogg and Juliano, 1997, Camel and Zolla, 2021). Rubisco is often characterised as having a slow turnover rate (Kcat) for CO_2 and poor specificity for CO_2 compared to O_2 (Sc/o) (but see Tcherkez et al. (2006)). Rubisco catalytic inefficiencies might limit plant photosynthetic performance in certain environmental conditions such as saturating irradiance and limiting CO_2 concentrations. Improving Rubisco kinetic traits is therefore a target for improving plant carbon uptake and crop yield. One strategy of doing this is screening the natural diversity of Rubisco kinetics and replacing of a plant's native Rubisco with a better enzyme (Ort et al., 2015, Hermida-Carrera et al., 2016, Orr et al., 2016, Sharwood et al., 2016, Galmés et al., 2019, Orr and Parry, 2020, Von Caemmerer, 2020, Iqbal et al., 2021, Lin et al., 2022). Although there has been some progress with this strategy, direct replacement of Rubiscos in crops is currently challenging due to both limited capacity to mass-screen Rubisco kinetics, and Rubisco chaperone incompatibilities between distant species (Kanevski et al., 1999, Whitney et al., 2011, Whitney et al., 2015, Wilson et al., 2016, Sharwood, 2017, Wilson et al., 2018, Zhou and Whitney, 2019, Gunn et al., 2020, Martin-Avila et al., 2020).

Given the resource-intensive nature of screening enzyme kinetics in the laboratory, modelling or *in silico* approaches, such as machine learning (ML), are being increasingly adopted to aid bioengineering efforts (Bedbrook et al., 2017, Yang et al., 2018, Li et al., 2019, Yang et al., 2019, Benes et al., 2020, Bonetta and Valentino, 2020, Zhu et al., 2020, Biswas et al., 2021, Wittmann et al., 2021, Brandes et al., 2022, Hsu et al., 2022). ML largely consists of 'supervised' tasks that involve training ML algorithms on previously seen protein sequences (e.g. enzyme sequence) with associated labels (e.g. catalytic activity). The trained model can then be used to predict labels of previously unseen but similar data inputs (Yang et al., 2019, Mazurenko et al., 2020, Newman and Furbank, 2021, Wittmann et al., 2021). Several examples exist of ML applications being used to screen enzyme properties, however no model exists which has predicted Rubisco kinetics from sequence variation (Romero et al., 2013, Yang et al., 2018, Greenhalgh et al., 2021, Hsu et al., 2022). The reasons for this may be that we do not know exactly which properties of the Rubisco protein determine Rubisco kinetics. Additionally, state-of-the-art ML algorithms such as neural networks usually require hundreds or thousands of labelled data to perform well that is not possible with the current size of Rubisco datasets.

Gaussian processes (GPs), a family of non-parametric, non-linear Bayesian models have shown to predict enzyme properties such as thermostability and activity given a limited amount of experimental data (Rasmussen and Williams, 2006, Yang et al., 2018, Yang et al., 2019, Deringer et al., 2021, Dutordoir et al., 2021). A GP finds non-linear functions $f(x1), f(x2)$ that map the relationship of similar labels (e.g. catalytic activity) with similar inputs $x1, x2$ (e.g. enzyme sequences), as encoded by a kernel function (Jokinen et al., 2018, Greenhalgh et al., 2021). The kernel function measures the similarity of the input data in the form of a covariance matrix. A key feature of a GP is that it can characterise the model uncertainty due to lack of similar data, which can be used to determine the quality of predictions.

With all ML techniques, protein sequences must be transformed into numerical representations and performance can suffer if the protein sequences are not encoded correctly. It is difficult to suggest *a priori* the best way to numerically represent protein sequences, as there are a variety of levels protein sequences can be represented, such as physiochemical properties of amino acids or the three-dimensional structure. Over the past decade, two classes of encoding schemes have been tested for mapping protein sequence-function relationships. A classical encoding scheme (or ‘one-hot encoding’) directly represents a protein sequence amino acids in binary notation and a ‘learned encoding’ scheme ,which involves training an unsupervised ML method on millions of unlabelled protein sequences (Yang et al., 2018, Alquraishi, 2021, Elnaggar et al., 2021, Rives et al., 2021, Wittmann et al., 2021). After the learned encoding scheme has been trained it can be reused to produce numerical vector representations of protein sequences (Elabd et al., 2020, Faulon and Faure, 2021, Wittmann et al., 2021). The learned encoding scheme assumes that all protein sequences follow a set of evolutionary rules or biophysical traits that govern the relationships between protein sequences that allow them to carry out a biological function (Elabd et al., 2020, Faulon and Faure, 2021, Wittmann et al., 2021). The vector representations from the learned encoding scheme capture the relationships between proteins from the learned sequence-space. As result, similar sequences will have similar vector representations and so can be assumed to have similar biological function by a downstream-supervised ML model such as a GP (Elabd et al., 2020, Faulon and Faure, 2021, Wittmann et al., 2021).

We think that the above ML processes could map the Rubisco sequence-function landscape for predicting unmeasured Rubisco kinetics. Previously, it was shown that Rubisco kinetic trade-offs exist between the Sc/o, Kcat and Michaelis-Menten constant for CO_2 (Kc), leading to the belief that Rubisco kinetics are heavily constrained within a low-dimensional landscape (Tcherkez et al., 2006, Savir et al., 2010). However, recent work highlighted the importance of phylogenetic constraints for Rubisco kinetics suggesting that closely related species are more likely to have similar kinetics (Flamholz et al., 2019, Bouvier et al., 2021); but see exceptions driven by a rapid evolution within recent adaptive radiations (Kapralov and Filatov, 2006, Kubien et al., 2008, Kapralov et al., 2011, Galmés et al., 2014a) Thus, similarity of Rubisco sequences might be among the many features that GPs with protein encoding schemes may use for interpolating uncharacterised Rubisco kinetics.

Here, we trained GPs with either a learned encoding scheme or classical encoding scheme on form IB Rubisco sequence and kinetic data from C_3 and C_4 plant species. We evaluated the performance of the ML frameworks using leave-one-out cross validation and found that the GPs with the learned encoding scheme outperformed the classical encoding scheme. Next, we subjected the GPs with the learned encoding scheme to another validation framework to detect overfitting. This involved removing species sharing the same genus during model training and using the unseen genus group to assess model performance; from here on referred to as ‘leave-genus-out’ cross validation. We found that the GPs with a learned encoding scheme generalised across plant genera well. Finally, we wanted to validate hundreds of predictions without experimental data. One strategy of doing this was grouping predictions by photosynthesis metabolism type and taxonomical group for which mechanisms have been hypothesised to constrain Rubisco kinetics.

Methods

Rubisco kinetics and sequence data

Rubisco large subunit harbouring the catalytic site is encoded by the RbcL gene ,which therefore has a major influence on Rubisco kinetic properties (Kellogg and Juliano, 1997, Camel and Zolla, 2021). 165 C_3 and C_4 plant Rubisco *in vitro* Kcat values (25^{O}C pH near 8), 170 *in vitro* Sc/o values and 170 *in vitro* Kc values as well as corresponding RbcL sequences were obtained from literature (Jordan and Ogren, 1983, Lehnherr et al., 1985, Uemura et al., 1997, Kubien et al., 2008, Savir et al., 2010, Viil et al., 2012, Galmés et al., 2014a, Galmes et al., 2014, Hermida-Carrera et al., 2016, Prins et al., 2016, Sharwood et al., 2016, Long et al., 2018, Flamholz et al., 2019). If studies reported overlapping *in vitro* kinetic data, the duplicate from the most recent study was kept and the other duplicate(s) discarded. Additional corrections were made to the data as follows: Standard errors (SE) with reported kinetic values such as Kcat, Kc and Sc/o were converted to standard deviations (SD) using the number of species and/or replicates. When the number of replicates and/or species were not reported, the number of measurements were assumed to be from one sample. When the number of replicates and/or species were reported as a range (e.g. n= 6-10) the mean number of samples was taken. Kc measurements under anoxygenic conditions were adjusted to ambient O_2 conditions ($Kc^{21\%O2}$) using the following equation: $\mathrm{Kc}^{21\%\mathrm{O2}} = \mathrm{Kc}^{0\%\mathrm{O2}} \cdot (1 + \frac{\mathrm{O_2}}{\mathrm{K_o}})$ (Von Caemmerer, 2000). Where '$Kc^{0\%O2}$ ' refers to Kc measured under anoxygenic conditions, 'O_2' refers to the ambient O_2 level and 'Ko' refers to the Rubisco Michaelis-Menten constant for O_2 (μM).

Model setup

Figure 1 shows a schematic diagram of the ML procedure. Just like a simple linear model, a GP can be used for regression or classification tasks (Rasmussen and Williams, 2006, Garnett, 2022). Here, since kinetics are continuous variables a GP regression was used. All ML tasks were performed using the python 'GPflow' module (version 2.1) and packaged into user-friendly Google COLAB notebooks (https://github.com/Iqbalwasim01/Mining-Rubisco-kinetics.git) (Matthews et al., 2017).

Protein encoding scheme

Two protein encoding schemes were tested before choosing a final encoding scheme. The classical encoding scheme (or one-hot encoding) expresses each amino acid as a 20 digit vector with the value '1' indicating the identity and position of the current amino acid out of 20 other amino acid types ,which are represented with the value '0' (Yang et al., 2018, Bonetta and Valentino, 2020, Elabd et al., 2020). The one-hot encoding is a relatively sparse and memory inefficient representation of protein sequences. For example, an RbcL with a length of 450 amino acids would result in a 9000 length vector. Further, ‘one-hot encoding’ requires that all RbcL sequences are aligned to the same length and each time a new sequence is added the alignment procedure must be repeated. Here, an alignment procedure was performed using the ‘msa’ R package with the ‘clustal omega’ alignment algorithm (Bodenhofer et al., 2015).

On the other hand, the learned encoding scheme takes inspiration from natural language processing and involves a semi-supervised ML model, learning basic underlying laws or rules of protein sequences that allow proteins to carry out a biological function (Yang et al., 2018, Bonetta and Valentino, 2020, Elabd et al., 2020, Wittmann et al., 2021). The Rives et al. (2021) learned encoding scheme also known as ESM-1b based on a neural network with a transformer architecture was adopted . Previous studies have shown that it predicts residue-residue contacts and secondary structure better than other transformers (Rao et al., 2019, Elnaggar et al., 2021). The learned encoding scheme summarised each RbcL sequence as a vector of length 1280. Once the RbcL sequences have been converted to either the classical or learned encoding, the encodings served as the direct inputs into the GP regression (Figure 1).

GP covariance structure

A GP regression defines a distribution over functions linking data inputs (e.g. RbcL sequence encodings) with labels (e.g. kinetics). The functions are encoded by a kernel function represented as a covariance matrix and mean ,which measure the similarity or nearness of input data (Rasmussen and Williams, 2006, Garnett, 2022). The kernel function makes the basic assumption that data inputs (e.g. RbcL sequences) ,which are closely related are more likely to have similar labels but some additional prior knowledge is required such as whether the functions are linear, smooth or rough. When the underlying nature is unknown a popular choice of kernel is the non-linear ‘Matern 5/2’ kernel ,which was used here (Rasmussen and Williams, 2006). A linear kernel function was also tested to demonstrate the need for the non-linear Matern 5/2 kernel. When data inputs consist of more than one numerical value, the kernel can be applied to each numerical value position allowing the GP regression to learn across multiple input positions known as an ‘additive kernel’ (Duvenaud et al., 2011). For instance, many phenomenon depend on the sum of parts such as the value of a car ,which can be better approximated by the sum of prices of individual car parts. Similarly, the amino acid sites in a protein sequence may convey greater information when protein sequences share a high degree of overall structural similarity. Therefore, this study first applied the kernel function to each learned encoding input position or classical encoding alignment position i.e. $K = k(x_1) + k(x_2)$ ... (Figure 1).

The performance with an additive kernel was then compared to a single kernel where the GP depends on all input positions simultaneously i.e. $K = k(x_1, x_2, \ldots)$. The reason for testing both kernel configurations is that if the encodings consist of many low-order interactions, the additive kernel can exploit this and improve model performance (e.g. see Figure 5 Duvenaud et al. (2011)), if not both the additive and single kernel configurations should give similar performance. Finally, during training the kernel hyperparameters such as the length scale 'l' and/or variance 'σ^2' were tuned to maximise the probability of observing the data points known as the marginal likelihood. Predictions for new data inputs were then obtained from drawing samples from the trained GP.

Leave-one-out cross validation

Performance of the GP regression was assessed using leave-one-out cross validation. Generally, any cross-validation involves splitting a dataset into training and testing datasets. The training dataset with input data (e.g. RbcL sequence encodings) and labels (e.g. kinetics) is used to fit the GP regression model parameters and the testing dataset with input data and labels is used to assess the performance of the trained GP regression to unseen data. Leave-one-out cross validation as the name implies involves holding out one labelled data input out of the training dataset and using the remainder of the dataset for fitting the GP model parameters and predicting the unseen labelled data input that was left out. For example, if a dataset consists of 170 data inputs with labels, the model would be trained on 169 data inputs with labels and the data input and label that was omitted would serve as the testing data set. Leave-one-out cross validation is carried out on each labelled data input, leaving a different labelled data input out of the training dataset each time. The predictions are gathered and performance metrics such as coefficient of determination (R^2) and mean absolute error (MAE) are calculated with the experimental data.

Leave-genus-out cross validation

The leave-one-out cross-validation aims to reduce the chance of model overfitting and provide a depiction of model performance to unseen data. We know patterns or biases can arise from training models on similar datasets that could give a misleading picture of model performance. For instance, it is well known that form IB Rubiscos from the same genus can have similar sequences and kinetic properties (Hermida-Carrera et al., 2016, Orr et al., 2016). This could have led to overoptimistic performance metrics during leave-one-out cross validation because at least one form IB variant from the same genus would have been left in the training dataset during model training. To see if the GP regression generalises across genera, attempts were made to split the data equally while ensuring that a genus group was left out of the training set each time. However, each genus group had unequal species numbers ,which made it difficult to create equally distributed testing/training splits while ensuring non-overlapping genus criteria. Instead, educated splits between the data were made by leaving a genus group out of the training data and then testing of the model on this omitted genus group. While the R^2 metric was used in the leave-one-out cross validation for assessing performance, it is not suitable for assessing all areas of predictive performance because it scales with the size of the dataset (i.e. the more data points there are the less sensitive the R^2 metric is to changes) and assumes values are strictly monotonically associated. Because each genus group contained unequal species numbers, were small and predictions may not be normally distributed or monotonically associated with experimental values, model performance was assessed with the MAE metric as well as direct comparison with the experimental means ± SD.

Benchmarking GP uncertainty estimates

A benefit of a GP is that a $'\sigma^2'$ estimate is provided with each prediction, which allows users to identify predictions with a high chance of being different from the training dataset. In other words, the lower the predicted σ^2 the nearer the prediction is to an example found in the training dataset. However, the GP σ^2 parameter is not explicitly dependent on the labels (i.e. kinetics) and is actually dependent on the data inputs (e.g. see equation 24 Deringer et al. (2021)). During training, the σ^2 parameter is implicitly mapped to the data labels via hyperparameter optimisation. Because the σ^2 parameter is a trainable part of the model, the reliability of the σ^2 estimates must be assessed against test data. Here, the quality of the predicted σ^2 estimates from cross validation was first assessed using the spearman rank correlation with the true errors (i.e. absolute errors between actual mean values and predicted mean values) (Greenman et al., 2022). Secondly, we assessed if the actual mean values fall within the 95% predicted confidence intervals (CIs) $(\pm 2\sigma)$ as demonstrated by Kompa et al. (2021) . This method involves two metrics: ‘coverage’, which is if the actual mean value falls within the predicted 95% CI and ‘width’, which is the full range of the predicted 95% confidence interval (4σ).

t-distributed stochastic neighbour embedding (t-SNE)

In this study protein encoding schemes convert protein sequences from their widely used amino acid format to sequences of numbers ,which cannot be understood using conventional protein sequence analysis methods such as multiple sequence alignments. To investigate how protein encoding schemes portray proteins, which ultimately determine their fate for prediction tasks, a dimensionality reduction method called t-distributed stochastic neighbour embedding (t-SNE) was applied (Maaten and Hinton, 2008). t-SNE projects the protein encodings into two-dimensions ,which allows patterns/clustering arising from the protein encodings to be visualised. t-SNE was performed on the RbcL classical and learned encodings with a perplexity of 20 and default learning rate parameters using the ‘sci-kit learn’ python module (version 1.0.2) (Pedregosa et al., 2011).

Assessing RbcL sequence-space predictions with trait data

Wild-type RbcL sequences from non-redundant protein databases were obtained (n 35,413) from a recent search (Davidi et al., 2020). Unknown species, sequences with lengths >500 or <450 and duplicates entries were omitted leaving 13,124 unique RbcL sequences. 9052 RbcL sequences identified as land plants (Embryophyta) remained. Using the fully trained GPs with the chosen encoding scheme, Rubisco kinetic predictions were obtained for 9052 land plants. Predictions were grouped by plant photosynthetic type (C3, C4 or CAM) and taxonomical group (Angiosperms, Bryophytes, Gymnosperms, and “Ferns”, the latter is a group that included Pteridophyta and Lycopodiophyta). Differences between groups were assessed using one-way ANOVA and Duncan’s post hoc test with the ‘DescTools’ R package (version 0.99.44).

While the sequence criteria of <500 and >450 was used to remove incomplete sequences, some sequences may still have several amino acids missing from the N-terminus and/or C-terminus or ambiguous amino acids ,which could have led to high predicted σ^2. To see if such sequences affected the distribution of predictions, predictions were restricted based on σ^2 estimates selected from cross validation if the σ^2 estimates were well calibrated. Otherwise, the influence of outliers was assessed by removing predictions outside the training dataset ranges. Predictions were grouped by plant photosynthetic type and taxonomical group as described before.

Results

GP performance with a learned encoding scheme compared with a classical encoding scheme

GPs with the learned encoding and classical encoding schemes were trained on form IB RbcL sequence and kinetic data. The performance of the two encoding schemes applied to a single and additive kernel configuration was assessed (Figure S1-S3). The GPs with the learned encodings applied to an additive non-linear Matern 5/2 kernel had the highest predictive ability (Figure 2) (R^2 0.79-0.86) compared with the classical encodings (R^2 0.60-0.74) and other kernel configurations (Figure S1-S3). These results justified the adoption of the learned encodings with the non-linear Matern 5/2 additive kernel for the final models (Figure 2).

GP performance with the learned encoding scheme for numerous plant genera

Form IB Rubisco variants included as part of the training data could have led to overoptimistic performance metrics shown in Figure 2 because at least one form IB Rubisco from the same genus may have been left in the training dataset during model training. Here, the GPs with the learned encoding scheme were assessed using another validation framework. This time form IB Rubiscos sharing the same genus were omitted from the model during training. The remaining data was used to train the model and the omitted genus group was used to assess the model performance.

The GPs with the learned encoding scheme displayed excellent performance. The majority of genus groups had Kcat predictions with a MAE <0.5 s^{-1} (Figure S4), $Kc^{21\%O2}$ predictions with a MAE <4.00 µM (Figure S5) and Sc/o predictions with a MAE <7.00 mol mol^{-1} (Figure S6).

Visualisation of the RbcL learned and classical encodings used during GP training

To investigate how the GPs learned to predict form IB kinetics, the RbcL sequence classical and learned encodings used for model training were visualised using t-distributed stochastic neighbour embedding (t-SNE) (Figure 3 and Figure S7). Both the classical and learned encodings show some sequences with higher Kcat, $Kc^{21\%O2}$ and Sc/o cluster together and some sequences with lower Kcat, $Kc^{21\%O2}$ and Sc/o cluster together. Differences between the RbcL classical and learned encodings are unclear for $Kc^{21\%O2}$ and Sc/o but more clustering in the learned encodings than the classical encodings can be seen for Kcat.

Assessing GP uncertainty estimates

Generally, it is assumed that GP predictions with high σ^2 most likely arises from parts of the trained GPs from which less or less similar training data was included. However, because the σ^2 estimates are a trainable part of the model, the reliability of the predicted σ^2 was assessed before guiding the selection of appropriate predictions.

Figure S8 and S9 demonstrates correlations between predicted σ^2 estimates and true error from leave-one-out and leave-genus out cross validation. No clear trend was observed between predicted σ^2 estimates and true error. Figure S10 shows uncertainty from leave-genus-out cross validation assessed using coverage and width. Most genus groups exhibit high coverage and varying average width (4σ) but some do not. As predicted mean values become increasingly out of distribution, ideal models should increase width indicating model uncertainty while coverage remains high.

Assessing RbcL sequence-space predictions with trait data

The final goal was to screen the kinetic properties of thousands of Rubisco variants *in silico* using the GPs with the learned encoding scheme. Predictions were made for 9052 unique RbcL sequences encoding Rubisco proteins from land plants. Grouping predictions by photosynthesis metabolism type revealed significant differences between Kcat, Sc/o and $Kc^{21\%O2}$ of C_3, C_4 and CAM groups (Figure S11). Grouping predictions by taxonomical group revealed significant differences between most groups except the Kcat of angiosperms and ferns, and $Kc^{21\%O2}$ of gymnosperms and bryophytes (Figure S12).

Because the predicted σ^2 estimates from cross validation showed no clear trend (Figure S8-S10), a criteria for determining the quality of predictions in the absence of experimental data could not be specified. Instead, the influence of outliers was assessed by removing predictions outside the ranges of the training dataset. Most kinetic predictions were within the range for Kcat (1.4, 7.1), $Kc^{21\%O2}$ (7, 42) and Sc/o (58, 121). (Figure 4 vs Figure S11, Figure 5 vs Figure S12). The overall trend in kinetics remained the same as before. For instance, Rubiscos from CAM and C_4 plants have a higher median Kcat than Rubiscos from C_3 plants. Similarly, the overall trend remained the same when grouping predictions by taxonomical type. For instance, angiosperms and ferns have a higher median Kcat than bryophytes and gymnosperms.

Discussion

This work presents a useful tool for screening and predicting plant Rubisco kinetics for engineering efforts as well as for fundamental studies on Rubisco evolution and adaptation. Advancements in protein language modelling has allowed the exploitation of existing plant Rubisco data for predicting Rubisco kinetics *in silico*. Further, our predictions followed well established trends observed by previous studies in plants with different photosynthetic types without *a priory* knowledge. For example, generally Rubiscos from C_4 plants have a higher Kcat, $Kc^{21\%O2}$ and lower Sc/o than Rubiscos from C_3 plants (Galmés et al., 2014b, Galmés et al., 2015, Hermida-Carrera et al., 2016, Prins et al., 2016, Galmés et al., 2019, Iñiguez et al., 2020). In contrast, CAM plants have a similar mean Kcat to that of C_4 plants (Hermida-Carrera et al., 2020, Iñiguez et al., 2020).

The kinetic properties of modern Rubiscos are believed to be shaped by changes in atmospheric CO_2 and O_2 concentrations and temperature over time (Tcherkez et al., 2006, Savir et al., 2010, Studer et al., 2014, Hermida-Carrera et al., 2016, Cummins et al., 2018, Tcherkez et al., 2018, Moore et al., 2021). C_4 and CAM plants both possess CCMs that enhance CO_2 concentration near the Rubisco active site (Raven and Beardall, 2014, Raven et al., 2017, Young and Hopkinson, 2017, Ruban et al., 2022). CCMs in C_4 and CAM plants may have first arisen in high O_2/CO_2 ratio environments and a decrease in O_2/CO_2 ratio over several million years led to the present day maintenance of high Kcat values to cope with higher mesophyll CO_2 concentrations (Cc) (Iñiguez et al., 2020). Because both C_4 and CAM plants are also found in high temperature environments, CCMs also help concentrate CO_2 near the active site when the gas solubility of atmospheric CO_2/O_2 ratio decreases with increasing temperature (Raven et al., 2017, Iñiguez et al., 2020). Despite the presence of CCMs in both C_4 and CAM plants and similar mean Kcat values, both groups had significantly different mean $Kc^{21\%O2}$ and Sc/o. C_4 plants may have evolved higher $Kc^{21\%O2}$ and lower Sc/o because of the adoption of the CCMs led to a reduced requirement for a higher Sc/o and lower $Kc^{21\%O2}$ (Iñiguez et al., 2020). On the other hand, unlike C_3 and C_4 plants, CAM plants have evolved to fix CO_2 over the course of a day in phases and are commonly found in drier climates (Leverett et al., 2021, Ruban et al., 2022). One possibility is that the temporal separation of CAM CO_2 fixation may hinder the use of CCMs during some periods leading to the requirement for a similar mean Sc/o to that of C_3 plants and lower mean $Kc^{21\%O2}$ (Iñiguez et al., 2020).

Additionally, land plant Rubiscos are characteristic of the ecological or taxonomical group from which they originated (Figure 5) (Galmés et al., 2014b). For instance, angiosperms has the largest distribution in kinetics because it is the largest and most diverse group of land plants comprising Rubiscos from C_3, C_4 and CAM plants.

What is unclear is how the GPs mapped the Rubisco sequence-function landscape. Projecting the classical and learned encodings suggests that some encodings with similar kinetics cluster together but some do not (Figure 3 and S7). Instead, the GPs may have found something ‘deeper’ about the relationship between RbcL encodings and kinetics during training. During training, when a single kernel function was applied over all encoding input positions the models performed poorly compared with an additive kernel. This suggests a complex relationship which depends on the sum of small functions rather than on a single large modelled function.

There are several strengths and limitations of the techniques used in this study. Firstly, one can assume that the training dataset only represented a fraction of all land plant Rubisco diversity. As a starting point the first logical step was to test the model on this currently available data before spending more time and resources on creating a more comprehensively rich training dataset that may reveal more subtle parts of the sequence-function landscape (Hsu et al., 2022). In fact, when removing predictions outside the ranges of the training dataset (e.g. Figure 4 vs Figure S4) there was no change in the kinetic trends suggesting predictions for most land plant Rubiscos are similar to the training dataset. We would be cautious about extending the current trained models to other Rubisco forms such as those found in bacteria and archaea, which exhibit greater sequence and kinetic diversity than form IB Rubiscos. For example, Davidi et al. (2020) identified form II Rubiscos with the fastest having a Kcat of 22 s^{-1} which is far greater than all known plant Rubiscos. As more experimental data becomes available we expect models on more Rubisco forms to be built. Secondly, the models in this study assumed that features of the RbcL determines the kinetic properties of form IB Rubiscos. While over the past few years this assumption is largely thought to be true because a) the active site is encoded by the RbcL sequence and b) the RbcL sequence is largely conserved over time as chloroplast-encoded genes evolved slower than nuclear-encoded genes (Kelly, 2021). It is now well established that the Rubisco small subunit encoded by the RbcS gene can influence catalysis too (Spreitzer et al., 2005, Genkov and Spreitzer, 2009, Atkinson et al., 2017, Martin-Avila et al., 2020, Lin et al., 2021, Sakoda et al., 2021). It would be interesting to see if incorporating RbcS sequences alongside RbcL sequences could improve the predictive power of our models. However, incorporating the RbcS *in silico* is further complicated by the existence of multiple RbcS genes located in the nucleus and different nuclear-encoded RbcS genes differentially influencing Rubisco kinetics in the same plant (Khumsupan et al., 2020, Martin-Avila et al., 2020). Further, the models in this paper can be used in thought experiments to predict the kinetics of novel Rubisco variants created *in silico* by manipulation of the Rubisco sequence potentially creating better enzymes. Lastly, the learned encoding scheme adopted in this study was a pre-trained neural network capable of predicting protein sequence features across numerous protein families without any knowledge of Rubisco kinetics. In future, we aim to improve model performance by making the neural network of the learned encoding scheme a trainable part of the GP models (also known as end-to-end learning) i.e. fine-tune the learned encoding scheme specifically for Rubisco sequence-function tasks.

Conclusion

Overall, this study is the first to demonstrate the prediction of land plant Rubisco kinetics from RbcL sequence data. This study provides plant biologists with a pre-screening tool for highlighting Rubisco species exhibiting better kinetics for crop engineering efforts. Going forward we expect more experimental data to become available, which will facilitate the development of richer models.

Supplementary data

Figure S1. Leave-one-out cross validation results for GPs using a single Matern 5/2 kernel. **Figure S2.** Leave-one-out cross validation results for GPs using an additive linear kernel. **Figure S3.** Leave-one-cross validation results for GPs using a single linear kernel. **Figure S4.** Leave-genus-out cross validation plots for Kcat. **Figure S5.** Leave-genus-out cross validation plots for Kc$^{21\%O2}$. **Figure S6.** Leave-genus-out cross validation plots for Sc/o. **Figure S7.** Visualization of the RbcL classical encodings used during GP training. **Figure S8.** Spearman rank correlations of the leave-one-out cross validation predicted uncertainties and true errors. **Figure S9.** Spearman rank correlations of the leave-genus-out cross validation predicted uncertainties and true errors. **Figure S10.** Leave-genus-out cross validation predicted uncertainties assessed using the coverage and width method. **Figure S11.** Box plots depicting kinetic predictions for all land plant Rubiscos grouped by photosynthesis metabolism type. **Figure S12.** Box plots depicting kinetic predictions for all land plant Rubiscos grouped by taxonomical type. **Table S1.** Rubisco experimental kinetics and Rubisco large subunit (RbcL) sequences for training Gaussian process models. **Table S2.** Kinetic predictions for all known land plant Rubisco large subunit (RbcL) sequences isolated from non-redundant protein databases. **Table S3.** Table S2 kinetic predictions within the training dataset ranges for Kcat (1.4, 7.1), Kc$^{21\%O2}$ (7, 42), and Sc/o (58, 121). **Table S4.** All wild-type Rubisco large subunit (RbcL) sequences isolated from non-redundant protein databases (Davidi et al., 2020).

Conflict of interest

The authors have no conflict of interests to declare.

Data availability

https://github.com/Iqbalwasim01/Mining-Rubisco-kinetics.git

Acknowledgements

W.I. would like to thank ONEplanet DTP Natural Environment Research Council (NERC) program for funding his PhD research (NE/S007512/1) . W.I. would like to thank Rebecca L Moore, Iain J Hope, Daniel Cowan-Turner and Bethan Morris for discussions on Rubisco.

Authors contribution

W.I and M.K. conceived the idea of the study. W.I. developed the models and performed analyses. Contents of the paper was developed and written by W.I. with input from M.K. and A.L.

References

Alquraishi, M. 2021. Machine learning in protein structure prediction. *Current Opinion in Chemical Biology,* 65, 1-8.

Atkinson, N., Leitão, N., Orr, D. J., Meyer, M. T., Carmo-Silva, E., Griffiths, H., Smith, A. M. & Mccormick, A. J. 2017. Rubisco small subunits from the unicellular green alga Chlamydomonas complement Rubisco-deficient mutants of Arabidopsis. *New Phytologist,* 214, 655-667.

Bar-On, Y. M. & Milo, R. 2019. The global mass and average rate of rubisco. *Proceedings of the National Academy of Sciences,* 116, 4738-4743.

Bedbrook, C. N., Yang, K. K., Rice, A. J., Gradinaru, V. & Arnold, F. H. 2017. Machine learning to design integral membrane channelrhodopsins for efficient eukaryotic expression and plasma membrane localization. *PLOS Computational Biology,* 13, e1005786.

Benes, B., Guan, K., Lang, M., et al. 2020. Multiscale computational models can guide experimentation and targeted measurements for crop improvement. *The Plant Journal,* 103, 21-31.

Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M. & Church, G. M. 2021. Low-N protein engineering with data-efficient deep learning. *Nature Methods,* 18, 389-396.

Bodenhofer, U., Bonatesta, E., Horejš-Kainrath, C. & Hochreiter, S. 2015. msa: an R package for multiple sequence alignment. *Bioinformatics,* 31, 3997-3999.

Bonetta, R. & Valentino, G. 2020. Machine learning techniques for protein function prediction. *Proteins: Structure, Function, and Bioinformatics,* 88, 397-413.

Bouvier, J. W., Emms, D. M., Rhodes, T., et al. 2021. Rubisco Adaptation Is More Limited by Phylogenetic Constraint Than by Catalytic Trade-off. *Molecular Biology and Evolution,* 38, 2880–2896.

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N. & Linial, M. 2022. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics,* 38, 2102-2110.

Camel, V. & Zolla, G. 2021. An Insight of RuBisCO Evolution through a Multilevel Approach. *Biomolecules,* 11, 1761.

Cummins, P. L., Kannappan, B. & Gready, J. E. 2018. Directions for Optimization of Photosynthetic Carbon Fixation: RuBisCO's Efficiency May Not Be So Constrained After All. *Frontiers in Plant Science,* 9.

Davidi, D., Shamshoum, M., Guo, Z., et al. 2020. Highly active rubiscos discovered by systematic interrogation of natural sequence diversity. *The EMBO Journal,* 39, e104081.

Deringer, V. L., Bartók, A. P., Bernstein, N., Wilkins, D. M., Ceriotti, M. & Csányi, G. 2021. Gaussian Process Regression for Materials and Molecules. *Chemical Reviews,* 121, 10073-10141.

Dutordoir, V., Salimbeni, H., Hambro, E., et al. 2021. GPflux: A Library for Deep Gaussian Processes. *arXiv preprint arXiv:2104.05674*.

Duvenaud, D., Nickisch, H. & Rasmussen, C. E. 2011. Additive gaussian processes. *arXiv preprint arXiv:1112.4394*.

Elabd, H., Bromberg, Y., Hoarfrost, A., Lenz, T., Franke, A. & Wendorff, M. 2020. Amino acid encoding for deep learning applications. *BMC Bioinformatics,* 21, 235.

Elnaggar, A., Heinzinger, M., Dallago, C., et al. 2021. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning. *bioRxiv*, 2020.07.12.199554.

Faulon, J.-L. & Faure, L. 2021. In silico, in vitro, and in vivo machine learning in synthetic biology and metabolic engineering. *Current Opinion in Chemical Biology,* 65, 85-92.

Flamholz, A. I., Prywes, N., Moran, U., Davidi, D., Bar-On, Y. M., Oltrogge, L. M., Alves, R., Savage, D. & Milo, R. 2019. Revisiting Trade-offs between Rubisco Kinetic Parameters. *Biochemistry,* 58, 3365-3376.

Galmés, J., Andralojc, P. J., Kapralov, M. V., Flexas, J., Keys, A. J., Molins, A., Parry, M. a. J. & Conesa, M. À. 2014a. Environmentally driven evolution of Rubisco and improved photosynthesis and growth within the C3 genus Limonium (Plumbaginaceae). *New Phytologist,* 203**,** 989-999.

Galmés, J., Capó-Bauçà, S., Niinemets, Ü. & Iñiguez, C. 2019. Potential improvement of photosynthetic CO2 assimilation in crops by exploiting the natural variation in the temperature response of Rubisco catalytic traits. *Current Opinion in Plant Biology,* 49**,** 60-67.

Galmes, J., Conesa, M. A., Diaz-Espejo, A., Mir, A., Perdomo, J. A., Niinemets, U. & Flexas, J. 2014. Rubisco catalytic properties optimized for present and future climatic conditions. *Plant Science,* 226**,** 61-70.

Galmés, J., Kapralov, M. V., Andralojc, P. J., Conesa, M., Keys, A. J., Parry, M. A. & Flexas, J. 2014b. Expanding knowledge of the Rubisco kinetics variability in plant species: environmental and evolutionary trends. *Plant, Cell & Environment,* 37**,** 1989-2001.

Galmés, J., Kapralov, M. V., Copolovici, L. O., Hermida-Carrera, C. & Niinemets, Ü. 2015. Temperature responses of the Rubisco maximum carboxylase activity across domains of life: phylogenetic signals, trade-offs, and importance for carbon gain. *Photosynthesis Research,* 123**,** 183-201.

Garnett, R. 2022. *Bayesian optimization,* In preparation, Cambridge University Press.

Genkov, T. & Spreitzer, R. J. 2009. Highly conserved small subunit residues influence rubisco large subunit catalysis. *Journal of Biological Chemistry,* 284**,** 30105-30112.

Greenhalgh, J. C., Fahlberg, S. A., Pfleger, B. F. & Romero, P. A. 2021. Machine learning-guided acyl-ACP reductase engineering for improved in vivo fatty alcohol production. *Nature Communications,* 12**,** 5825.

Greenman, K. P., Soleimany, A. & Yang, K. K. 2022. Benchmarking Uncertainty Quantification for Protein Engineering. *ICLR2022 Machine Learning for Drug Discovery.*

Gunn, L. H., Avila, E. M., Birch, R. & Whitney, S. M. 2020. The dependency of red Rubisco on its cognate activase for enhancing plant photosynthesis and growth. *Proceedings of the National Academy of Sciences,* 117**,** 25890-25896.

Hermida-Carrera, C., Fares, M. A., Font-Carrascosa, M., et al. 2020. Exploring molecular evolution of Rubisco in C3 and CAM Orchidaceae and Bromeliaceae. *BMC Evolutionary Biology,* 20**,** 11.

Hermida-Carrera, C., Kapralov, M. V. & Galmés, J. 2016. Rubisco Catalytic Properties and Temperature Response in Crops. *Plant Physiology,* 171**,** 2549-2561.

Hsu, C., Nisonoff, H., Fannjiang, C. & Listgarten, J. 2022. Learning protein fitness models from evolutionary and assay-labeled data. *Nature Biotechnology*.

Iñiguez, C., Capó-Bauçà, S., Niinemets, Ü., Stoll, H., Aguiló-Nicolau, P. & Galmés, J. 2020. Evolutionary trends in RuBisCO kinetics and their co-evolution with CO2 concentrating mechanisms. *The Plant Journal,* 101**,** 897-918.

Iqbal, W. A., Miller, I. G., Moore, R. L., Hope, I. J., Cowan-Turner, D. & Kapralov, M. V. 2021. Rubisco substitutions predicted to enhance crop performance through carbon uptake modelling. *Journal of Experimental Botany,* 72**,** 6066-6075.

Jokinen, E., Heinonen, M. & Lähdesmäki, H. 2018. mGPfusion: predicting protein stability changes with Gaussian process kernel learning and data fusion. *Bioinformatics,* 34**,** i274-i283.

Jordan, D. B. & Ogren, W. L. 1983. Species variation in kinetic properties of ribulose 1,5-bisphosphate carboxylase/oxygenase. *Archives of Biochemistry and Biophysics,* 227**,** 425-433.

Kanevski, I., Maliga, P., Rhoades, D. F. & Gutteridge, S. 1999. Plastome engineering of ribulose-1, 5-bisphosphate carboxylase/oxygenase in tobacco to form a sunflower large subunit and tobacco small subunit hybrid. *Plant Physiology,* 119**,** 133-142.

Kapralov, M. V. & Filatov, D. A. 2006. Molecular Adaptation during Adaptive Radiation in the Hawaiian Endemic Genus Schiedea. *PLOS ONE,* 1**,** e8.

Kapralov, M. V., Kubien, D. S., Andersson, I. & Filatov, D. A. 2011. Changes in Rubisco Kinetics during the Evolution of C4 Photosynthesis in Flaveria (Asteraceae) Are Associated with Positive Selection on Genes Encoding the Enzyme. *Molecular Biology and Evolution,* 28**,** 1491-1503.

Kellogg, E. & Juliano, N. 1997. The structure and function of RuBisCO and their implications for systematic studies. *American Journal of Botany,* 84**,** 413.

Kelly, S. 2021. The economics of organellar gene loss and endosymbiotic gene transfer. *Genome Biology,* 22**,** 345.

Khumsupan, P., Kozlowska, M. A., Orr, D. J., Andreou, A. I., Nakayama, N., Patron, N., Carmo-Silva, E. & Mccormick, A. J. 2020. Generating and characterizing single- and multigene mutants of the Rubisco small subunit family in Arabidopsis. *Journal of Experimental Botany,* 71**,** 5963-5975.

Kompa, B., Snoek, J. & Beam, A. L. 2021. Empirical frequentist coverage of deep learning uncertainty quantification procedures. *Entropy,* 23**,** 1608.

Kubien, D. S., Whitney, S. M., Moore, P. V. & Jesson, L. K. 2008. The biochemistry of Rubisco in Flaveria. *J Exp Bot,* 59**,** 1767-1777.

Lehnherr, B., Mächler, F. & Nösberger, J. 1985. Influence of Temperature on the Ratio of Ribulose Bisphosphate Carboxylase to Oxygenase Activities and on the Ratio of Photosynthesis to Photorespiration of Leaves. *J Exp Bot,* 36**,** 1117-1125.

Leverett, A., Hurtado Castaño, N., Ferguson, K., Winter, K. & Borland, A. M. 2021. Crassulacean acid metabolism (CAM) supersedes the turgor loss point (TLP) as an important adaptation across a precipitation gradient, in the genus Clusia. *Functional Plant Biology,* 48**,** 703-716.

Li, G., Dong, Y. & Reetz, M. T. 2019. Can Machine Learning Revolutionize Directed Evolution of Selective Enzymes? *Advanced Synthesis & Catalysis,* 361**,** 2377-2386.

Lin, M. T., Orr, D. J., Worrall, D., Parry, M. A., Carmo‐Silva, E. & Hanson, M. R. 2021. A procedure to introduce point mutations into the Rubisco large subunit gene in wild‐type plants. *The Plant Journal,* 106**,** 876-887.

Lin, M. T., Salihovic, H., Clark, F. K. & Hanson, M. R. 2022. Improving the efficiency of Rubisco by resurrecting its ancestors in the family Solanaceae. *Science Advances,* 8**,** eabm6871.

Long, B. M., Hee, W. Y., Sharwood, R. E., et al. 2018. Carboxysome encapsulation of the CO2-fixing enzyme Rubisco in tobacco chloroplasts. *Nature Communications,* 9**,** 3570.

Maaten, L. V. D. & Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research,* 9**,** 2579-2605.

Martin-Avila, E., Lim, Y.-L., Birch, R., Dirk, L. M. A., Buck, S., Rhodes, T., Sharwood, R. E., Kapralov, M. V. & Whitney, S. M. 2020. Modifying Plant Photosynthesis and Growth via Simultaneous Chloroplast Transformation of Rubisco Large and Small Subunits. *The Plant Cell,* 32**,** 2898-2916.

Matthews, A. G. D. G., Van Der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrá, P., Ghahramani, Z. & Hensman, J. 2017. GPflow: A Gaussian Process Library using TensorFlow. *Journal of Machine Learning Research,* 18**,** 1-6.

Mazurenko, S., Prokop, Z. & Damborsky, J. 2020. Machine Learning in Enzyme Engineering. *ACS Catalysis,* 10**,** 1210-1223.

Moore, C. E., Meacham-Hensold, K., Lemonnier, P., Slattery, R. A., Benjamin, C., Bernacchi, C. J., Lawson, T. & Cavanagh, A. P. 2021. The effect of increasing temperature on crop photosynthesis: from enzymes to ecosystems. *Journal of Experimental Botany,* 72**,** 2822-2844.

Newman, S. J. & Furbank, R. T. 2021. Explainable machine learning models of major crop traits from satellite-monitored continent-wide field trial data. *Nature Plants,* 7**,** 1354–1363.

Orr, D. J., Alcântara, A., Kapralov, M. V., Andralojc, P. J., Carmo-Silva, E. & Parry, M. a. J. 2016. Surveying Rubisco Diversity and Temperature Response to Improve Crop Photosynthetic Efficiency. *Plant Physiology,* 172**,** 707-717.

Orr, D. J. & Parry, M. a. J. 2020. Overcoming the limitations of Rubisco: fantasy or realistic prospect? *Journal of Plant Physiology,* 254**,** 153285.

Ort, D. R., Merchant, S. S., Alric, J., et al. 2015. Redesigning photosynthesis to sustainably meet global food and bioenergy demand. *Proceedings of the National Academy of Sciences,* 112**,** 8529-8536.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research,* 12**,** 2825-2830.

Prins, A., Orr, D. J., Andralojc, P. J., Reynolds, M. P., Carmo-Silva, E. & Parry, M. A. 2016. Rubisco catalytic properties of wild and domesticated relatives provide scope for improving wheat photosynthesis. *J Exp Bot,* 67**,** 1827-38.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P. & Song, Y. S. 2019. Evaluating protein transfer learning with TAPE. *Advances in neural information processing systems,* 32**,** 9689.

Rasmussen, C. E. & Williams, C. 2006. *Gaussian processes for machine learning,* Cambridge, The MIT Press.

Raven, J. A. 2013. Rubisco: still the most abundant protein of Earth? *New Phytologist,* 198**,** 1-3.

Raven, J. A. & Beardall, J. 2014. CO2 concentrating mechanisms and environmental change. *Aquatic Botany,* 118**,** 24-37.

Raven, J. A., Beardall, J. & Sánchez-Baracaldo, P. 2017. The possible evolution and future of CO2-concentrating mechanisms. *Journal of Experimental Botany,* 68**,** 3701-3716.

Rives, A., Meier, J., Sercu, T., et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences,* 118**,** e2016239118.

Romero, P. A., Krause, A. & Arnold, F. H. 2013. Navigating the protein fitness landscape with Gaussian processes. *Proceedings of the National Academy of Sciences,* 110**,** E193-E201.

Ruban, A. V., Murchie, E. & Foyer, C. H. 2022. *Photosynthesis in Action,* London, Academic Press.

Sakoda, K., Yamamoto, A., Ishikawa, C., Taniguchi, Y., Matsumura, H. & Fukayama, H. 2021. Effects of introduction of sorghum RbcS with rice RbcS knockdown by RNAi on photosynthetic activity and dry weight in rice. *Plant Production Science,* 24**,** 346-353.

Savir, Y., Noor, E., Milo, R. & Tlusty, T. 2010. Cross-species analysis traces adaptation of Rubisco toward optimality in a low-dimensional landscape. *Proceedings of the National Academy of Sciences,* 107**,** 3475-3480.

Sharwood, R. E. 2017. Engineering chloroplasts to improve Rubisco catalysis: prospects for translating improvements into food and fiber crops. *New Phytologist,* 213**,** 494-510.

Sharwood, R. E., Ghannoum, O., Kapralov, M. V., Gunn, L. H. & Whitney, S. M. 2016. Temperature responses of Rubisco from Paniceae grasses provide opportunities for improving C3 photosynthesis. *Nature Plants,* 2**,** 16186.

Spreitzer, R. J., Peddi, S. R. & Satagopan, S. 2005. Phylogenetic engineering at an interface between large and small subunits imparts land-plant kinetic properties to algal Rubisco. *Proceedings of the National Academy of Sciences,* 102**,** 17225-17230.

Studer, R. A., Christin, P.-A., Williams, M. A. & Orengo, C. A. 2014. Stability-activity tradeoffs constrain the adaptive evolution of RubisCO. *Proceedings of the National Academy of Sciences,* 111**,** 2223-2228.

Tcherkez, G. G., Bathellier, C., Farquhar, G. D. & Lorimer, G. H. 2018. Commentary: Directions for Optimization of Photosynthetic Carbon Fixation: RuBisCO's Efficiency May Not Be So Constrained After All. *Frontiers in Plant Science,* 9**,** 183.

Tcherkez, G. G. B., Farquhar, G. D. & Andrews, T. J. 2006. Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases may be nearly perfectly optimized. *Proceedings of the National Academy of Sciences,* 103**,** 7246-7251.

Uemura, K., Anwaruzzaman, Miyachi, S. & Yokota, A. 1997. Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase from Thermophilic Red Algae with a Strong Specificity for CO2Fixation. *Biochemical and Biophysical Research Communications,* 233**,** 568-571.

Viil, J., Ivanova, H. & Pärnik, T. 2012. Specificity factor of Rubisco: estimation in intact leaves by carboxylation at different CO2/O2 ratios. *Photosynthetica,* 50**,** 247-253.

Von Caemmerer, S. 2000. *Biochemical Models of Leaf Photosynthesis,* Australia, CSIRO.

Von Caemmerer, S. 2020. Rubisco carboxylase/oxygenase: From the enzyme to the globe: A gas exchange perspective. *Journal of Plant Physiology***,** 153240.

Whitney, S. M., Birch, R., Kelso, C., Beck, J. L. & Kapralov, M. V. 2015. Improving recombinant Rubisco biogenesis, plant photosynthesis and growth by coexpressing its ancillary RAF1 chaperone. *Proceedings of the National Academy of Sciences,* 112**,** 3564.

Whitney, S. M., Sharwood, R. E., Orr, D., White, S. J., Alonso, H. & Galmés, J. 2011. Isoleucine 309 acts as a C4 catalytic switch that increases ribulose-1, 5-bisphosphate carboxylase/oxygenase (rubisco) carboxylation rate in Flaveria. *Proceedings of the National Academy of Sciences,* 108**,** 14688-14693.

Wilson, R. H., Alonso, H. & Whitney, S. M. 2016. Evolving Methanococcoides burtonii archaeal Rubisco for improved photosynthesis and plant growth. *Scientific Reports,* 6**,** 22284.

Wilson, R. H., Martin-Avila, E., Conlan, C. & Whitney, S. M. 2018. An improved Escherichia coli screen for Rubisco identifies a protein–protein interface that can enhance CO2-fixation kinetics. *Journal of Biological Chemistry,* 293**,** 18-27.

Wittmann, B. J., Johnston, K. E., Wu, Z. & Arnold, F. H. 2021. Advances in machine learning for directed evolution. *Current Opinion in Structural Biology,* 69**,** 11-18.

Yang, K. K., Wu, Z. & Arnold, F. H. 2019. Machine-learning-guided directed evolution for protein engineering. *Nature methods,* 16**,** 687-694.

Yang, K. K., Wu, Z., Bedbrook, C. N. & Arnold, F. H. 2018. Learned protein embeddings for machine learning. *Bioinformatics,* 34**,** 2642-2648.

Young, J. N. & Hopkinson, B. M. 2017. The potential for co-evolution of CO2-concentrating mechanisms and Rubisco in diatoms. *Journal of Experimental Botany,* 68**,** 3751-3762.

Zhou, Y. & Whitney, S. 2019. Directed evolution of an improved Rubisco; in vitro analyses to decipher fact from fiction. *International journal of molecular sciences,* 20**,** 5019.

Zhu, X.-G., Ort, D. R., Parry, M. a. J. & Von Caemmerer, S. 2020. A wish list for synthetic biology in photosynthesis research. *Journal of Experimental Botany,* 71**,** 2219-2225.

Figure 1. Schematic diagram showing steps involved in training a Gaussian process (GP) regression. (A) Rubisco large subunit (RbcL) sequences can be converted to either a binary representation (classical encodings) which explicitly represents the amino acids or learned encodings (such as: Rives et al. (2021)) which involves another machine learning method-learning key features of each sequence (such as physiochemical properties or secondary structures) and storing these features as numerical vectors. The encoded RbcL sequences are stored in a kernel which describes the similarity between the encoded sequences. A kernel function can be applied to each input feature of the encodings. For example, $k(x_1)$ would encode the first numerical input for the learned encodings or the first alignment position for the classical encodings. Alternatively, input features can vary simultaneously using a single kernel function. (B) During model training, hyperparameters such as the length scale (l) and/or variance (σ^2) are optimised to find functions ($f(x)$) that describe the relationship between the RbcL encodings and associated labels (e.g. turnover rate: Kcat). The l describes the horizontal distances between $f(x)$, and σ^2 the vertical distance (i.e. noise and signal). As such, GPs provide a flexible framework for explaining numerous relationships.

Figure 2. Comparison between predicted and actual carboxylation turnover rate (Kcat : s^{-1}), Michaelis-Menten constant for CO_2 at ambient O_2 ($Kc^{21\%O2}$: µM) and specificity for CO_2 over O_2 (Sc/o: mol mol^{-1}) at $25^{O}C$. Determined using leave-one-out cross-validation with the learned encoding scheme (Rives et al., 2021) (green) and classical encoding scheme (orange). The better performance of the learned encoding with an additive non-linear kernel justified the adoption of this method over classical for the final machine learning tasks.

Figure 3. Visualization of the Rubisco large subunit (RbcL) learned encodings used in the fully trained Gaussian process (GP) models. Each data point represents an RbcL learned encoding with (A) carboxylation turnover rate (Kcat: s^{-1}), (B) Michaelis-Menten constant for CO_2 at ambient atmospheric O_2 ($Kc^{21\%O2}$: µM) and (C) specificity for CO_2 over O_2 (Sc/o: mol mol^{-1}).

Figure 4. Box plots depict (A) carboxylation turnover rate (Kcat: s^{-1}), (B) Michaelis-Menten constant for CO_2 at ambient atmospheric O_2 ($Kc^{21\%O2}$: μM) and (C) specificity for CO_2 over O_2 (Sc/o: mol mol^{-1}) predictions made for the form IB (plants) Rubisco large subunit (RbcL) sequence-space using the fully trained Gaussian process (GP) models with the learned encoding scheme. Shown are predictions within the ranges of the training dataset for Kcat (1.4, 7.1), $Kc^{21\%O2}$ (7, 42) and Sc/o (58, 121). Predictions were grouped by photosynthesis metabolism type ($C_{3,}$ C_4 or CAM). Box plot horizontal lines show the median value, and the box and whisker represent the 25th and 75th percentile and minimum to maximum distributions of the data. Significant differences from the one-way ANOVA with Duncan's post hoc test are shown for groups: *** $p<0.001$, ** $p<0.01$, * $p<0.05$, n.s., non significant.

Figure 5. Box plots depict (A) carboxylation turnover rate (Kcat: s^{-1}), (B) Michaelis-Menten constant for CO_2 at ambient atmospheric O_2 ($Kc^{21\%O2}$: μM) and (C) specificity for CO_2 over O_2 (Sc/o: mol mol^{-1}) predictions made for the form IB (plants) Rubisco large subunit (RbcL) sequence-space using the fully trained Gaussian process (GP) models with the learned encoding scheme. Shown are predictions within the ranges of the training dataset for Kcat (1.4, 7.1), $Kc^{21\%O2}$ (7, 42) and Sc/o (58, 121). Predictions were grouped by taxonomical type (Angiosperms, 'Ferns' (including Pteridophytes and Lycopodiophytes), Gymnosperms or Bryophytes). Box plot horizontal lines show the median value, and the box and whisker represent the 25th and 75th percentile and minimum to maximum distributions of the data. Significant differences from the one-way ANOVA with Duncan's post hoc test are shown for groups: *** $p<0.001$, ** $p<0.01$, * $p<0.05$, n.s., non significant.