

Method for enzyme fitness data collection

Technical field of the invention

The present invention relates to a method for predicting or determining fitness of enzyme variants comprising generating a DNA variant library of a given enzyme template and linking a unique DNA barcode to each DNA variant, wherein the DNA variants are subjected to long-read and short-read sequencing. In particular, the present invention relates to a method for estimating fitness of enzyme variants and/or optimizing a nucleotide sequence encoding an enzyme variant by combining DNA barcodes, selection pressure based on enzyme catalytic activity, long-read sequencing, and short-read sequencing.

Background of the invention

While machine learning and deep learning have been applied successfully toward many protein-related problems, for example structure prediction and enzyme classification, machine learning has not been particularly powerful for predicting enzyme activities. This is due to a lack of data. Enzyme data collection is typically time- and resource-intensive. Although many advanced machine learning algorithms have been developed over the recent years, machine learning cannot yield good performance without sufficient data. Data for application in machine learning for predicting enzyme fitness should be of high quality, such as comprising full-length enzyme sequence, and in large amounts.

Yang *et al.* (2021) (Defining protein variant functions using high-complexity mutagenesis libraries and enhanced mutant detection software ASMv1.0.) describes a method for screening of a variant expression library for a gene of interest for characterizing the functions of thousands of single-amino-acid substitutions in a single experiment. The method comprises characterizing protein fitness of several variants by attaching barcodes to the gene of interest, amplification and expression in cells, growth selection, and deep sequencing. Further, while Yang *et al.* mentioned that long read sequencing is possible for associating a set of unique external barcodes to each of the variants in a saturation variant library, so that the screen deconvolution work can be achieved by direct next-generation sequencing of the barcodes, they are doubtful about the

feasibility of this approach, and are silent about applying this approach to obtaining enzyme activity data.

A review paper by Wei H, Li X. (2023) (Deep mutational scanning: A versatile tool in systematically mapping genotypes to phenotypes. Front Genet. 2023 Jan 12)

5 proposed the possibility of constructing a barcoded mutant variant library, cell growth selection, associating each barcode with corresponding variant using long-read sequencing, and afterwards performing deep sequencing of the barcodes. Wei H, Li X. are silent in respect of coupling cell growth to the catalytic activity of enzymes, and are silent about using such an approach for enzyme activity data
10 collection. Wei H, Li X. are also silent about using data obtained in such a manner to predict enzyme sequences or sequence libraries with potentially improved activity.

Fowler DM *et al.* (2014) (Measuring the activity of protein variants on a large scale using deep mutational scanning. Nat Protoc. 2014 Sep) describes analysis of
15 big dataset relating to protein activity within a short time period. The technique comprises attachment of a unique DNA barcode on each variant and combines selection with high-throughput DNA sequencing, for a library of tens to hundreds of thousands of variants of some protein and imposes a selection for function, including growth-based selections. High-throughput DNA sequencing measures
20 the frequency of each variant during the selection experiment. The barcodes are used to create a database by performing deep sequencing and the frequency of each variant is identified by deep sequencing the barcode only. The association between the barcodes and the enzyme sequences needs to be established individually through Sanger sequencing, limiting the capacity of the number of
25 variants. Fowler DM *et al.* are silent about using high-throughput long-read sequencing such as through the Nanopore or PacBio platforms for establishing such association.

WO 2022/125686 A1 provides a process for determining a coding DNA sequence for a specific dose-response curve for a genetic sensor.

30

Zhuoxing Wu *et al.* (Expression level is a major modifier of the fitness landscape of a protein coding gene. Nat Ecol Evol. 2022) describes an analysis of expression level of the fitness landscape of protein coding gene.

There is no disclosure for obtaining full-length (>100 amino acids) enzyme sequence-activity data through a combination of 1) coupling of cell growth to the catalytic activity of an enzyme; 2) molecular barcoding; 3) long-read sequencing; 4) deep sequencing, to obtain data for individual labeling of enzyme sequences with their fitness levels, nor is there any disclosure of using data obtained in such a manner to predict enzyme sequences or sequence libraries with potentially improved activity.

Hence, an improved method for predicting enzyme fitness and thereby optimizing a nucleotide sequence encoding an enzyme variant would be advantageous, and in particular a more efficient and/or reliable method for estimating enzyme fitness and/or optimizing a nucleotide sequence encoding an enzyme variant comprising high quality data in large amounts, which is advantageous for predicting enzyme sequences with improved fitness or activity through machine learning.

15 Summary of the invention

The invention presented relates to a method for estimating fitness of enzyme variants and/or for optimizing a nucleotide sequence encoding an enzyme variant. The method relates to extracting data relating to DNA sequence variants encoding enzyme variant and their fitness. The method comprises generating a DNA variant library of a given enzyme and linking a unique DNA barcode on each DNA variant. The DNA variant library is transformed into a selected strain for non-selective amplification and expression of the library. The amplified DNA variant library is extracted and subjected to long-read sequencing for generating information that associates the unique barcodes to their corresponding enzyme variants. Further, the strain comprising the DNA variant library is subjected to a selection pressure favoring enzyme variants with increased fitness as manifested by increased catalytic activity. The growth rate of the strain under selection pressure is a reflection of enzyme fitness and activity. Subsequently, the DNA variant library is extracted and subjected to short-read deep sequencing, wherein only the barcode region is sequenced. The frequency of each barcode is used as a reflection of the fitness of the corresponding enzyme variant and thereby allowing optimization of a nucleotide sequence encoding an improved enzyme variant. The barcode is used for identifying the corresponding enzyme variant.

The enzyme sequence-activity data generated from the invention will be of high quality comprising accurate full-length of the enzyme sequences, and in large amounts being a capacity of more than 10^6 data points per run with current technologies, which far exceeds the state-of-the-art. This will greatly improve the application of machine learning for enzyme activity-related tasks.

Example 1 shows the construction of the selection strain, where the activity and fitness of SAM-dependent methyltransferases is coupled to cell growth by providing precursors for cysteine synthesis. The SAM-dependent methyltransferase used in this example is *Arabidopsis thaliana* N-acetylserotonin O-methyltransferase (atASMT), and the target catalytic activity is the conversion of protocatechuic acid (PCA) into vanillic acid.

Example 2 shows the creation of a library of DNA variants linked with unique DNA barcodes. The DNA library can be obtained through error-prone PCR or other methods. The variants in the DNA library are linked with unique DNA barcodes through a one-cycle PCR process. The barcoded DNA library is subsequently assembled into a plasmid vector for expression in *E. coli*.

Example 3 shows the construction of an *in vivo* library by transforming the plasmid library, containing the DNA library for enzyme variants, into the selection strain. The *in vivo* library is further amplified by non-selective growth for different subsequent uses.

Example 4 shows Nanopore sequencing to generate data that associates the DNA variant sequences with their associated DNA barcodes. This example shows extraction of the plasmid library after growing an aliquot of the strain from Example 3 with antibiotic selection for the plasmid, but without selection for the enzyme variants. The region including the target enzyme expression cassette and barcode is isolated by restriction digestion and sequenced through amplicon sequencing with Nanopore long-read sequencing.

Example 5 shows growth of the selection strain comprising the DNA library with selection biased towards higher target enzyme activity. The cells proliferate with a growth rate positively correlated with the activity of the target enzyme variant.

Sampling during this process and the subsequent short-read sequencing to characterize the change of variant abundance/frequency, allow the calculation of the growth rate of the cell population harboring a specific enzyme variant.

- 5 Example 6 shows short-read sequencing of the barcode region for obtaining the frequency change of the DNA barcodes for cell samples over the course of selective growth. Short-read sequencing of the barcode region is performed to obtain the frequency information of the barcodes for each cell sample.
- 10 Example 7 shows processing of long-read sequencing data and short-read sequencing data obtained from Example 4 and 6, respectively, in order to generate the full DNA sequence-activity relationship. The DNA barcode is used to link the full-length DNA sequence to the corresponding growth rate.
- 15 Example 8 shows training of a machine learning model to predict amino acid sequence fitness using obtained sequence-fitness data. The top-5 most highly ranked variants of the original atASMT sequence were identified.

Example 9 shows improvement of sample preparation to improve the size,
20 diversity, and quality of the collected data.

Example 10 shows improvement of the quality and output size of the consensus sequence generated from Nanopore data, and to improve the statistical rigor of the fitness estimation for sequence variants.

25

Example 11 shows generation of a series of variant sequences using the MoCHI suite, which identifies beneficial and non-beneficial mutations by considering both individual and epistatic effects. The resulting library combines highly beneficial mutations and epistatic interactions while avoiding less beneficial ones.

30

Example 12 shows experimental validation of the data obtained from method of the invention to evaluate the performance of sequences generated based on machine learned information from the data. The estimated enzyme fitness levels obtained in the high-throughput manner are positively correlated with results in
35 the bioconversion assay. In addition, many generated sequences based on

learned mutational effects shows significant activity. These results indicate the great value of using the high-throughput dataset obtained by the method of the present invention for enzyme engineering.

5 Thus, an object of the present invention relates to the provision of an improved method for estimating enzyme fitness and/or optimizing a nucleotide sequence encoding an enzyme variant by combining barcodes, selective growth, long-read, and short-read sequencing.

In particular, it is an object of the present invention to provide a method that
10 solves the above-mentioned problems of the prior art with an improved method for determining enzyme fitness allowing optimization of a nucleotide sequence encoding an enzyme variant, which provides data of high quality, such as comprising full-length enzyme sequence, and in large amounts, which is further suitable for application in machine learning for predicting enzyme fitness and/or
15 obtaining optimized/improved enzyme variant.

Thus, one aspect of the invention relates to a method for estimating fitness of enzyme variants, the method comprising

- a) providing a DNA variant library, preferably a plasmid library, the library
20 comprises DNA variants coding for enzyme variants, wherein each DNA variant is linked to one or more unique DNA barcodes, preferably through a one-cycle PCR process;
- b) transforming the DNA variant library into a host cell for expressing the enzyme variants from the library;
 - 25 I. growing the host cell comprising the DNA variant library, preferably under conditions in the absence of a selection pressure favoring enzyme variants with increased activity; and
 - II. growing the host cell comprising the DNA variant library, under selection pressure favoring enzyme variants with increased catalytic
30 activity;
- c) isolating DNA variant library from the host cells grown under condition I.) and isolating at least the barcode region of the DNA variant library from the host cells grown under condition II.);
- d) subjecting the isolated DNA variant library grown under condition I.) to
35 long-read sequencing, thereby obtaining first set of information, such as a

first database, linking a specific DNA barcode to a specific DNA variant coding for a specific enzyme variant,

- e) subjecting the isolated DNA variant library grown under condition II.) to short-read deep sequencing, thereby obtaining a second set of information,
5 such as a second database, linking a specific DNA barcode to a frequency of the specific DNA variant coding for a specific enzyme variant or a rate of change of the specific DNA variant coding for a specific enzyme variant;
 - f) comparing the first set of information from step d) to the second set of information from step e), thereby correlating specific enzyme variants to
10 the frequency information, such as the frequency of the variant (such as in the entire population of enzyme variants); and
 - g) optionally estimating the enzyme fitness of the specific enzyme variant from the determined frequency information.
- 15 A further aspect of the invention relates to a method for optimizing a nucleotide sequence encoding an enzyme variant, the method comprising
- a) providing a DNA variant library, preferably a plasmid library, the library comprises DNA variants coding for enzyme variants, wherein each DNA variant is linked to one or more unique DNA barcodes, preferably through a
20 one-cycle PCR process;
 - b) transforming the DNA variant library into a host cell for expressing the enzyme variants from the library;
 - I. growing the host cell comprising the DNA variant library, preferably under conditions in the absence of a selection pressure favoring
25 enzyme variants with increased activity; and
 - II. growing the host cell comprising the DNA variant library, under selection pressure favoring enzyme variants with increased catalytic activity;
 - c) isolating DNA variant library from the host cells grown under condition I.) and isolating at least the barcode region of the DNA variant library from the
30 host cells grown under condition II.);
 - d) subjecting the isolated DNA variant library grown under condition I.) to long-read sequencing, thereby obtaining first set of information, such as a first database, linking a specific DNA barcode to a specific DNA variant
35 coding for a specific enzyme variant,

- e) subjecting the isolated DNA variant library grown under condition II.) to short-read deep sequencing, thereby obtaining a second set of information, such as a second database, linking a specific DNA barcode to a frequency of the specific DNA variant coding for a specific enzyme variant or a rate of change of the specific DNA variant coding for a specific enzyme variant;
- 5 f) comparing the first set of information from step d) to the second set of information from step e), thereby correlating specific enzyme variants to the frequency information, such as the frequency of the variant; and
- g) obtaining the optimized nucleotide sequence encoding an enzyme variant
- 10 from the determined frequency information.

Another aspect of the present invention relates to use of at least one barcoded DNA sequence encoding an enzyme variant for predicting enzyme fitness of said enzyme variant, wherein the at least one barcoded DNA sequence is subjected to long-read sequencing and short-read sequencing.

15

Yet another aspect of the present invention relates to use of at least one barcoded DNA sequence encoding an enzyme variant for optimizing said DNA sequence encoding said enzyme variant, wherein the at least one barcoded DNA sequence is subjected to long-read sequencing and short-read sequencing.

20

In an aspect, the invention relates to a polypeptide according to SEQ ID NO. 3 comprising one or more of the following mutations P72S, K132R, D147E, I103N, P286L, R77C, G105S, K187R, M215K, P231S, E46D, S84I, and/or P302L.

25

Another aspect of the present invention relates to processor system programmed to operate according to a machine learning (ML) algorithm for estimating fitness of enzyme variants, the machine learning (ML) algorithm being trained, and/or being trainable, on data obtained by a method according to the first aspect of the invention.

30

A further aspect of the present invention relates to use of a machine learning (ML) algorithm trained on data comprising a correlation of a specific nucleotide sequence encoding an enzyme variant to frequency information for optimizing a

35

nucleotide sequence encoding an enzyme variant, or to guide the generation of an improved nucleotide sequence encoding an enzyme variant with given fitness properties.

- 5 Another aspect of the invention relates to use of a machine learning (ML) algorithm trained on data obtained by a method according to the first aspect of the invention for optimizing a nucleotide sequence encoding an enzyme variant, or to guide the generation of an improved nucleotide sequence encoding an enzyme variant with given fitness properties.

10

In yet another aspect, the invention relates to use of a machine learning (ML) algorithm trained on data obtained by a method according to the first aspect of the invention to predict fitness for any enzyme sequence, or to guide the generation of enzyme sequences with given fitness properties.

15

An aspect of the invention relates to a system suitable for executing an algorithm, such as machine learning (ML) algorithm, for optimizing a nucleotide sequence encoding an enzyme variant, the system being trained, and/or being trainable, on data obtained by a method according to the first aspect of the invention, the

20 system being arranged for:

- receiving a first set of information (1SI), such as a first database, said first of information being obtained, or being obtainable, from d) subjecting the isolated DNA variant library grown under condition I.) to long-read sequencing, thereby
- 25 obtaining said first set of information, such as a first database, linking a specific DNA barcode to a specific DNA variant coding for a specific enzyme variant,
- receiving a second set of information (2SI), such as a second database, said second of information being obtained, or being obtainable, from e) subjecting the
- 30 isolated DNA variant library grown under condition II.) to short-read deep sequencing, thereby obtaining said second set of information, such as a second database, linking a specific DNA barcode to a frequency of the specific DNA variant coding for a specific enzyme variant or a rate of change of the specific DNA variant coding for a specific enzyme variant;

35

wherein the system comprises:

- a first comparison module (1CM) for comparing said first set of information (1SI) to said second set of information (2SI), thereby correlating specific enzyme variants to the frequency information, such as frequency of the variant, using the
5 DNA barcode as an identifier; and
- an optional second comparison module (2CM) for correlating determined frequency information to the enzyme fitness of the specific enzyme variant.

Yet another aspect of the invention relates to a method for training a machine
10 learning (ML) system for optimizing a nucleotide sequence encoding an enzyme variant according to the invention, the method comprises the steps of:

- receiving training data comprising a first set of information (1SI), such as a first database, and a second set of information (2SI), such as a second database,
- training the ML system for optimizing a nucleotide sequence encoding an enzyme
15 variant using said training data, and
- validating the ML system using correlated specific enzyme variants to the frequency information, such as frequency of the variant.

A further aspect relates to a computer-readable storage medium comprising data
20 obtained or obtainable by a method according to the first aspect.

An aspect of the invention relates to neural network model obtained or obtainable by a method for training a machine learning (ML) system according to the invention.
25

Brief description of the figures

Figure 1 shows a workflow of the library generation, barcoding, and *in vivo* library preparation,

30 Figure 2 shows a sequence-fitness data collection workflow,

Figure 3 shows distribution of relative enzyme fitness for the aggregated dataset,

Figure 4 shows the distribution of growth rates for **A)** variants with premature stop codons using the aggregated dataset, based on variants with at least 7 reads and **B)** variants with premature stop codons using the aggregated dataset, based on variants with at least 15 reads. C) illustrates comparable levels of dispersion for the WT2,

Figure 5 shows growth assay results. A strong positive correlation was observed between the estimated fitness in dataset and the measured fitness in the growth assay, with R^2 value of 0.87, and

10

Figure 6 shows the bioconversion results. **A)** vanillic acid produced in bioconversion experiment by all cloned ASMT variants. The dataset includes ASMT variants identified from the selected population, and novel ASMT variants generated using first- and second-order models. ASMT-WT refers to the original wild-type ASMT sequence. ASMT1-ASMT37 refer to ASMT variants from the dataset. The rest of the ASMT sequences were generated with the second-order model. **B)** shows vanillic acid production plotted against predicted growth rate for all cloned ASMT variants identified during selection. Error bars show one standard deviation. 95% confidence bands for the best-fit line are depicted with dashed lines. The variance of the data is considered when calculating the R^2 value. Significance of deviation of regression slope from zero: $p < 0.001$ ($\alpha = 0.05$) (Graphpad Prism). **C)** shows vanillic acid production plotted against predicted growth rate for functional, cloned ASMT variants identified during selection with detectable levels of vanillic acid in bioconversion experiment. Error bars show one standard deviation. 95% confidence bands for the best-fit line are depicted with dashed lines. The variance of the data is considered when calculating the R^2 value. Significance of deviation of regression slope from zero: $p < 0.001$ ($\alpha = 0.05$) (Graphpad Prism). **D)** shows vanillic acid production plotted against predicted growth rate for functional, cloned ASMT variants identified during selection with detectable levels of vanillic acid in bioconversion experiment. Only data points where $SD < 0.5$ for vanillic acid production and $SD < 0.0152$ for predicted growth rate are included. Error bars show one standard deviation. 95% confidence bands for the best-fit line are depicted with dashed lines. The variance of the data is considered when calculating the R^2 value. Significance of deviation of regression slope from zero: $p = 0.003$ ($\alpha = 0.05$) (Graphpad Prism).

The present invention will now be described in more detail in the following.

Detailed description of the invention

Definitions

- 5 Prior to discussing the present invention in further details, the following terms and conventions will first be defined:

Long-read sequencing

- Long-read sequencing is a group of DNA/RNA sequencing techniques that enables the sequencing of much longer DNA fragments, such as having more than 601
10 bases to more than 20,000 bases, than traditional short-read deep sequencing methods. Long-read sequencing tends to have lower accuracy and lower read depth than short-read deep sequencing.

- Long-read sequencing technologies may be selected from the group consisting of Pacific Biosciences' (PacBio) single-molecule real-time (SMRT) sequencing and
15 Oxford Nanopore Technologies' (ONT) nanopore sequencing, Illumina Complete Long Read sequencing, TELL-Seq.

Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) are also known as third generation sequencing technologies.

Short-read sequencing

- 20 Short-read sequencing is sequencing of shorter fragments of DNA or RNA. In these types of sequencing, DNA or RNA is broken into smaller fragments, before being sequenced. With short read sequencing, DNA or RNA can be sequenced with high read number, and with length limited to 50-300 per direction.

- Short-read sequencing is a form of next-generation sequencing (NGS), wherein
25 the platform for short-read sequencing may be selected from the group consisting of Illumina, such as HiSeq2000, HiSeq2500, HiSeq4000, HiSeqX10, and NovaSeq6000, and MGI, such as BGISEQ-500, and MGISEQ-T7.

Short-read sequencing is also known as deep sequencing; hence the terms may be used interchangeably.

- 30 Enzyme

Enzymes are biological molecules (such as proteins and RNAs) that catalyze chemical reactions. The molecules that enzymes act upon are called substrates, and the molecules that enzymes convert the substrates into are called products.

DNA variant library

A DNA variant library is a pool of DNA variants encoding enzyme variants that are expected to have different sequences with different fitness levels. The DNA variants are obtained through various methods, such as error-prone PCR based on
5 given templates, or assembly of DNA fragments that differ from each other in specific region or regions.

Enzyme variants

Enzyme variants are the protein product from translation of the DNA variants. The enzyme variants are expected to have different fitness levels or enzyme activity
10 due to their different sequences.

Enzyme fitness

Enzyme fitness is a measure of how well a given enzyme performs a specific function. It can be used interchangeably with *enzyme activity* in the context of this invention. It encompasses properties such as expression level or ease of
15 expression in a given host cell, turnover number (unimolecular rate constant, K_{cat}), Michaelis constant (K_M), solubility, and stability, substrate specificity, product specificity, thermotolerance.

Thus, in an embodiment, the enzyme fitness is selected from the group consisting
20 of enzyme activity, expression level, turnover number such as unimolecular rate constant (K_{cat}), Michaelis constant (K_M), enzyme solubility, enzyme stability, substrate specificity, product specificity, thermotolerance, and combinations thereof.

Rate of change

25 The rate of change is to be understood as the change of abundance or frequency of the DNA/enzyme variants, or the change of abundance or frequency of the cells that express specific enzyme variants.

DNA barcode

The DNA barcode is (a) DNA fragment(s) that is attached to a DNA variant, often
30 randomly. The DNA barcode is linked to the DNA variants as an extension of the DNA sequence of the DNA variant via phosphodiester bonds. As an example, a one-cycle PCR step can be performed on a DNA variant to attach a unique barcode to the DNA variant.

DNA barcodes serve three purposes: 1) for generating accurate full-length DNA sequence reading from long-read sequencing based on consensus generation for the DNA reads with the same barcodes; 2) for profiling the frequency change of DNA variants during growth-based selection using short-read sequencing; and 3) 5 for associating the frequency information obtained using short-read sequencing to the full-length DNA variant sequence information obtained using long-read sequencing.

Selection pressure

A selection strain comprising a DNA variant library encoding enzyme variants of a 10 specific template enzyme may be subjected to a selection pressure biased toward higher target enzyme activity. The choice of specific selection pressure depends on the specific type of target enzyme. As an example, a selection strain can be constructed by deleting endogenous gene(s) responsible for synthesizing cysteine in the host cells, while introducing an alternative pathway for synthesizing 15 cysteine, part of which is the target enzyme to be selected. In this case, in a certain range, higher target enzyme activity will lead to improved growth of the selection strain in a given media where the introduced alternative pathway could be used.

The selection strain may also be subjected to selection pressure with antibiotics, 20 such as chloramphenicol, ampicillin, and the like, which does not bias toward higher target enzyme activity.

Genetic modification

In this invention, genetic modification specifically refers to changes introduced to the genome of the host cells besides the introduction of the DNA variants. Genetic 25 modifications serve the purpose of rendering the host cells susceptible to a certain selection pressure. As one example, genetic modifications may be used to delete the endogenous cysteine synthesis pathway and introduce a heterologous cysteine synthesis pathway such that there is selection pressure for higher activity for the enzymes in the heterologous pathway. As another example, genetic modifications 30 may be used to introduce a sensor that could sense the concentration of the product of a target enzyme, and control the expression of an essential gene, such that higher target enzyme activity leads to higher expression level of the essential gene.

Growth rate

The growth rate of the selection strain is determined by the change in the number of cells over a specific period of time. Under selection pressure, the growth rate of a cell variant of the selection strain signifies the activity or fitness of a specific enzyme variant that the cell variant contains. Hence, a high growth rate correlates with a high enzyme activity or enzyme fitness.

Optical density (OD)

Optical density (OD) is a measure of how much light is absorbed by a material as it passes through it. It is defined as the logarithmic ratio of the intensity of incident light (I_o) to the intensity of transmitted light (I_t) through the material. In the context of cell cultures, optical density OD is commonly used to estimate the concentration or density of cells in a culture. This is typically done using a spectrophotometer, which measures how much light is scattered by the cells in the sample. When light passes through a cell culture, cells scatter the light. The spectrophotometer measures this scattering, which is reported as optical density. The more cells present, the higher the OD value.

MoCHI

MoCHI (Faure, Andre & Lehner, Ben. (2024). *MoCHI: neural networks to fit interpretable models and quantify energies, energetic couplings, epistasis and allostery from deep mutational scanning data*) is a software tool that allows the parameterisation of arbitrarily complex models using deep mutational scanning (DMS) data. MoCHI simplifies the task of building custom models from measurements of mutant effects on any number of phenotypes. It allows the inference of free energy changes, as well as pairwise and higher-order interaction terms (energetic couplings) for specified biophysical models. When a suitable user-specified mechanistic model is not available, global nonlinearities (epistasis) can be estimated directly from the data. MoCHI also builds upon and leverages theory on ensemble (or background-averaged) epistasis to learn sparse predictive models that can incorporate higher-order epistatic terms and are informative of the genetic architecture of the underlying biological system. The combination of DMS and MoCHI allows biophysical measurements to be performed at scale, including the construction of complete allosteric maps of proteins.

Storage medium

A storage medium is to be understood as a physical material or device used to store digital data. It allows data to be saved, retrieved, and managed. A storage medium may be, but not limited to:

- 5 - a magnetic storage, such as Hard Disk Drive (HDD), magnetic tape, floppy disk, zip disk, super disk,
- flash memory, such as USB flash drives, SD cards, Solid State Drive (SSD),
- optical disc, or
- network storage, such as Cloud storage, Network-Attached Storage (NAS).

10

The invention presented is in the framework of directed evolution, meaning the screening and selection of enzyme variant libraries based on the metrics of the enzymes. The aim is not to simply find the best-performing variant, but also the fitness information of each enzyme variant in the libraries, in order to obtain an

15 information-rich dataset useful for machine learning. It combines four essential parts:

- 1) **Coupling cell growth to the catalytic activity of an enzyme.** This enables the use of cell growth rate as a reflection of enzyme activity/fitness.

20

- 2) **Attaching unique DNA barcodes to the DNA sequence of each enzyme variant.** A crucial step in the library creation procedure is the attachment of one or more unique DNA barcodes to the DNA encoding each enzyme variant.

25

- 3) **Using long-read sequencing technologies (e.g., Oxford Nanopore, PacBio) to sequence the variant library.** This allows the creation of a DNA variant sequence-barcode database that shows, which barcode corresponds to which specific DNA variant encoding a specific enzyme
- 30 variant.

The DNA barcode additionally serves the purpose of correcting the relatively high error rate of long-read sequencing: by reading multiple copies of a DNA variant-barcode fragment, different sequencing reads containing the same DNA barcode sequence will be assumed to have the

35 same sequence, and the alignment of these different reads containing

sequencing errors will generate a consensus sequence with high sequence accuracy.

- 5 4) **Using deep sequencing technologies (e.g., Illumina sequencing) to profile the frequency change of barcodes during the library selection process.** The frequency, or the change rate of the frequency or abundance of each barcode, will be used as a proxy for the fitness and activity of the corresponding enzyme. Through the DNA sequence-barcode database generated through step 3), the identity of the corresponding
10 enzyme sequence will be obtained for each DNA barcode. This way, the activity for each enzyme variant is obtained.

With this workflow, the enzyme sequence-activity dataset will have important desirable features:

- 15 A. A full-length sequence, which allows characterization of the effects of the interactions of different mutations at different positions of the enzyme sequence, termed epistasis;
 B. High sequence accuracy through barcode-based consensus generation;
 C. Quantitative activity labelling of each individual enzyme variant;
20 D. Given the available technologies, a capacity of $10^5 - 10^7$ data points during each run.

The combination of these features is highly valuable for machine learning-assisted directed evolution.

25 **A method for estimating fitness of enzyme variants**

One aspect of the present invention relates to a method for estimating fitness of enzyme variants, the method comprising

- 30 a) providing a DNA variant library, preferably a plasmid library, the library comprises DNA variants coding for enzyme variants, wherein each DNA variant is linked to one or more unique DNA barcodes, preferably through a one-cycle PCR process;
 b) transforming the DNA variant library into a host cell for expressing the enzyme variants from the library;

- I. growing the host cell comprising the DNA variant library, preferably under conditions in the absence of a selection pressure favoring enzyme variants with increased activity; and
- II. growing the host cell comprising the DNA variant library, under
5 selection pressure favoring enzyme variants with increased catalytic activity;
- c) isolating DNA variant library from the host cells grown under condition I.) and isolating at least the barcode region of the DNA variant library from the host cells grown under condition II.);
- 10 d) subjecting the isolated DNA variant library grown under condition I.) to long-read sequencing, thereby obtaining first set of information, such as a first database, linking a specific DNA barcode to a specific DNA variant coding for a specific enzyme variant,
- e) subjecting the isolated DNA variant library grown under condition II.) to
15 short-read deep sequencing, thereby obtaining a second set of information, such as a second database, linking a specific DNA barcode to a frequency of the specific DNA variant coding for a specific enzyme variant or a rate of change of the specific DNA variant coding for a specific enzyme variant;
- f) comparing the first set of information from step d) to the second set of
20 information from step e), thereby correlating specific enzyme variants to the frequency information, such as the frequency of the variant (such as in the entire population of enzyme variants); and
- g) optionally estimating the enzyme fitness of the specific enzyme variant from the determined frequency information.

25

Another aspect of the invention relates to a method for optimizing a nucleotide sequence encoding an enzyme variant, the method comprising

- a) providing a DNA variant library, preferably a plasmid library, the library comprises DNA variants coding for enzyme variants, wherein each DNA
30 variant is linked to one or more unique DNA barcodes, preferably through a one-cycle PCR process;
- b) transforming the DNA variant library into a host cell for expressing the enzyme variants from the library;

- I. growing the host cell comprising the DNA variant library, preferably under conditions in the absence of a selection pressure favoring enzyme variants with increased activity; and
- II. growing the host cell comprising the DNA variant library, under selection pressure favoring enzyme variants with increased catalytic activity;
- c) isolating DNA variant library from the host cells grown under condition I.) and isolating at least the barcode region of the DNA variant library from the host cells grown under condition II.);
- d) subjecting the isolated DNA variant library grown under condition I.) to long-read sequencing, thereby obtaining first set of information, such as a first database, linking a specific DNA barcode to a specific DNA variant coding for a specific enzyme variant,
- e) subjecting the isolated DNA variant library grown under condition II.) to short-read deep sequencing, thereby obtaining a second set of information, such as a second database, linking a specific DNA barcode to a frequency of the specific DNA variant coding for a specific enzyme variant or a rate of change of the specific DNA variant coding for a specific enzyme variant;
- f) comparing the first set of information from step d) to the second set of information from step e), thereby correlating specific enzyme variants to the frequency information, such as the frequency of the variant; and
- g) obtaining the optimized nucleotide sequence encoding an enzyme variant from the determined frequency information.
- In an embodiment, the method further comprises a step between step d) and e) of subjecting the isolated DNA variant library grown under condition II.) to long-read sequencing.
- The long-read sequencing may be performed of post-selection library after selection, in addition to the long-read sequencing of pre-selection library. When sequencing coverage is not sufficient compared to the library size, the long-read sequencing of post-selection library enriches the reads for those with highest activities, thereby enriching the data with high activity, which contains more information than those with less activities.

As shown in the Examples of the present invention, studies have been conducted, documenting that the method of the invention indeed works as claimed.

Step A

- 5 In an embodiment of the invention, the identification of the enzyme variant fitness is for determining the fitness of a portion of enzyme variants in the DNA variant library, preferably a portion of the variants with the highest activity.

In an embodiment, the identification of the optimized nucleotide sequence
10 encoding an enzyme variant comprises determining the fitness of a portion of enzyme variants in the DNA variant library, preferably a portion of the variants with the highest activity.

As seen throughout the Examples, the optimized or improved enzyme variants are selected based on the fitness of the enzyme variants, preferably variants with the
15 highest fitness/activity. Example 12 shows validation of selected enzyme variants with predicted improved fitness and activity.

In an embodiment of the invention, the enzyme fitness is selected from the group consisting of enzyme activity, expression level, turnover number such as
20 unimolecular rate constant (K_{cat}), Michaelis constant (K_M), enzyme solubility, enzyme stability, substrate specificity, product specificity, thermotolerance, and combinations thereof.

In an embodiment of the invention, the DNA variant library is prepared by the
25 steps:

- a) providing at least one DNA sequence encoding an enzyme variant;
- b) linking one or more unique DNA barcodes on the at least one DNA sequence encoding an enzyme variant and obtaining a barcode-DNA library; and
- 30 c) assembling the barcode-DNA library and obtaining a DNA variant library.

As seen in Example 2, the library generation and barcoding are performed successfully as outlined above. One DNA sequence may be linked to more than one DNA barcode.

In an embodiment of the invention, the at least one DNA sequence of step a) is generated by PCR, preferably error-prone PCR based on given templates, or assembly of DNA fragments that differ from each other in specific region or regions.

5 As seen in Example 2, the DNA library is generated via PCR successfully. However, for the present invention, the method for generating the DNA library, by introducing nucleotide mutational changes in a template DNA sequence encoding a template enzyme, is not limited to PCR.

10 In an embodiment of the invention, the assembly of the barcode-DNA library is by Golden Gate assembly.

As seen in Example 2, the DNA library is assembled into a vector by Golden Gate assembly, however for the present invention, the method for assembling the DNA library into a vector is not limited to Golden Gate assembly, hence the DNA library

15 may be assembled by various DNA assembly of PCR methods.

In an embodiment of the invention, the unique DNA barcode is a random nucleotide sequence.

DNA barcodes serve three purposes: 1) for generating accurate full-length DNA
20 sequence reading from long-read sequencing based on consensus generation for the DNA reads with the same barcodes; 2) for profiling the frequency change of DNA variants during growth-based selection using short-read sequencing; and 3) for associating the frequency information obtained using short-read sequencing to the full-length DNA variant sequence information obtained using long-read
25 sequencing.

In an embodiment of the invention, the unique DNA barcode has a length in the range 10-300 DNA nucleotides, preferably 15-100 DNA nucleotides, most preferably 20-30 DNA nucleotides.

30 The length of the unique DNA barcode should have a sufficient length for the purpose of identifying the associated DNA sequence of an enzyme variant as well as determining the frequency change of the barcode of the enzyme variants.

In an embodiment of the invention, the DNA variant is linked to the unique DNA barcode via a covalent bond, such as an extension of the DNA variant coding for the enzyme variant, such as a phosphodiester bond.

The unique DNA barcode may be linked to the DNA variant as an extension of the
5 DNA variant, hence the DNA barcode may be attached to the DNA variant via a phosphodiester bond. The DNA variant may have more than one DNA barcode linked.

In an embodiment of the invention, the DNA library is a vector library, such as a
10 plasmid library.

The DNA library is assembled into a vector for obtaining a vector library, which is expressed in a host cell. As seen in Example 2, the barcoded DNA library is assembled onto a cloning vector for expression in *E. coli*. Alternatively, the DNA variants can be assembled with homology arms for integration onto the genome
15 of the host cell. The type of vector is not limited to a specific vector and the general way of assembling the DNA variants is not limited to a specific way for the present invention.

In an embodiment, the DNA variant library comprises at least 100.000 DNA
20 variants, such as at least 300.000 DNA variants, preferably at least 800.000 DNA variants, more preferably at least 1 million DNA variants.

In an embodiment, the DNA variant library comprises 100.000-5 million DNA variants, such as 300.000-5 million DNA variants, such as 500.000-4 million DNA
25 variants, preferably 800.000-4 million DNA variants, more preferably 1-3 million DNA variants, most preferably 3 million DNA variants.

As seen in Example 9, the DNA variant library comprises the above mentioned size.

30 In an embodiment, the enzyme is an enzyme for which the enzyme activity and/or enzyme fitness correlates with growth of the host cell, such as growth during selection pressure.

The enzyme for the determination of the fitness of the enzyme or for optimizing a nucleotide sequence encoding an enzyme variant according to the present
35 invention may be any enzyme for which the enzyme activity and enzyme fitness

correlates with growth of the host cell during growth-based selection pressure favoring enzyme variants with increased activity.

In an embodiment of the invention, the enzyme is selected from the group
5 consisting of oxidoreductases, transferases, hydrolase, lyases, isomerases, ligases, translocases, or a plurality thereof.

As shown in the Examples, the enzyme may be SAM-dependent methyltransferase.

In an embodiment, the enzyme is N-acetylserotonin O-methyltransferase (ASMT).
10 As seen in the Examples, the enzyme is ASMT.

Step B

In an embodiment of the invention, the host cell is a strain selected from the group consisting of a bacterial cell, a yeast cell, and a eukaryotic cell.

The host cell may be any cell, which can express the DNA library and be subjected
15 to selection pressure. The appropriate host cell may be chosen based on the specific enzyme to be expressed. For the present invention, the host cell is not limited to a specific type.

In an embodiment of the invention, the host cell is a bacterial cell, preferably *E.*
20 *coli*.

As shown in Example 2, the host cell may be *E. coli*, especially for the expression of SAM-dependent methyltransferase as presented in the Examples.

In an embodiment of the invention, the selection pressure dependency is achieved
25 by genetic modification of the host genome, represented preferably by the following table:

Selection pressure	Rationale	Example
Sensor-based approach	The activity of the target enzyme leads to increased concentration of a molecule, which is sensed by another	Andon, James S., ByungUk Lee, and Tina Wang. "Enzyme directed evolution using genetically encodable

	molecule (for example, a transcriptional regulator) that controls the concentration of a third molecule (for example, product of an essential gene)	biosensors." Organic & Biomolecular Chemistry 20.30 (2022): 5891-5906.
Metabolic coupling	The activity of the target enzyme leads to increased concentration of a molecule which is a building block for cell growth.	Luo, Hao, et al. "Coupling S-adenosylmethionine-dependent methylation to growth: Design and uses." Plos biology 17.3 (2019): e2007050.
Substrate stress relief	The activity of the target enzyme leads to decreased concentration of a molecule which inhibits cell growth.	Zhan, Chunjun, et al. "Improved polyketide production in C. glutamicum by preventing propionate-induced growth inhibition." Nature metabolism 5.7 (2023): 1127-1140.

The selection pressure is dependent on the enzyme’s catalytic activity for converting substrates to products. Depending on the enzyme to be expressed in the host cell, the selection pressure is chosen accordingly to favor the enzyme

5 variants with increased activity.

In an embodiment, the host cell is grown under selection pressure at an optical density (OD) of 0.01-0.5, such as 0.3-0.5, preferably 0.05-0.5.

As seen in Example 9, the host cells are grown at the mentioned optical density.

In a further embodiment, the host cell is grown in a colorless media. In the non-selective and selective media, coloration of the media may be avoided, since colour of the media can result in interference with OD reading.

5

Step D

In an embodiment of the invention, the long-read sequencing is sequencing of at least 601 nucleotides, such as at least 1000 nucleotides, such as at least 5000 nucleotides, such as at least 10,000 nucleotides, such as at least 20,000

10 nucleotides.

In an embodiment of the invention, the long-read sequencing is selected from the group consisting of third generation sequencing, such as Oxford Nanopore sequencing, PacBio sequencing, Illumina Complete Long Read sequencing, TELL-

15 Seq.

Long-read sequencing enables sequencing of the DNA barcode and the DNA variant encoding an enzyme variant, hence the whole barcode-DNA variant is sequenced for obtaining the sequence of the DNA barcode for identifying the associated enzyme variant.

20

Step E

In an embodiment of the invention, the short-read sequencing is sequencing of nucleotide sequences in the range 10-600 nucleotides, preferably in the range 30-500, most preferably in the range 50-300 nucleotides.

25

In an embodiment of the invention, the short-read sequencing is selected from the group consisting of deep sequencing, Next-Generation Sequencing (NGS), such as Illumina sequencing.

Short-read sequencing is a high-accuracy sequencing method, and for the present invention, short-read sequencing is performed on the barcode region for obtaining information relating to the abundance or frequency of the barcode present during selection, and optionally at different time points. Hence, the enzyme activity can be determined based on the number of barcodes for each enzyme variant.

Further, the rate of change of abundance of DNA variants can be determined from sampling host cells expressing the DNA library under selection pressure at

35

different time points and subsequently isolating the DNA for short-read sequencing (See Example 5).

In an embodiment of the invention, the frequency information of the specific DNA
5 variant coding for a specific enzyme variant is a reflection of the enzyme fitness of said specific enzyme variant.

The frequency information from short-read sequencing of the DNA variants, or the rate of change of the variants that can be calculated based on the frequency information, is used as a proxy for the fitness of the corresponding enzyme. The
10 enzyme fitness includes expression level or ease of expression in a given host cell, turnover number (unimolecular rate constant, K_{cat}), Michaelis constant (K_M), solubility, and stability, substrate specificity, product specificity, thermotolerance.

In an embodiment of the invention, the rate of change of the specific DNA variant
15 coding for a specific enzyme variant is determined from the change of the amount of the specific DNA variant over a specific period of time. The rate of change can be obtained based on the barcode frequency information and total cell abundance over time.

Sampling of the host cells under selection pressure, and subsequent short-read
20 sequencing of the barcode region to characterize the change of variant abundance/frequency, will allow the calculation of the growth rate of the cell population harboring a specific enzyme variant. This growth rate will be used as a proxy for the activity of the cognate enzyme variant in the cell population.

25 Step F

In an embodiment of the invention, in step f) the comparison of the first set of information from step d) to the second set of information from step e), is by correlating DNA barcodes from the second set of information to the enzyme variants of the first set of information via the enzyme variant specific DNA
30 barcode, thereby obtaining a prediction of enzyme fitness of the enzyme variants and/or an optimized nucleotide sequence encoding an enzyme variant via the frequency of the DNA barcodes from the second set of information or the rate of change of the specific DNA variant coding for a specific enzyme variant.

Via the first set of information, the identity of the cognate enzyme variant is
35 known for each barcode. Hence, the second set of information, which comprises

the sequence of the barcodes and frequency of barcodes only, is compared to the first set of information, wherein each barcode of the second set of information is matched with each barcode of the first set of information, and thereby determining the corresponding enzyme variants for the barcodes of the second set
5 of information. As the frequency of each barcode is known from the second set of information, the frequency of corresponding enzymes variants is obtained.

Step G

In an embodiment of the invention, the enzyme fitness of the specific enzyme
10 variant and/or an optimized nucleotide sequence encoding an enzyme variant is reflected by a growth rate of the host cell.

Growth rate is obtained by characterizing the change of variant abundance/frequency, which will allow the calculation of the growth rate of the cell population harboring a specific enzyme variant. This growth rate will be used
15 as a proxy for the activity of the cognate enzyme variant in the cell population.

In an embodiment of the invention, the growth rate is determined from growing the host cell from step b(I.) under selection pressure favoring enzyme variants with increased fitness or activity.

20 As the selection pressure favors enzyme variants with increased fitness or activity, it is to be expected that host cells expressing the enzyme variants with highest fitness or activity will exhibit a higher growth rate. As mentioned above, the growth rate can be determined based on the rate of change of variant abundance.

25 In an embodiment, the DNA nucleotide sequence coding for enzyme variants comprises a promoter such as J23100 (SEQ ID NO. 15) or J23110 (SEQ ID NO. 16).

As seen in Examples 3 and 9, the promoters used are J23100 or J23110.

Use of barcoded DNA sequence for predicting enzyme fitness

30 A second aspect of the present invention relates to use of at least one barcoded DNA sequence encoding an enzyme variant for predicting enzyme fitness of said enzyme variant and/or for optimizing said DNA sequence encoding said enzyme variant, wherein the at least one barcoded DNA sequence is subjected to long-read sequencing and short-read sequencing.

As can be seen in the Examples, the DNA library encoding enzyme variants is barcoded for later long-read and short-read sequencing. The long-read sequencing of the barcoded DNA results in a first set of information or first database, which is the association of the DNA barcode sequence and its
5 corresponding enzyme variant. The short-read sequencing of the DNA barcode allows for quantification of the DNA barcode at specific time points and thereby obtaining a second set of information comprising the frequency of each barcode or change rate of variant abundance. The comparison of the first set of information and second set of information allows for assessing the activity or fitness of each
10 enzyme variant. Thus, the assessment of activity or fitness of each enzyme variant allows for optimizing DNA sequences encoding enzyme variants.

In an embodiment of the invention, the at least one barcoded DNA sequence for short-read sequencing has been obtained from a host cell subjected to selection
15 pressure favoring enzyme variants with increased fitness.

The host cells being subjected to a selection pressure favoring enzyme variant with increased fitness or activity will exhibit a higher growth rate when expressing enzyme variants with increased activity. When sampling the cells for isolation of the DNA for short-read sequencing, the abundance of each of the enzyme variant
20 can be determined.

Polypeptide

As shown in Example 8, 11 and 12, the inventing team has identified a number of optimized variants of atASMT having a high fitness score for the conversion of PCA to vanillic acid in *E. coli*.

25 Thus, in an aspect, the invention relates to a polypeptide according to SEQ ID NO. 3 comprising one or more of the following mutations P72S, K132R, D147E, I103N, P286L, R77C, G105S, K187R, M215K, P231S, E46D, S84I, and/or P302L.

In an embodiment, the polypeptide comprises one or more of the following combinations of mutations:

- 30
- P72S, K132R, and D147E,
 - P72S, I103N, and P286L,
 - R77C, G105S, and K187R,

- R77C, M215K, and P231S, and/or
- E46D, S84I, and P302L.

In yet an embodiment, the polypeptide is a polypeptide according to any of SEQ ID NOs 10-14.

5 In another embodiment, the polypeptide is an isolated polypeptide.

Nucleic acids, vectors and host cells

An aspect of the invention relates to nucleic acid molecules encoding the polypeptides according to the invention.

Yet an aspect relates to a vector, such as a plasmid comprising the nucleic acid
10 molecule according to the invention.

Yet another aspect relates to a host cell comprising the vector according to the invention.

Software/AI

A further aspect of the present invention relates to a processor system
15 programmed to operate according to a machine learning (ML) algorithm for estimating fitness of enzyme variants and/or for optimizing a nucleotide sequence encoding an enzyme variant, the machine learning (ML) algorithm being trained, and/or being trainable on data comprising a correlation of a specific nucleotide sequence encoding an enzyme variant to frequency information.

20

Another aspect of the present invention relates to a processor system programmed to operate according to a machine learning (ML) algorithm for estimating fitness of enzyme variants and/or for optimizing a nucleotide sequence encoding an enzyme variant, the machine learning (ML) algorithm being trained,
25 and/or being trainable, on data obtained by a method according to the first or second aspect of the invention.

Yet a further aspect of the invention relates to use of a machine learning (ML) algorithm trained on data comprising a correlation of a specific nucleotide
30 sequence encoding an enzyme variant to frequency information to predict fitness

for any enzyme sequence, or to guide the generation of enzyme sequences with given fitness properties.

In yet another aspect, the invention relates to use of a machine learning (ML)
5 algorithm trained on data obtained by a method according to the first aspect of the invention to predict fitness for any enzyme sequence, or to guide the generation of enzyme sequences with given fitness properties.

Another aspect of the invention relates to use of a machine learning (ML)
10 algorithm trained on data comprising a correlation of a specific nucleotide sequence encoding an enzyme variant to frequency information for optimizing a nucleotide sequence encoding an enzyme variant, or to guide the generation of an improved nucleotide sequence encoding an enzyme variant with given fitness properties.

15 A further aspect of the present invention relates to use of a machine learning (ML) algorithm trained on data obtained by a method according to the second aspect for optimizing a nucleotide sequence encoding an enzyme variant, or to guide the generation of an improved nucleotide sequence encoding an enzyme variant with
20 given fitness properties.

An aspect of the present invention relates to a system suitable for executing an algorithm (such as machine learning (ML) algorithm) for estimating fitness of enzyme variants and/or for optimizing a nucleotide sequence encoding an enzyme
25 variant, the (machine learning) system being trained, and/or being trainable, on data provided according to the first or second aspect, the system being arranged for:

- receiving a first set of information (1SI), such as a first database, said first of
30 information being obtained, or being obtainable, from d) subjecting the isolated DNA variant library grown under condition I.) to long-read sequencing, thereby obtaining said first set of information, such as a first database, linking a specific DNA barcode to a specific DNA variant coding for a specific enzyme variant,

- receiving a second set of information (2SI), such as a second database, said second of information being obtained, or being obtainable, from e) subjecting the isolated DNA variant library grown under condition II.) to short-read deep sequencing, thereby obtaining said second set of information, such as a second
5 database, linking a specific DNA barcode to a frequency of the specific DNA variant coding for a specific enzyme variant or a rate of change of the specific DNA variant coding for a specific enzyme variant;

wherein the system comprises:

- 10 -a first comparison module (1CM) for comparing said first set of information (1SI) to said second set of information (2SI), thereby correlating specific enzyme variants to the frequency information, such as frequency of the variant; and
- an optional second comparison module (2CM) for correlating determined frequency information to the enzyme fitness of the specific enzyme variant.

15

In an embodiment, the algorithm is MoCHI. In Example 11, a series of variant enzyme sequences were generated using the MoCHI. However, in principle any machine learning-based models of protein fitness can be used.

- 20 Advantageously, the invention may also relate to a method for training a machine learning (ML) system for estimating fitness of enzyme variants and/or for optimizing a nucleotide sequence encoding an enzyme variant, such as according to the previous aspects. Thus, yet an aspect of the invention relates to a method comprises the steps of:

- 25 -receiving training data comprising a first set of information (1SI), such as a first database, and a second set of information (2SI), such as a second database,
-training the system for estimating fitness of enzyme variants using said training data, and
- validating the system using correlated specific enzyme variants to the frequency
30 information, such as frequency of the variant.

In an embodiment, the ML system comprises:

- a ML model configured to assign a score to a nucleotide sequence encoding an enzyme variant,

- a comparison component configured to compare the score with a threshold for predicting a growth rate of a host cell expressing said nucleotide sequence encoding the enzyme variant, and
- a featuring component configured to receive an output from the
5 comparison component, wherein the output indicates a predicted fitness, such as activity, of said enzyme variant encoded by said nucleotide sequence.

In an embodiment, if the score is above the threshold, it is indicative of an
10 improved growth rate of a host cell expressing said enzyme variant compared to a growth rate of a host cell expressing a corresponding wild-type enzyme.

In yet an embodiment, if the score is below the threshold, it is indicative of a
decreased growth rate of a host cell expressing said enzyme variant compared to
15 a growth rate of a host cell expressing a corresponding wild-type enzyme.

In a preferred embodiment, the threshold is 0.

Preferably the system and/or algorithm and/or method is implemented on a
20 computer, thus being computer-implemented.

A further aspect relates to a computer-readable storage medium comprising data comprising a correlation of a specific nucleotide sequence encoding an enzyme variant to enzyme fitness, such as enzyme activity.

25

Another aspect relates to a computer-readable storage medium comprising data obtained by a method according to the invention.

An aspect of the invention relates to a neural network model obtained or
30 obtainable by the method for training a machine learning (ML) system according to a previous aspect.

Below is given a list of some no-limiting types of algorithms that are particularly suited for machine learning (ML) system and/or training of a (ML) system using
35 enzyme data:

1. Deep Learning Algorithms: Deep learning is a subset of machine learning where artificial neural networks, algorithms inspired by the human brain, learn from large amounts of data. Deep learning algorithms are capable of learning to
5 represent the world as a nested hierarchy of concepts, with each concept defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones. The skilled reader is referred to for example University of Illinois at Urbana-Champaign; "AI predicts enzyme function better than leading tools." ScienceDaily. ScienceDaily, 30 March 2023.
10 <www.sciencedaily.com/releases/2023/03/230330172121.htm>.

2. Contrastive Learning: This is a type of unsupervised learning approach that trains models to learn similar features from similar data points and different features from different data points. An AI tool named 'CLEAN' was recently
15 reported to use this algorithm to predict enzyme function, cf. Gupta, R., Srivastava, D., Sahu, M. et al. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol Divers 25, 1315–1360 (2021) for more details.

20 3. Artificial Neural Networks (ANNs): ANNs are computing systems vaguely inspired by the biological neural networks that constitute animal brains. An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain.

25 4. Support Vector Machines (SVMs): SVMs are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis.

5. Generative Adversarial Networks (GANs): GANs are a class of artificial
30 intelligence algorithms used in unsupervised machine learning, implemented by a system of two neural networks contesting with each other in a zero-sum game framework.

These algorithms can be used individually or in combination, depending on the
35 specific requirements of the type of enzyme data. The invention according to this

aspect can be implemented by means of hardware, software, firmware or any combination of these. The invention or some of the features thereof can also be implemented as software running on one or more data processors and/or digital signal processors.

5

The individual elements of an embodiment of the invention may be physically, functionally and logically implemented in any suitable way such as in a single unit, in a plurality of units or as part of separate functional units. The invention may be implemented in a single unit, or be both physically and functionally distributed

10 between different units and processors.

It should be noted that embodiments and features described in the context of one of the aspects of the present invention also apply to the other aspects of the invention. Thus, for example the embodiments/claims relating to the method of

15 the invention also apply to the embodiments relating to the use-

embodiments/claims. Hence, individual features mentioned in different claims, may possibly be advantageously combined, and the mentioning of these features in different claims does not exclude that a combination of features is not possible and advantageous.

20

Although the present invention has been described in connection with the specified embodiments, it should not be construed as being in any way limited to the presented examples. The scope of the present invention is to be interpreted in the light of the accompanying claim set. In the context of the claims, the terms

25 "comprising" or "comprises" do not exclude other possible elements or steps. Also, the mentioning of references such as "a" or "an" etc. should not be construed as excluding a plurality. The use of reference signs in the claims with respect to elements indicated in the figures shall also not be construed as limiting the scope of the invention.

30

All patent and non-patent references cited in the present application, are hereby incorporated by reference in their entirety.

The invention will now be described in further details in the following non-limiting

35 examples.

Examples

Example 1 – Coupling enzyme activity to cell growth

Aim of study

The goal of this step is to link the growth rate of a cell to the activity of the target enzyme, such that the growth rate of a cell is positively or negatively correlated to the activity of the enzyme variant that the cell contains. This enables the use of growth rate as a metric for the enzyme activity.

Materials and methods

In this example, the selection strain is constructed based on WO2018037098A1, where the activity of SAM-dependent methyltransferases is coupled to cell growth by providing precursors for cysteine synthesis. The specific strain used in this example is *E. coli* BW25113 with FolE (T198I) YnbB (V197A) Δ tnaA Δ cysE Δ cfa ompW- Δ (26 bp intergenic region)>yciE::cys3-cys4.

Results and conclusion

The SAM-dependent methyltransferase used here is *Arabidopsis thaliana* N-acetylserotonin O-methyltransferase (atASMT) (**SEQ ID NO: 3**), and the target catalytic activity is the conversion of protocatechuic acid (PCA) into vanillic acid. (EC 2.1.1.1.).

Example 2 – Library generation and barcoding

Aim of study

This step creates a library of DNA variants with unique DNA barcodes attached. The DNA variants encode enzyme variants that are expected to have different fitness levels. The barcodes will serve two purposes in later steps: 1) for accurate full-length sequence reading from long-read sequencing; 2) for using the short barcodes to profile the frequency change of variants during growth-based selection. (Figure 1)

Materials and methods

Error prone PCR

Error prone PCR was performed on the template DNA using GeneMorph II Random Mutagenesis Kit according to manufacturer's instructions, with 1 ng of template

DNA (**SEQ ID NO: 2**) and 30 cycles. The primers used may be found in **SEQ ID NO: 4 and 5**. The forward primer additionally contains sequences used for the assembly of the variants onto the cloning vector in a later step. Several reactions were run in parallel to increase the amount of products. The PCR products were
5 gel-purified to remove residual primers.

Barcoding with one-cycle PCR

- A one-cycle PCR step was performed on the error prone PCR products to attach unique barcodes to the DNA variants. Only one primer is used (**SEQ ID NO: 6**),
10 which contains:
- 1) Sequence corresponding to the reverse primer for the DNA variants.
 - 2) A region with 25 nt of fully random sequences, which serves as barcodes. The possibility space of this region is 4^{25} , or $\sim 10^{15}$, which ensures a high chance that the barcode attached to each variant will be different from all other attached
15 barcodes in a library of less than 10^{10} variants.
 - 3) Sequences used for assembly of the variants with the cloning vector.
- The reaction is performed in one cycle so that each sequence will only have one corresponding barcode. All of purified "error prone PCR" products are used in the reaction (separated into several reactions as needed). 2X Q5 hot-start master mix
20 is used. The reaction is done with 150s of 98 °C for denaturation, 150 seconds at the annealing temperature (68 °C in this case), and 8 min of 72 °C for extension. The prolonged time for each step is to maximize the extent of the reaction for all molecules.
- 25 The one-cycle PCR product is purified through gel purification, which also serves the purpose of removing the leftover single strand in the process.
- Assembly of the barcoded library with cloning vector (**SEQ ID NO: 1**) to obtain the plasmid library. A control library comprising non-mutated template DNA may also be constructed (**SEQ ID NO: 7**).
- 30 The cloning vector, pSD221 (Rep101 ori, cat marker for chloramphenicol resistance), was constructed with a linker for Golden Gate assembly. The linker has two BsaI digestion sites on the two ends of the linker region. The barcoded library is ligated to pSD221 via Golden Gate assembly in several parallel reactions. Afterwards, the reactions were pooled together and purified with AMPure XP beads

(Beckman Coulter) according to manufacturer's PCR purification protocol with 0.9 bead-to-reaction volume. The entire library is eluted to ~25 μ L of water.

Results and conclusion

- 5 The DNA library is obtained through error prone PCR based on an existing enzyme template. The DNA library can alternatively be generated through other means, e.g., DNA synthesis with heterogenous components.
- All the PCR products are tagged with unique barcodes through a one-cycle PCR process. The barcoded library is subsequently assembled onto the vector for
- 10 expression in *E. coli*. In other cases, the library could be assembled with homology arms for integration onto the chromosome.

Example 3 – Construction and amplification of *in vivo* library

Aim of study

- 15 In this step, an *in vivo* library is constructed by transforming the plasmid library (containing the DNA library for enzyme variants) into the selection strain. The *in vivo* library is further amplified by non-selective growth for subsequent uses. (Figure 1)

20 Materials and methods

Transformation of the plasmid library

- Electrocompetent cells for the selection strain were prepared according to commonly used protocols. Several electroporations were done in parallel to transform the full amount of the plasmid library. Immediately after
- 25 electroporation, the cells are added to 2x YT media (~1:10 volume ratio) without antibiotics and incubated for 1 hour at 37 °C for recovery.
- After recovery, the transformant library is stocked in glycerol stock to -80°C. A small volume of the library is serially diluted and plated onto Chloramphenicol plates to estimate the library size per volume of the stocked library.

30

In vivo library amplification

- Given the technologies in 2023, the library size that can be characterized by this invention is one to several millions per run, limited by long-read sequencing capacity. Supposing a limit of 1 million, if the full library size exceeds this limit,
- 35 only a portion of the full library, corresponding to 1 million variants, should be

used for subsequent steps; if the full library size is below 1 million, then the full library can be used for subsequent steps.

An amount of the library based on the consideration above is grown in 2x YT media with chloramphenicol selection overnight for amplification so that aliquots
5 of the same library can be used in subsequent steps. The amount of media should allow sufficient amplification of the library (>100x to ensure that aliquots will be representative of the full library) and sufficient number of aliquots as seeding cultures for subsequent steps. The amplified library is aliquoted and stocked as glycerol stocks at -80°C.

10

Results and conclusion

With these steps, the barcoded enzyme library now exists as *in vivo* library in the selection strain, and aliquots of the same library can be used for the next steps.

15 Example 4 – Long read sequencing

Aim of study

This step uses long-read sequencing technologies (exemplified with Oxford Nanopore, PacBio, etc) to generate data that associates the variant sequences with their associated barcodes. Within a certain library size limit and sufficient
20 read depth, most variant-barcode fragments will be read multiple times, which allows the generation of high-accuracy consensus sequences for the variants based on the clustering of the reads with the same barcode sequences. (Figure 2)

Materials and methods

25 Plasmid extraction and digestion

One aliquot of the *in vivo* library was grown in rich media (2x YT) with chloramphenicol in large volume at 37 °C with shaking. The large volume ensures sufficient plasmid output for Nanopore sample preparation. The plasmid was extracted and purified through maxiprep (ZymoPURE™ II Plasmid Maxiprep Kit)
30 according to manufacturer's instructions.

The extracted plasmid was digested with SalI and SpeI (both from NEB), two unique restriction sites on the plasmid sequence. SalI is 71 bp upstream of the expression promoter, while SpeI is 46 bp downstream of the barcode region. The digestion reaction was run on an agarose gel, and the fragment with length
35 corresponding to the target enzyme gene-barcode cassette was purified.

Nanopore sequencing

This step was done with the R10.4.1 flow cell on a GridION instrument, with sample preparation done using Ligation Sequencing Kit V14 according to the
5 protocol "Ligation sequencing amplicons V14 (SQK-LSK114)" from Oxford Nanopore. Additional materials and instruments were used according to the protocol: NEBNext® Companion Module for Oxford Nanopore Technologies® Ligation Sequencing, Quibit fluorometer and the associated consumables, AMPure XP beads and magnetic rack for 1.5 ml Eppendorf tubes, 1.5 ml Eppendorf DNA
10 LoBind tubes.

After sample preparation, the sample was loaded onto a GridION and run for 39 hours with 9.48 M reads generated (longer run time can be done for higher read output). 100% of the reads were called, and 15.76 Gb bases were called (based on min Q score of 7, though the overall Q score is generally above 12 throughout
15 the run).

Results and conclusion

The plasmid library was extracted after growing an aliquot of the library with antibiotic selection for the plasmid, but without selection for the enzyme variants.
20 The lack of selection in the growth condition avoids bias for different enzyme variants, which maximized the coverage of the library for sequence-barcode association. The region including the target enzyme expression cassette and barcode was isolated by restriction digestion and sequenced through amplicon sequencing with Nanopore sequencing. The sequenced expression cassette
25 includes the promoter and RBS region, which allows the effect of potential mutation in the promoter and RBS on expression level during the enzyme library preparation process to be accounted for.

The long-read Nanopore sequencing generates data that associates the DNA variant sequences with their associated DNA barcodes.

30

Example 5 – Growth-based selection and sampling

Aim of study

In this step, the selection strain comprising the DNA library is grown with selection bias toward higher target enzyme activity. Cells will proliferate with a
35 growth rate positively correlated with the activity of the target enzyme variant.

Sampling during this process, and the subsequent deep sequencing to characterize the change of variant abundance/frequency, will allow the calculation of the growth rate of the cell population harboring a specific enzyme variant. This growth rate will be used as a proxy for the activity of the cognate enzyme variant in this cell population. (Figure 2)

Materials and methods

Prepare media

Prepare non-selective and selective minimal media. The non-selective media consists of 1X M9 salts, 0.1 mM CaCl₂, 2 mM MgSO₄, 0.2% glucose, 1X trace minerals solution, 1X Wolfe's vitamin solution, 25 µg/mL chloramphenicol, 50 mg/l cysteine. The selective media differs from the non-selective media by containing no cysteine, but 1 mM protocetachuaic acid and 1 mM methionine.

Growth-based selection and sampling

An aliquot of the library is grown overnight in non-selective media at 37 °C with shaking. The next day, the culture is washed twice with the selection media, and inoculated into the selection media at an initial OD below 0.1, at 30 °C. Another portion of the overnight culture is spun down, washed twice with water, and the cell pellet is stored at -20 °C, which serves as sample 0 and reflects the initial frequency distribution of the variants.

During growth, three other samples were taken at OD of 0.12, 0.22, 0.39. The cells were washed with water twice, and stored at -20 °C, serving as samples 1, 2, 3.

Results and conclusion

An aliquot of the library is grown overnight non-selectively and seeded into selection media after washing. Samples were taken at time 0, and during the exponential growth phase. The growth rate of cells with a specific variant is assumed to remain constant in the exponential phase and correlate positively with the activity of the enzyme variant.

The distribution of variants in each sample can be captured in the next step through deep sequencing. The OD information will allow the calculation of the expected total cell number given constant volume. Combined, they will allow the calculation of growth rate enabled by each variant.

Example 6 – Deep sequencing**Aim of study**

This step prepares the cell pellet samples for deep sequencing by amplifying the
5 barcode region of the variants cassettes, and deep sequencing of the barcode
region is performed to obtain the frequency information of the barcodes for each
sample. (Figure 2)

Materials and methods**10 Sample preparation for deep sequencing**

The cell pellets are thawed, and water is added to the cell pellets to a final
concentration around 0.3 OD per 10 μ L. Thermolysis is performed to lyse the cells
by heating the cell solution at 95 °C for 10 min. The resulting mixture is spun
down, and the supernatant is kept as PCR template for barcode amplification.

15 The forward primer (**SEQ ID NO: 8**) is directly upstream of the barcode region.
The reverse primer (**SEQ ID NO: 9**) is downstream of the barcode region for a
total length of 352 bp PCR product. The PCR reaction (50 μ L) is set up as follows:
25 μ L Kapa HiFi Hotstart ReadyMix, 1.5 μ L of 10 μ M forward primer, 1.5 μ L of 10
 μ M reverse primers, 5 μ L lysis supernatant, 1.5 μ L DMSO, 15.5 μ L water. The
20 reactions were run 22 cycles according to manufacturer's instructions. The PCR
products were purified with AMPure XP beads with 0.9 bead-to-reaction volume
ratio. The purified products were quality checked with Qubit.

Deep sequencing

25 In this example Deep sequencing is exemplified with Illumina sequencing.
>1 μ g was sent to GENEWIZ Germany GmbH (Azenta Life Sciences) for Illumina
sequencing. The sequencing platform was Illumina NovaSeq 2x150 bp
sequencing, with estimated data output of ~29M paired-end reads (incl. 30% PhiX
spike-in) per sample.

30

Results and conclusion

Through deep sequencing of the barcode region, the frequency change of the
barcodes for samples over the course of growth is captured.

Example 7 – Bioinformatics processing to determine the sequence-activity relationship

Aim of study

The long-read sequencing data and deep sequencing data are processed in this step to generate the full sequence-activity relationship. Specifically, it includes: 1) the processing of deep sequencing data to obtain growth rate (proxy for enzyme activity) for each barcode; 2) the processing of long-read sequencing data to generate consensus sequence for each variant by clustering the reads with the same barcode sequence, assuming the same barcode corresponds to the same variant sequence. The consensus sequence alleviates the high error rate issue of current long-read sequencing methods and allows high accuracy of the variant sequence data. The barcode is used to link the full-length consensus sequence to the corresponding growth rate. (Figure 2)

Materials and methods

Barcode-based clustering of long-read sequencing data

All Nanopore files are consolidated into a single FASTQ file. The barcode of each sequence is identified by 8 bp upstream and downstream of the barcode region. All barcodes are collected into a single file. Given the bidirectional nature of Nanopore reads, barcodes are extracted for both the forward and reverse complements. Length validation of 25 bases is again performed for each barcode. All barcodes extracted from the reverse complement are changed to conform to the top strand and the 5' to 3' end direction, meaning that the reverse complement of these barcodes was taken. The output is two files, one containing all sequences and the other containing all barcodes from the Nanopore data set is generated.

Sequences with the same barcodes (accounting for reverse complements) are grouped together as a cluster. Each corresponding barcode is also saved in a separate file for technical purposes.

Barcode count from deep sequencing data

For each raw FASTQ file, barcodes are counted by parsing all sequences in the file. The barcode is identified by starting 31 nt upstream of the barcode and 4 nt downstream of the barcode. A length validation is performed to ensure that the

extracted sequence is precisely 25 nt. The output is the count of occurrences of each distinct barcode for each Illumina sample.

Each experiment has counts extracted at two exponential growth time points.

These are combined into one csv file with the counts at time point 1 and 2 for 5 each barcode.

Consensus sequence generation for barcodes

The consensus sequence for each barcode is generated with medaka using r1041_e82_400bps_sup_v4.2.0 (<https://github.com/nanoporetech/medaka>). If 10 compute is limited, consensus sequence generation is done only for barcodes of interest based on the Illumina sequence processing results.

Amino acid sequence generation and mutation analysis

The full variant sequence is segregated into promoter-RBS sequences, and the 15 coding sequences.

The coding sequence is translated into amino acid sequence using BioPython. The mutations at the nucleotide and amino acid levels for each consensus sequence is subsequently generated.

20 Growth rate calculation for each variant sequence

OD measurements at each time point are combined with barcode counts and the time differential (Δt) to calculate the growth rate (μ) for each variant:

$$x_1 = \frac{OD_1 \times \text{Count_df1}}{\text{Sum of Count_df1}}$$

$$x_2 = \frac{OD_2 \times \text{Count_df2}}{\text{Sum of Count_df2}}$$

$$\mu = \frac{\log_2(x_2) - \log_2(x_1)}{\Delta t}$$

25

x is the amount of OD for the variant at a given time point.

Results and conclusion

The frequency of each barcode can be obtained from the Illumina data. Combined with the time and OD information (OD provides information about the total cell amount), the growth rate is calculated for each barcode.

The Nanopore data is processed by barcode-based clustering and consensus

5 sequence generation. The consensus sequence will have higher accuracy than the raw reads after error correction based on the multiple raw reads.

Combining all the information above, the growth rate and full-length sequence information is obtained for individual variants.

Example 8 – Training machine learning model with sequence-fitness data 10 and predicting fitness of sequences not seen in the data

Aim of study

A machine learning model is trained to predict amino acid sequence fitness using the sequence-fitness data obtained from the procedure of the previous Examples.

This model incorporates a pretrained embedding model and a specialized

15 ensemble of regressors that is trained on the embeddings of the sequence-fitness data.

Materials and methods

Model Training Process:

1) Embedding: Each sequence from the sequence-fitness dataset was processed
20 using the pretrained embedding model 'esm1_t34_670M_UR50S', sourced from a public Github repository [<https://github.com/facebookresearch/esm#available-models>].

2) Ensemble Training: A diverse ensemble of regression models, including neural networks, Bayesian regression, and decision trees, was trained on a selected
25 subset of data points. Each data point comprises an embedded amino acid sequence paired with its fitness. The code to run the ensemble is from the public github repository [<https://github.com/fhalab/MLDE>].

3) Selection: The top 5 models in the ensemble, selected based on their generalization error, are retained for inference.

Inference Process:

- 1) Embedding: A set of variant sequences of atASMT are embedded using the same pretrained model. Each variant contains three substitution mutations of the original sequence. These substitution mutations are chosen independently and
- 5 randomly from the mutations seen in the training data. A total of 1000 variants are generated.
- 2) Fitness Prediction: The prediction of fitness is conducted by the trained ensemble of regression models. These predictions are then averaged to yield a consolidated fitness estimate for each variant.
- 10 3) Selecting a useful subset as a library: The variants are ranked by the predicted fitness score. A subset of the highest ranked sequences is then selected, yielding advantageous variants of the original atASMT sequence (**SEQ ID NO. 3**).

Results and conclusion

The top-5 most highly ranked variants of the 1000 examined sequences are **SEQ**

15 **ID NO. 10-14**. These sequences are predicted to have the highest fitness level given the experimental conditions, based on the machine learning model trained with the data obtained through the procedure in the preceding examples.

The top-5 most highly ranked variants (**SEQ ID NO. 10-14**) comprise the following mutations on the original atASMT sequence (**SEQ ID NO. 3**):

SEQ ID NO.	Mutations
SEQ ID NO. 10	P72S, K132R, D147E
SEQ ID NO. 11	P72S, I103N, P286L
SEQ ID NO. 12	R77C, G105S, K187R
SEQ ID NO. 13	R77C, M215K, P231S
SEQ ID NO. 14	E46D, S84I, P302L

20

Example 9 – Modified sample collection procedure

Aim of study

This study shows several steps during sample preparation to improve the size, diversity, and quality of the collected data.

Materials and methods

Library generation and barcoding

Compared with Example 2, the template used in this case is the plasmid DNA
5 extracted from the selected culture from Example 5, rather than the WT
sequence. This is to increase the diversity of the template, as well as to have
more sequences with higher activities, since the template sequences has a higher
proportion of templates with improved initial activities compared with WT. Other
steps are done in the same manner as in Example 2.

10

Construction and amplification of *in vivo* library

Compared with Example 3, the promoter controlling the expression of the variants
were changed from J23100 (**SEQ ID NO. 15**) to a weaker J23110 (**SEQ ID NO.
16**) in order to raise the selection ceiling of the selection range. The size of the
15 library used for subsequent steps is chosen as around 3 million, rather than 1
million. This exceeds the capacity of one Nanopore flow cell to obtain close to 10
reads per sequence but will allow the collection of more high-activity sequence
data based on a subsequent step. Other steps were performed in the same way as
Example 3.

20

Growth-based selection and sampling

Compared with Example 5, several changes were made to improve the selection
process. In the non-selective and selective media, trace minerals solution and
Wolfe's vitamin solution were not added, to avoid the coloration of the media and
25 interference with OD reading when protocatechuic acid is present.

The initial OD of the selection culture was set at 0.1 rather than lower, so as to
avoid extended selection time a high enough OD could be reached for sample
collection. One dilution was made when the OD reached ~0.35, to a second initial
OD of 0.05.

30 Samples were taken at time point 0 hr (OD 0.1, denoted time point 0), 43 hr (OD
~0.24, denoted time point 1), 50 hr (OD ~0.35, denoted time point 2; dilution
into 0.05 initial OD was done at this point), 67 hr (OD ~0.27, denoted time point
3), 71 hr (OD ~0.45, denoted time point 4). All were done in triplicates.

Other steps were performed in the same manner as in Example 5.

35

Long read sequencing

Due to limited data output from single Nanopore runs, three runs were performed on the unselected library to obtain sufficient reads.

In addition, plasmids from one post-selection culture were also processed and
5 sequenced in one Nanopore run. This enriches the reads for the top variants, which eventually leads to more data points for variants with higher activities. The resulting data from these runs were combined for processing. Other steps were performed in the same manner as Example 4.

10 Deep sequencing

Compared with Example 6, the amplicons were barcoded with custom barcodes according to the TruSeq UDI scheme. Other conditions were done in the same way as seen in Example 6.

15 **Results and conclusion**

Nanopore sequencing data and Illumina sequencing data were obtained for the samples.

20.8 million reads in total were obtained from Nanopore sequencing, including 19.3 million in non-selected library, and 1.5 million in selected library.

20 20-33 million Illumina reads were obtained for each sample.

Example 10 – Optimized bioinformatics processing to determine the sequence-fitness relationship

Aim of study

25 Compared with Example 7, in this example, additional procedures were taken to improve the quality and output size of the consensus sequence generated from Nanopore data, and to improve the statistical rigor of the fitness estimation for sequence variants.

30

Materials and methods

Mapping Sequence to Growth Rate

Using custom scripts, Illumina reads with a mean -q score below 30 were
35 removed. The barcodes were extracted from the reads producing a CSV file with

the abundances of each unique barcode across all time points. Subsets of the dataset were generated each containing barcode abundances at time point pairs 0–1, 0–2, 0–3, and 0–4.

The DiMSum suite was used to compute the growth rates for each of the dataset
5 subsets with the following parameters –maxSubstitutions 100 –mixedSubstitutions T –indels all –startStage 4 –stopStage 5. The result is a mapping of each barcode to a set of growth rates.

ONT reads were basecalled using the Dorado v5.0.0 sup-accuracy basecaller.

Reads containing multiple concatenated coding regions were split into multiple
10 reads using a custom script. Filtering based on read quality and read length was performed using Chopper with parameters -q 9 -l 800 –maxlength 2000.

UMIC-seq was then used to extract the barcodes from the filtered ONT reads.

Separately, barcodes extracted from Illumina reads were clustered using
mmseqs2 easy-linclust with parameters –min-seq-id 0.88 -c 0.88, and the most
15 abundant barcode in each cluster was collected in a table. Using vsearch –usearch_global with parameters –query_cov 0.84 and –id 0.88, barcodes derived from ONT reads were clustered against the barcodes in the table. For clusters containing seven or more IDs, the reads were collected in a FASTA file.

Using a custom pipeline, a consensus sequence was generated for each cluster.

20 minimap2 was used to align the reads in each cluster to the reference sequence, and racon was used to generate a draft assembly. In three additional rounds, minimap2 and racon were used with the generated draft assembly replacing the reference sequence as the template. The final draft assembly was then polished with medaka (ONT) using one round of medaka_consensus, producing a
25 consensus sequence for each cluster.

The consensus sequences, along with their corresponding barcodes, were compiled into a CSV file. The growth rate scores were then linked to the consensus sequences using the barcode as the identifier.

30 Nanopore read count requirement evaluation

The error rate of the coding region depends in part on the number of reads in the cluster used to produce it. A lower bound for the coding region error rate can be provided based on the read count of the cluster. Self-consistency was estimated by computing the coding region of a set of 700 clusters with 25 reads or more.

35 Having 25 reads or more in a cluster implies the consensus sequence will be

highly accurate. Sub-sampling the clusters, clusters were generated ranging from 7 to 15 in size. For each cluster size between 7 and 15, 5 subsets were created. The coding regions computed from the subsets were compared to those computed from the full cluster. The frequency of mismatched coding regions was then
 5 estimated for cluster sizes 7 to 15, respectively.

Post Processing the Dataset

The coding region and promoter region were extracted using a custom script. The consensus sequence was aligned to the reference sequence, and all bases
 10 between the first and last base of the promoter and coding regions were extracted, respectively.

MoCHI requires the information of the template sequence (wild type) from which the library is derived from. Since the template for the construction of the library was a mixture of variants, wildtype (WT2) coding sequence in this example was
 15 defined as a sequence 6 nucleotide mutations distant from the coding region of the original WT. These mutations were chosen as they are shared by the majority of the variants.

An aggregated version of the dataset was created in which, for each variant v , an aggregated growth rate λ_v^- is produced from the set of growth rates. Using the
 20 equation below,

$$\lambda_v^- = \frac{\sum_{(t_1, t_2) \in T_v} \frac{\lambda_{v, (t_1, t_2)}}{\sigma_{v, (t_1, t_2)}^2}}{\sum_{(t_1, t_2) \in T_v} \frac{1}{\sigma_{v, (t_1, t_2)}^2}}$$

the aggregated growth rate λ_v^- for a variant v was computed using only the time-
 25 step pairs T_v for which the variant has both $\lambda_{v, (t_1, t_2)}$ and $\sigma_{v, (t_1, t_2)}^2$ values.

Results and conclusions

The Nanopore read count requirement evaluation results show around 3.5% error rate for clusters of size 7, decreasing significantly to 1.9% for clusters of size 10
 30 and 1.3% for clusters of size 15.

The resulting dataset contains 571,043 records for variants with at least 7 reads. The distribution of relative enzyme fitness for the aggregated dataset is illustrated

in **Figure 3**. Relative enzyme fitness is defined as the growth rate normalized to the median wild-type (WT2) growth rate, quantifying each variant's growth as a multiple of this baseline.

A non-aggregated version of the dataset was also used, where growth rate
5 derives only from time step pair 0-4. Using a specific time step pair has the advantage that all growth rates have an associated variance score. Time step pair 0-4 was used as it showed the highest correlation for predicted growth rate of variants between the replicates, with an average correlation of 0.917. The dataset contains 464,014 records.

10 The number of variants were found with a statistically significant higher growth rate than the WT2. To have a measure of growth rate variance, the non-aggregated dataset corresponding to time step pair 0-4 was used. Variants with the WT2 coding and promoter region were used to compute a single score for the WT2 growth rate and variance. Non-WT2 variants were then compared with the
15 WT2 using one-tailed p-values, calculated based on the z-scores, with the False Discovery Rate (FDR) correction applied to account for multiple testing. Using the critical value of 0.01, a total of 12,115 variants were found to have a higher growth rate than the WT2.

The quality of the sequence-growth rate relationship was assessed in part by
20 plotting the growth rate distribution of variants that are inactive. Any variant is assumed to be inactive if its coding region has a stop codon before or at the active site. Only variants with a non-mutated coding and promoter region were considered to be WT2. In **Figure 4A** below, the distribution of growth rates is shown for variants with premature stop codons using the aggregated dataset,
25 based on variants with at least 7 reads.

In **Figure 4B** below, the distribution of growth rates is shown for variants with premature stop codons using the aggregated dataset, based on variants with at least 15 reads.

The results show reasonably symmetrical distributions around 0 with few outliers.

30 Comparable levels of dispersion are observed for the WT2, as illustrated in **Figure 4C**.

Example 11 - Learning from the data to generate sequence variants

Aim of study

The data from Example 10 was used to generate a series of variant sequences using the MoCHI suite, which identifies beneficial and non-beneficial mutations by considering both individual and epistatic effects. The resulting library combines highly beneficial mutations and epistatic interactions while avoiding less beneficial ones.

Materials and methods

Training a Second-Order Model Using MoCHI

MoCHI (Faure, Andre & Lehner, Ben. (2024). *MoCHI: neural networks to fit interpretable models and quantify energies, energetic couplings, epistasis and allostery from deep mutational scanning data*) was used to fit a second-order model predicting variant growth rates based on additive effects of individual mutations and pairwise epistatic interactions. The model training parameters were set to `-min_observed 4`, `-max_interaction_order 2`, `-learn_rate 0.02`, and `-transformation ReLU`, with 10-fold cross-validation. A filtered version of the non-aggregated dataset, corresponding to time step pair 0-4, was used as the training set to ensure an accurate measure of variance. Variants containing indels, promoter region changes, or stop codons before the last active site were excluded, resulting in a dataset of 103,731 records. MoCHI provides a score for each single mutation and pairwise epistatic interaction inferred from the training set, as well as a score for the WT2. Using the defined parameters, a positive score indicates a beneficial effect on the growth rate, while a negative score indicates a detrimental effect.

Generating Variants Using the Second-Order MoCHI Model

Scores obtained for mutations and pairwise epistatic interactions were used to generate a library of high-growth-rate variants. The variants are generated at amino acid level. A greedy iterative sampling algorithm was employed to identify a set of mutations that, when applied to the WT2, maximize the predicted growth rate according to the MoCHI model. To generate variants, the sum of additive traits were optimized. A function f is defined as the sum of the mean single mutation terms $\underline{S}$ and the mean pairwise epistasis terms $\underline{E}$, as shown in equations below:

$$f(x) = \sum_{m \in x} \underline{S}_m + \sum_{\{m_i, m_j\} \subset x} \underline{E}_{m_i m_j}$$

Since the second-order model was trained using ten-fold cross-validation, up to ten different scores are available for each single mutation or pairwise epistasis term.

- 5 The sampling algorithm operates in two phases. In the first phase, an empty set of mutations (the nominated set) is initialized. During each iteration, a position not yet occupied by a mutation in the nominated set is randomly sampled. All mutations with an available score (candidate mutations) are then evaluated using equation below:

10
$$\Delta f_{\text{phase1}}(m_{\text{new}}, N) = f(N \cup \{m_{\text{new}}\}) - f(N)$$

From the set of candidates with a positive Δf_{phase1} (beneficial mutations), one mutation was sampled and added to the nominated set. The probability distribution for sampling was determined by applying the softmax function to the Δf_{phase1} values of all beneficial mutations, using a temperature T_{max} . This process

- 15 was repeated until the desired number of mutations was reached.

In the second phase, mutations in the nominated set are iteratively removed and replaced with better ones, if available. During each iteration, a position not yet occupied by a mutation in the nominated set is randomly sampled. All possible combinations of adding a candidate mutation and removing an existing one from

- 20 the nominated set are evaluated using equation below:

$$f_{\text{phase2}}(m_{\text{old}} \rightarrow m_{\text{new}}, N) = f((N \setminus \{m_{\text{old}}\}) \cup \{m_{\text{new}}\}) - f(N)$$

For each candidate m_{new} , the mutation swap $m_{\text{old}} \rightarrow m_{\text{new}}$ with the highest Δf_{phase2} was added to the set of beneficial swaps, provided Δf_{phase2} was greater than 0. From this set, one swap was sampled, adding m_{new} and removing

- 25 m_{old} from the nominated set. The probability distribution was determined by applying the softmax function to the Δf_{phase2} values of all beneficial swaps.

Sampling was performed using a temperature T_i , which followed a linear schedule from T_{max} to T_{min} . The iteration process stopped when for all positions no more beneficial swaps were available or the maximum number of iterations, I_{max} , was

- 30 reached.

A subset of the generated variants was selected for synthesis. Variants with a negative pairwise epistatic interaction term below a threshold—set at half the

negative score of the WT2—were excluded. Additionally, variants were removed if less than one-third of their potential epistatic effects were accounted for by the MoCHI model. From the filtered pool, those offering the best balance of diversity and growth rate were selected for synthesis.

- 5 200 sequences were generated each with 7 and 15 mutations using $I_{max} = 1000$, and 50 sequences each with 25 and 35 mutations using $I_{max} = 1500$. The temperature parameters, T_{max} and T_{min} , were set to 5 and 0.5, respectively. Duplicate sequences were removed, and filters were applied, reducing the total number of sequences to 115. 3 sequences were selected each with 7 and 15
10 mutations, and 2 sequences each with 25 and 35 mutations.

To convert amino acid sequences to DNA sequences for synthesis, the coding region of the WT2 was used as the template, and introduced the amino acid mutations using the most frequent codon for each desired amino acid in *E. coli*.

15 Deterministic Naive Variant Generation Using the Second-Order Model

As a comparison, a library with minimal consideration of epistasis was generated. Sequences were generated by sequentially adding mutations from the set of available mutations, in decreasing order of score, to the nominated set until the desired number of mutations was reached. The mutation set included all single
20 mutations and pairs provided by the MoCHI model. Single mutations were assigned the score provided by MoCHI, while pair scores were calculated by combining the individual mutation scores and the epistatic term, based on the equation below:

$$pair(m_i, m_j) = \frac{S_{m_i} + S_{m_j} + E_{m_i m_j}}{2}$$

- 25 Pairwise epistasis with mutations already in the nominated set was ignored. When a pair was selected, both mutations were added to the set. In the final iteration, only single mutations were considered. Sequences were designed according to the algorithms described above.

30

Results and conclusions

In total, the second-order model produced through MoCHI processing provided 3,386 single mutation terms and 90,831 pairwise epistasis terms, achieving an R^2

of 0.74, implying the model explains 74% of the variance in growth rate on the held-out test data for each fold.

Example 12 - Experimental validation of the data and the generated

5 sequences

Aim of study

Experimental work was done to evaluate how well the data obtained from the invented method aligns with assays with lower throughput, as well as to evaluate the performance of sequences generated based on learned information from the
10 data.

Materials and methods

Strain constructions

Multiple variants with different estimated fitness levels from the dataset in
15 Example 10 were selected for validation experiments. They were obtained by PCR amplification from the miniprepplasmids of non-selected or selected libraries. They were subsequently assembled into pSD293 under the control of J23110 (**SEQ ID NO. 16**). Sequences were verified by Sanger sequencing after the strains containing the resulting expression plasmids were obtained.
20 For growth assays, strains were constructed by introducing the expression plasmids into the selection strain. For bioconversion assays, strains were constructed with E. coli BW25113 instead.

Sanger sequencing for sequence validation

25 A series of variants with generated consensus sequences in the high throughput dataset chosen to be Sanger sequencing-verified. 9 out of 10 with indels, and 32 out of 37 without indels were successfully cloned, which were sent for sequencing.

Growth assay

30 The selection strain expressing variants of ASMT under the control of J23110(**SEQ ID NO. 16**) were grown overnight in 2xYT media supplemented with cysteine supplemented with 25 µg/mL chloramphenicol. These precultures were washed with equal volume of 0.85% NaCl. 5 µl of the washed precultures were inoculated into 245 µl of the same selection media as in Example 9. The cultures were grown

in a Growth Profiler and growth was monitored over time. Growth rate was estimated with the built-in software of Growth Profiler.

Bioconversion assay

- 5 *E. coli* BW25113 strains expressing variants of ASMT under the control of J23110 (**SEQ ID NO. 16**) were grown overnight in 2xYT media supplemented with 25 µg/mL chloramphenicol in 96-deepwell plates placed in a MaxQ 8000 incubator set to 30 °C and 300 rpm. On the next day, each overnight culture was used to inoculate two new 500 µL cultures in new 96-deelwell plates with 2xYT
- 10 supplemented with chloramphenicol to an initial OD of 0.05. The new plates were placed back in the MaxQ 8000 incubator at 30 °C, 300 rpm for 5 hours when the OD600 of the cultures was measured to be between 3.0 and 5.3. The plates were put on ice, and cell mass was normalized by transferring between 262 and 410 µL culture to a new plate, ensuring a final UOD600 of 1.23 for all wells. The plates
- 15 were then centrifuged at 1500 ×g for 15 minutes to pellet the cells. The supernatant was removed, and the cell pellets were stored at -70 °C overnight. Next day, the cell pellets were equilibrated to room temperature, and 70 µL of room-temperature B-PER Complete Bacterial Protein Extraction Reagent was then added to each pellet. The resuspended pellets were incubated at room
- 20 temperature with gentle rocking (40 rpm) for 15 minutes. After lysis, 50 µL of B-PER lysate was mixed with 50 µL of bioconversion buffer consisting of an aqueous solution of 1 mM protocatechuic acid (PCA), 2 mM L-methionine (L-met) and 2 mM ATP disodium salt. The 100 µL bioconversion reactions were incubated in sealed microtiter plates for 120 minutes at 30 °C and
- 25 246 rpm in a Labnet 311DS Environmental Shaking Incubator. Reactions were stopped by heating the microtiter plates in a 90 °C water bath for 5 minutes. The microtiter plates were centrifuged at 1500 ×g for 15 minutes to pellet cell debris. The supernatant was filtered into HPLC sample plates through AcroPrep™ Advance Plates, Short Tip (Cytiva, 97052-096) and diluted 1:1 with blank buffer
- 30 to ensure sufficient volume for HPLC. Blank buffer components were 0.5 mM PCA, 1 mM L-met, 1 mM ATP disodium salt.
- Samples were analyzed on an Ultimate 3000 HPLC with sPFP column coupled to a DAD-3000 module. The sample injection volume was 10 µL, and the eluents were 10 mM ammonium formate and 100% acetonitrile. The ratio of acetonitrile was
- 35 initially 5%, then 60% at 5 minutes, 90% at 5.5 minutes, and 5% at 7.6 minutes.

The flow rate was constant a 0.7 ml/min. The length of the method was 10 minutes. Vanillic acid was observed as a peak in absorbance at 277 nm at a retention time of 3.7 minutes.

Graphpad Prism 10.1.12 was used to generate graphs and perform linear regression and related analyses.

Results

Sanger sequencing validations

The consensus sequences obtained based on Nanopore reads were perfectly aligned with the Sanger sequencing results in all 41 cases, including 9 with indels and 32 without indels.

Growth assay results

Strong positive correlation was observed between the estimated fitness in dataset and the measured fitness in the growth assay, with R^2 value of 0.87, as shown in **Figure 5**.

Bioconversion assay results

The bioconversion results are shown in **Figures 6A-6D** and in **Table 1**.

ASMT-WT (**SEQ ID NO. 2**) refers to the original wild-type ASMT sequence. ASMT1-ASMT37 refer to ASMT variants from the dataset as seen in Example 10. The rest of the ASMT sequences were generated with the second-order model. It can be seen in **Figure 6A** that the original ASMT WT led to a vanillic concentration of 0.66 mg/L (95% CI [0.50, 0.81], SD=0.11, n=2). The highest vanillic acid concentration observed for the tested variants in the dataset was with ASMT17 at 3.18 mg/L (95% CI [2.33, 4.03], SD=0.61, n=2), 380% times higher than WT (p=0.13, with two-tailed t-test).

The highest titers of vanillic acid were observed for the ASMT variant "seq39_7mut" generated with the second-order model. The mean concentration of vanillic acid in the bioconversion supernatant was 4.23 mg/L (95% CI [3.08, 5.38], SD=0.83, n=2). This is 541% higher than the WT (p=0.09, with one-tailed t-test), and 33% higher (p=0.49, with one-tailed t-test) than the best tested sequence from the dataset.

Seq39_7mut is 7 AA mutations from the WT2, 11 AA from the original WT, and at least 5 AA mutations from any sequence in the training dataset used for MoCHI.

Figure 6B shows vanillic acid production plotted against predicted growth rate for all cloned ASMT variants identified during selection.

Figure 6C shows vanillic acid production plotted against predicted growth rate for functional, cloned ASMT variants identified during selection with detectable levels of vanillic acid in bioconversion experiment.

Figure 6D shows vanillic acid production plotted against predicted growth rate for functional, cloned ASMT variants identified during selection with detectable levels of vanillic acid in bioconversion experiment.

Table 1: *Estimated growth rate and vanillic acid production in bioconversion experiment by all cloned ASMT variants. The dataset includes ASMT variants identified from the selected population ASMT1-ASMT37, and novel ASMT variants generated using the second-order model. Standard deviation (SD) is indicated. N/A is non-available data.*

Variant name	Estimated growth rate [1/h]	SD for estimated growth rate [1/h]	Vanillic acid production [mg/L]	SD for vanillic acid production [mg/L]
ASMT-WT	N/A	N/A	0.66	0.11
ASMT1	0.1667	0.00152	2.02	0.38
ASMT2	0.1619	0.00088	2.54	0.32
ASMT3	0.1623	0.00076	2.59	0.22
ASMT4	0.1596	0.00093	2.23	1.12
ASMT5	0.1321	0.00075	1.41	0.09
ASMT6	0.1086	0.00116	1.56	0.72
ASMT7	0.0953	0.00112	1.17	0.14

ASMT8	0.0689	0.00096	2.24	0.11
ASMT9	0.0265	0.00013	1.68	0.59
ASMT10	0.0184	0.00094	1.84	0.46
ASMT11	0.0008	0.00171	1.37	0.94
ASMT12	-0.0082	0.00232	0.00	0.00
ASMT15	0.0034	0.00387	0.00	0.00
ASMT17	0.1591	0.00096	3.18	0.61
ASMT18	0.1432	0.00144	2.06	0.07
ASMT20	0.1017	0.00106	2.29	0.83
ASMT21	0.0672	0.00083	1.91	1.11
ASMT22	0.0631	0.00131	2.47	1.54
ASMT24	0.0376	0.00147	2.24	0.56
ASMT25	0.0323	0.00129	1.28	0.88
ASMT26	0.0265	0.00013	1.10	0.02
ASMT27	0.0299	0.00179	1.36	0.05
ASMT28	0.0265	0.00013	1.29	0.31
ASMT29	0.0128	0.00139	0.76	0.16
ASMT30	0.0157	0.00213	0.90	0.07
ASMT31	0.0095	0.00141	0.92	0.07
ASMT32	-0.0027	0.00681	0.00	0.00
ASMT33	-0.0085	0.00323	0.00	0.00
ASMT34	-0.0057	0.00437	0.00	0.00

ASMT35	-0.0088	0.00419	0.00	0.00
ASMT36	-0.0118	0.00452	0.00	0.00
ASMT37	-0.0121	0.00371	0.00	0.00
seq9_7mut	N/A	N/A	2.12	0.17
seq29_25mut	N/A	N/A	0.00	0.00
seq31_15mut	N/A	N/A	0.30	0.04
seq31_35mut	N/A	N/A	0.00	0.00
seq32_7mut	N/A	N/A	1.47	0.25
seq39_7mut	N/A	N/A	4.23	0.83
seq48_25mut	N/A	N/A	0.49	0.04
seq103_15mut	N/A	N/A	0.63	0.04
seq139_15mut	N/A	N/A	0.75	0.26
seq149_15mut	N/A	N/A	0.68	0.07
seq49_35mut	N/A	N/A	0.54	0.09
seq69_7mut	N/A	N/A	0.00	0.00

Conclusion

The validation experiment results for the samples data points in the dataset show that,

- 5 1) The consensus sequences generated from Nanopore sequencing reads with the same barcodes have very low error rates (0% based on the 41 that were checked).
- 2) The growth rate data obtained in the high-throughput manner is highly correlated with the actual growth rate data ($R^2 = 0.87$).
- 10 3) The estimated fitness levels obtained in the high-throughput manner are positively correlated with results in the bioconversion

assay. The overall correlation was strong. Those with low predicted fitness (below 0.01) led to no detectable product formation, while those above that threshold led to significant product formation in almost all cases.

- 5 In addition, many generated sequences based on learned mutational effects showed significant activity, despite not having been seen in the dataset. In particular, one generated sequence showed potentially higher product formation than any tested variants. These results indicate the great value of using the high-throughput dataset obtained with the method for enzyme engineering.

10

Sequence listing

SEQ ID NO. 1	Plasmid backbone	SEQ ID NO. 9	Reverse primer for barcode amplification
SEQ ID NO. 2	atASMT DNA (ASMT-WT)	SEQ ID NO. 10	atASMT P72S, K132R, D147E
SEQ ID NO. 3	atASMT amino acid (ASMT-WT)	SEQ ID NO. 11	atASMT P72S, I103N, P286L
SEQ ID NO. 4	Forward primer for atASMT	SEQ ID NO. 12	atASMT R77C, G105S, K187R
SEQ ID NO. 5	Reverse primer for ASMT	SEQ ID NO. 13	atASMT R77C, M215K, P231S
SEQ ID NO. 6	One-cycle PCR primer	SEQ ID NO. 14	atASMT E46D, S84I, P302L
SEQ ID NO. 7	Library plasmid with non-mutated atASMT	SEQ ID NO. 15	J23100 promoter
SEQ ID NO. 8	Forward primer for barcode amplification	SEQ ID NO. 16	J23110 promoter

Items of the invention

1. A method for estimating fitness of enzyme variants, the method comprising
 - a) providing a DNA variant library, preferably a plasmid library, the library comprises DNA variants coding for enzyme variants, wherein each DNA
5 variant is linked to one or more unique DNA barcodes, preferably through a one-cycle PCR process;
 - b) transforming the DNA variant library into a host cell for expressing the enzyme variants from the library;
 - I. growing the host cell comprising the DNA variant library, preferably
10 under conditions in the absence of a selection pressure favoring enzyme variants with increased activity; and
 - II. growing the host cell comprising the DNA variant library, under selection pressure favoring enzyme variants with increased catalytic activity;
 - 15 c) isolating DNA variant library from the host cells grown under condition I.) and isolating at least the barcode region of the DNA variant library from the host cells grown under condition II.);
 - d) subjecting the isolated DNA variant library grown under condition I.) to long-read sequencing, thereby obtaining first set of information, such as a
20 first database, linking a specific DNA barcode to a specific DNA variant coding for a specific enzyme variant,
 - e) subjecting the isolated DNA variant library grown under condition II.) to short-read deep sequencing, thereby obtaining a second set of information, such as a second database, linking a specific DNA barcode to a frequency of
25 the specific DNA variant coding for a specific enzyme variant or a rate of change of the specific DNA variant coding for a specific enzyme variant;
 - f) comparing the first set of information from step d) to the second set of information from step e), thereby correlating specific enzyme variants to the frequency information, such as the frequency of the variant; and
 - 30 g) optionally estimating the enzyme fitness of the specific enzyme variant from the determined frequency information.
2. A method for optimizing a nucleotide sequence encoding an enzyme variant, the method comprising

- a) providing a DNA variant library, preferably a plasmid library, the library comprises DNA variants coding for enzyme variants, wherein each DNA variant is linked to one or more unique DNA barcodes, preferably through a one-cycle PCR process;
- 5 b) transforming the DNA variant library into a host cell for expressing the enzyme variants from the library;
- I. growing the host cell comprising the DNA variant library, preferably under conditions in the absence of a selection pressure favoring enzyme variants with increased activity; and
- 10 II. growing the host cell comprising the DNA variant library, under selection pressure favoring enzyme variants with increased catalytic activity;
- c) isolating DNA variant library from the host cells grown under condition I.) and isolating at least the barcode region of the DNA variant library from the
- 15 host cells grown under condition II.);
- d) subjecting the isolated DNA variant library grown under condition I.) to long-read sequencing, thereby obtaining first set of information, such as a first database, linking a specific DNA barcode to a specific DNA variant coding for a specific enzyme variant,
- 20 e) subjecting the isolated DNA variant library grown under condition II.) to short-read deep sequencing, thereby obtaining a second set of information, such as a second database, linking a specific DNA barcode to a frequency of the specific DNA variant coding for a specific enzyme variant or a rate of change of the specific DNA variant coding for a specific enzyme variant;
- 25 f) comparing the first set of information from step d) to the second set of information from step e), thereby correlating specific enzyme variants to the frequency information, such as the frequency of the variant; and
- g) obtaining the optimized nucleotide sequence encoding an enzyme variant from the determined frequency information.
- 30
3. The method according to any of items 1 or 2, wherein the method further comprises a step between step d) and e) of subjecting the isolated DNA variant library grown under condition II.) to long-read sequencing.

4. The method according to any of items 1 or 3, wherein identification of the enzyme variant fitness comprises determining the fitness of a portion of enzyme variants in the DNA variant library, preferably a portion of the variants with the highest activity.

5

5. The method according to any of items 2 or 3, wherein identification of the optimized nucleotide sequence encoding an enzyme variant comprises determining the fitness of a portion of enzyme variants in the DNA variant library, preferably a portion of the variants with the highest activity.

10

6. The method according to any of items 1-5, wherein the enzyme fitness is selected from the group consisting of enzyme activity, expression level, turnover number such as unimolecular rate constant (K_{cat}), Michaelis constant (K_M), enzyme solubility, enzyme stability, substrate specificity, product specificity,

15 thermotolerance, and combinations thereof.

7. The method according to any of the preceding items, wherein the DNA variant library is prepared by the steps:

- a) providing at least one DNA sequence encoding an enzyme variant;
- 20 b) linking one or more unique DNA barcodes on the at least one DNA sequence encoding an enzyme variant and obtaining a barcode-DNA library; and
- c) assembling the barcode-DNA library and obtaining a DNA variant library.

25 8. The method according to item 7, wherein the at least one DNA sequence of step a) is generated by PCR, preferably error prone PCR.

9. The method according to items 7 or 8, wherein the assembly of the barcode-DNA library is by Golden Gate assembly.

30

10. The method according to any of the preceding items, wherein the unique DNA barcode is a random nucleotide sequence.

11. The method according to any of the preceding items, wherein the unique DNA barcode has a length in the range 10-300 DNA nucleotides, preferably 15-100 DNA nucleotides, most preferably 20-30 DNA nucleotides.
- 5 12. The method according to any of the preceding items, wherein the DNA variant is linked to the unique DNA barcode via a covalent bond, such as an extension of the DNA variant coding for the enzyme variant, such as a phosphodiester bond.
13. The method according to any of the preceding items, wherein the DNA library
10 is a vector library, such as a plasmid library.
14. The method according to any of the preceding items, wherein the enzyme is an enzyme for which the enzyme activity and/or enzyme fitness correlates with growth of the host cell, such as growth during selection pressure.
- 15 15. The method according to any of the preceding items, wherein the enzyme is selected from the group consisting of oxidoreductases, transferases, hydrolase, lyases, isomerases, ligases, translocases, or a plurality thereof.
- 20 16. The method according to any of the preceding items, wherein the host cell is a strain selected from the group consisting of a bacterial cell, a yeast cell, and a eukaryotic cell.
17. The method according to any of the preceding items, wherein the host cell is a
25 bacterial cell, preferably *E. coli*.
18. The method according to any of the preceding items, wherein the selection pressure is selected from the group consisting of sensor-based approach, metabolic coupling, and substrate stress relief.
- 30 19. The method according to any of the preceding items, wherein the long-read sequencing is sequencing of at least 601 nucleotides, such as at least 1000 nucleotides, such as at least 5000 nucleotides, such as at least 10,000 nucleotides, such as at least 20,000 nucleotides.

20. The method according to any of the preceding items, wherein the long-read sequencing is selected from the group consisting of third generation sequencing, such as Oxford Nanopore sequencing, PacBio sequencing, Illumina Complete Long Read sequencing, TELL-Seq.

5

21. The method according to any of the preceding items, wherein the short-read sequencing is sequencing of nucleotide sequences in the range 10-600 nucleotides, preferably in the range 30-500, most preferably in the range 50-300 nucleotides.

10

22. The method according to any of the preceding items, wherein the short-read sequencing is selected from the group consisting of deep sequencing, Next-Generation Sequencing (NGS), such as Illumina sequencing.

15 23. The method according to any of the preceding items, wherein the frequency information of the specific DNA variant coding for a specific enzyme variant is a reflection of the enzyme fitness of said specific enzyme variant.

24. The method according to any of the preceding items, wherein the rate of
20 change of the specific DNA variant coding for a specific enzyme variant is determined from the change of an amount of the specific DNA variant over a specific period of time.

25. The method according to any of the preceding items, wherein in step f) the
25 comparison of the first set of information from step d) to the second set of information from step e), is by correlating DNA barcodes from the second set of information to the enzyme variants of the first set of information via the enzyme variant specific DNA barcode, thereby obtaining a prediction of enzyme fitness of the enzyme variants and/or an optimized nucleotide sequence encoding an
30 enzyme variant via the frequency of the DNA barcodes from the second set of information or the rate of change of the specific DNA variant coding for a specific enzyme variant.

26. The method according to any of the preceding items, wherein the enzyme fitness of the specific enzyme variant and/or an optimized nucleotide sequence encoding an enzyme variant is reflected by a growth rate of the host cell.

5 27. The method according to item 22, wherein the growth rate is determined from growing the host cell from step b(I.) under selection pressure favoring enzyme variants with increased fitness or activity.

28. The method according to any of the preceding items, wherein the DNA
10 nucleotide sequence coding for enzyme variants comprises a promoter such as SEQ ID NO. 15 or SEQ ID NO.16.

29. The method according to any of the preceding items, wherein the DNA variant library comprises 100.000-5 million DNA variants, such as 300.000-5 million DNA
15 variants, such as 500.000-4 million DNA variants, preferably 800.000-4 million DNA variants, more preferably 1-3 million DNA variants, most preferably 3 million DNA variants.

30. The method according to any of the preceding items, wherein the host cell is
20 grown under selection pressure at an optical density (OD) of 0.01-0.5, such as 0.3-0.5, preferably 0.05-0.5.

31. The method according to any of the preceding items, wherein the host cell is grown in a colorless media.

25

32. Use of at least one barcoded DNA sequence encoding an enzyme variant for predicting enzyme fitness of said enzyme variant, wherein the at least one barcoded DNA sequence is subjected to long-read sequencing and short-read sequencing.

30

33. Use of at least one barcoded DNA sequence encoding an enzyme variant for optimizing said DNA sequence encoding said enzyme variant, wherein the at least one barcoded DNA sequence is subjected to long-read sequencing and short-read sequencing.

35

34. The use according to any of items 32 or 33, wherein the at least one barcoded DNA sequence for short-read sequencing has been obtained from a host cell subjected to selection pressure favoring enzyme variants with increased fitness.

35. A polypeptide according to SEQ ID NO. 3 comprising one or more of the following mutations P72S, K132R, D147E, I103N, P286L, R77C, G105S, K187R, M215K, P231S, E46D, S84I, and/or P302L.

36. The polypeptide according to item 35 comprising one or more of the following combinations of mutations:

- P72S, K132R, and D147E,
- 10 - P72S, I103N, and P286L,
- R77C, G105S, and K187R,
- R77C, M215K, and P231S, and/or
- E46D, S84I, and P302L.

37. The polypeptide according to any of items 35 or 36, the polypeptide is a polypeptide according to any of SEQ ID NOs 10-14.

38. A processor system programmed to operate according to a machine learning (ML) algorithm for estimating fitness of enzyme variants, the machine learning (ML) algorithm being trained, and/or being trainable on data comprising a correlation of a specific nucleotide sequence encoding an enzyme variant to frequency information.

39. A processor system programmed to operate according to a machine learning (ML) algorithm for estimating fitness of enzyme variants, the machine learning (ML) algorithm being trained, and/or being trainable, on data obtained by a method according to any of items 1-31.

40. A processor system programmed to operate according to a machine learning (ML) algorithm for optimizing a nucleotide sequence encoding an enzyme variant, the machine learning (ML) algorithm being trained, and/or being trainable on data comprising a correlation of a specific nucleotide sequence encoding an enzyme variant to frequency information.

41. A processor system programmed to operate according to a machine learning (ML) algorithm for optimizing a nucleotide sequence encoding an enzyme variant, the machine learning (ML) algorithm being trained, and/or being trainable, on
5 data obtained by a method according to any of items 1-31.

42. Use of a machine learning (ML) algorithm trained on data comprising a correlation of a specific nucleotide sequence encoding an enzyme variant to frequency information to predict fitness for any enzyme sequence, or to guide the
10 generation of enzyme sequences with given fitness properties.

43. Use of a machine learning (ML) algorithm trained on data obtained by a method according to any of items 1-31 to predict fitness for any enzyme sequence, or to guide the generation of enzyme sequences with given fitness
15 properties.

44. Use of a machine learning (ML) algorithm trained on data comprising a correlation of a specific nucleotide sequence encoding an enzyme variant to frequency information for optimizing a nucleotide sequence encoding an enzyme
20 variant, or to guide the generation of an improved nucleotide sequence encoding an enzyme variant with given fitness properties.

45. Use of a machine learning (ML) algorithm trained on data obtained by a method according to any of items 1-31 for optimizing a nucleotide sequence
25 encoding an enzyme variant, or to guide the generation of an improved nucleotide sequence encoding an enzyme variant with given fitness properties.

46. A system suitable for executing an algorithm, such as machine learning (ML) algorithm, for estimating fitness of enzyme variants, the system being trained,
30 and/or being trainable, on data obtained by a method according to any of items 1-31, the system being arranged for:

- receiving a first set of information (1SI), such as a first database, said first of information being obtained, or being obtainable, from d) subjecting the isolated
35 DNA variant library grown under condition I.) to long-read sequencing, thereby

obtaining said first set of information, such as a first database, linking a specific DNA barcode to a specific DNA variant coding for a specific enzyme variant,

- receiving a second set of information (2SI), such as a second database, said
5 second of information being obtained, or being obtainable, from e) subjecting the isolated DNA variant library grown under condition II.) to short-read deep sequencing, thereby obtaining said second set of information, such as a second database, linking a specific DNA barcode to a frequency of the specific DNA variant coding for a specific enzyme variant or a rate of change of the specific
10 DNA variant coding for a specific enzyme variant;

wherein the system comprises:

- a first comparison module (1CM) for comparing said first set of information (1SI) to said second set of information (2SI), thereby correlating specific enzyme
15 variants to the frequency information, such as frequency of the variant; and
- an optional second comparison module (2CM) for correlating determined frequency information to the enzyme fitness of the specific enzyme variant.

47. A system suitable for executing an algorithm, such as machine learning (ML)
20 algorithm, for optimizing a nucleotide sequence encoding an enzyme variant, the system being trained, and/or being trainable, on data obtained by a method according to any of items 1-31, the system being arranged for:

- receiving a first set of information (1SI), such as a first database, said first of
25 information being obtained, or being obtainable, from d) subjecting the isolated DNA variant library grown under condition I.) to long-read sequencing, thereby obtaining said first set of information, such as a first database, linking a specific DNA barcode to a specific DNA variant coding for a specific enzyme variant,
- 30 - receiving a second set of information (2SI), such as a second database, said second of information being obtained, or being obtainable, from e) subjecting the isolated DNA variant library grown under condition II.) to short-read deep sequencing, thereby obtaining said second set of information, such as a second database, linking a specific DNA barcode to a frequency of the specific DNA

variant coding for a specific enzyme variant or a rate of change of the specific DNA variant coding for a specific enzyme variant;

wherein the system comprises:

- 5 -a first comparison module (1CM) for comparing said first set of information (1SI) to said second set of information (2SI), thereby correlating specific enzyme variants to the frequency information, such as frequency of the variant, using the DNA barcode as an identifier; and
- an optional second comparison module (2CM) for correlating determined
- 10 frequency information to the enzyme fitness of the specific enzyme variant.

48. The system suitable for executing an algorithm according to any of items 46 or 47, wherein the algorithm is MoCHI.

- 15 49. A method for training a machine learning (ML) system for estimating fitness of enzyme variants according to item 46, the method comprises the steps of:
- receiving training data comprising a first set of information (1SI), such as a first database, and a second set of information (2SI), such as a second database,
 - training the ML system for estimating fitness of enzyme variants using said
- 20 training data, and
- validating the ML system using correlated specific enzyme variants to the frequency information, such as frequency of the variant.

50. A method for training a machine learning (ML) system for optimizing a
- 25 nucleotide sequence encoding an enzyme variant according to item 47, the method comprises the steps of:
- receiving training data comprising a first set of information (1SI), such as a first database, and a second set of information (2SI), such as a second database,
 - training the ML system for optimizing a nucleotide sequence encoding an enzyme
- 30 variant using said training data, and
- validating the ML system using correlated specific enzyme variants to the frequency information, such as frequency of the variant, using a barcode as an identifier.

51. The method for training a machine learning (ML) system according to any of items 49 or 50, wherein the ML system comprises:

- a ML model configured to assign a score to a nucleotide sequence encoding an enzyme variant,
- 5 • a comparison component configured to compare the score with a threshold for predicting a growth rate of a host cell expressing said nucleotide sequence encoding the enzyme variant, and
- a featuring component configured to receive an output from the comparison component, wherein the output indicates a predicted fitness,
10 such as activity, of said enzyme variant encoded by said nucleotide sequence.

52. The method for training a machine learning (ML) system according to item 51, wherein if the score is above the threshold, it is indicative of an improved growth
15 rate of a host cell expressing said enzyme variant compared to a growth rate of a host cell expressing a corresponding wild-type enzyme.

53. The method for training a machine learning (ML) system according to item 51, wherein if the score is below the threshold, it is indicative of a decreased growth
20 rate of a host cell expressing said enzyme variant compared to a growth rate of a host cell expressing a corresponding wild-type enzyme.

54. The method for training a machine learning (ML) system according to any of items 50-53, wherein the threshold is 0.

25

55. The processor system according to any of items 38-41, the use according to any of items 42-45, the system according to any of items 46-48, and/or the method according to any of items 49-54, being implemented on a computer, such as being computer-implemented.

30

56. A computer-readable storage medium comprising data comprising a correlation of a specific nucleotide sequence encoding an enzyme variant to enzyme fitness, such as enzyme activity.

57. A computer-readable storage medium comprising data obtained by a method according to any of items 1-31.

58. A neural network model obtained or obtainable by a method according to any
5 of items 49-54.

59. A method for converting a substrate into a product comprising:

- 10 a) selecting a parent enzyme capable of enzymatically converting the substrate into the product;
- b) evaluating the fitness of variants of the parent enzyme with respect to enzymatically converting the substrate into the product using the method, algorithm, system, model and/or storage medium of any of the preceding items;
- 15 c) identifying and selecting a nucleotide sequence encoding an enzyme variant having improved fitness compared to the parent enzyme;
- d) transforming the nucleotide sequence into a host cell;
- e) expressing the nucleotide sequence in the host cell to produce the variant enzyme;
- 20 f) optionally isolating the variant enzyme; and
- g) contacting the enzymes variant with the substrate, optionally in the host cell, at conditions to allow the variant enzyme to convert the substrate into the product.

25 60. The method according to item 2, wherein

- the nucleotide sequence from step g) is cloned into an expression host cell;
- the enzyme variant encoded by the nucleotide sequence is expressed from the host cell and optionally purified; and
- 30 - the expressed enzyme variant is incubated with its enzymatic substrate to convert the substrate into the product of the enzymatic reaction.

Claims

1. A method for optimizing a nucleotide sequence encoding an enzyme variant, the method comprising
- 5 a) providing a DNA variant library, preferably a plasmid library, the library comprises DNA variants coding for enzyme variants, wherein each DNA variant is linked to one or more unique DNA barcodes, preferably through a one-cycle PCR process;
 - b) transforming the DNA variant library into a host cell for expressing the enzyme variants from the library;
 - 10 I. growing the host cell comprising the DNA variant library, preferably under conditions in the absence of a selection pressure favoring enzyme variants with increased activity; and
 - II. growing the host cell comprising the DNA variant library, under selection pressure favoring enzyme variants with increased catalytic
 - 15 activity;
 - c) isolating DNA variant library from the host cells grown under condition I.) and isolating at least the barcode region of the DNA variant library from the host cells grown under condition II.);
 - d) subjecting the isolated DNA variant library grown under condition I.) to
 - 20 long-read sequencing, thereby obtaining first set of information, such as a first database, linking a specific DNA barcode to a specific DNA variant coding for a specific enzyme variant,
 - e) subjecting the isolated DNA variant library grown under condition II.) to short-read deep sequencing, thereby obtaining a second set of information,
 - 25 such as a second database, linking a specific DNA barcode to a frequency of the specific DNA variant coding for a specific enzyme variant or a rate of change of the specific DNA variant coding for a specific enzyme variant;
 - f) comparing the first set of information from step d) to the second set of information from step e), thereby correlating specific enzyme variants to
 - 30 the frequency information, such as the frequency of the variant; and
 - g) obtaining the optimized nucleotide sequence encoding an enzyme variant from the determined frequency information.

2. The method according to claim 1, wherein the method further comprises a step between step d) and e) of subjecting the isolated DNA variant library grown under condition II.) to long-read sequencing.
- 5 3. The method according to any of claims 1 or 2, wherein identification of the optimized nucleotide sequence encoding an enzyme variant comprises determining the fitness of a portion of enzyme variants in the DNA variant library, preferably a portion of the variants with the highest activity.
- 10 4. The method according to claim 3, wherein the enzyme fitness is selected from the group consisting of enzyme activity, expression level, turnover number such as unimolecular rate constant (K_{cat}), Michaelis constant (K_M), enzyme solubility, enzyme stability, substrate specificity, product specificity, thermotolerance, and combinations thereof.
- 15 5. The method according to any of the preceding claims, wherein the DNA variant library is prepared by the steps:
- a) providing at least one DNA sequence encoding an enzyme variant;
 - b) linking one or more unique DNA barcodes on the at least one DNA
 - 20 sequence encoding an enzyme variant and obtaining a barcode-DNA library; and
 - c) assembling the barcode-DNA library and obtaining a DNA variant library.
6. The method according to claim 5, wherein the at least one DNA sequence of
- 25 step a) is generated by PCR, preferably error prone PCR.
7. The method according to claims 5 or 6, wherein the assembly of the barcode-DNA library is by Golden Gate assembly.
- 30 8. The method according to any of the preceding claims, wherein the unique DNA barcode is a random nucleotide sequence.
9. The method according to any of the preceding claims, wherein the unique DNA barcode has a length in the range 10-300 DNA nucleotides, preferably 15-100
- 35 DNA nucleotides, most preferably 20-30 DNA nucleotides.

10. The method according to any of the preceding claims, wherein the DNA variant is linked to the unique DNA barcode via a covalent bond, such as an extension of the DNA variant coding for the enzyme variant, such as a phosphodiester bond.
11. The method according to any of the preceding claims, wherein the DNA library is a vector library, such as a plasmid library.
12. The method according to any of the preceding claims, wherein the enzyme is selected from the group consisting of oxidoreductases, transferases, hydrolase, lyases, isomerases, ligases, translocases, or a plurality thereof.
13. The method according to any of the preceding claims, wherein the host cell is a strain selected from the group consisting of a bacterial cell, a yeast cell, and a eukaryotic cell.
14. The method according to any of the preceding claims, wherein the host cell is a bacterial cell, preferably *E. coli*.
15. The method according to any of the preceding claims, wherein the selection pressure is selected from the group consisting of sensor-based approach, metabolic coupling, and substrate stress relief.
16. The method according to any of the preceding claims, wherein the long-read sequencing is sequencing of at least 601 nucleotides, such as at least 1000 nucleotides, such as at least 5000 nucleotides, such as at least 10,000 nucleotides, such as at least 20,000 nucleotides.
17. The method according to any of the preceding claims, wherein the long-read sequencing is selected from the group consisting of third generation sequencing, such as Oxford Nanopore sequencing, PacBio sequencing, Illumina Complete Long Read sequencing, TELL-Seq.

18. The method according to any of the preceding claims, wherein the short-read sequencing is sequencing of nucleotide sequences in the range 10-600 nucleotides, preferably in the range 30-500, most preferably in the range 50-300 nucleotides.

5

19. The method according to any of the preceding claims, wherein the short-read sequencing is selected from the group consisting of deep sequencing, Next-Generation Sequencing (NGS), such as Illumina sequencing.

10 20. The method according to any of the preceding claims, wherein the frequency information of the specific DNA variant coding for a specific enzyme variant is a reflection of the enzyme fitness of said specific enzyme variant.

21. The method according to any of the preceding claims, wherein the rate of
15 change of the specific DNA variant coding for a specific enzyme variant is determined from the change of an amount of the specific DNA variant over a specific period of time.

22. The method according to any of the preceding claims, wherein in step f) the
20 comparison of the first set of information from step d) to the second set of information from step e), is by correlating DNA barcodes from the second set of information to the enzyme variants of the first set of information via the enzyme variant specific DNA barcode, thereby obtaining an optimized nucleotide sequence encoding an enzyme variant via the frequency of the DNA barcodes from the
25 second set of information or the rate of change of the specific DNA variant coding for a specific enzyme variant.

23. The method according to any of the preceding claims, wherein the enzyme fitness of the specific enzyme variant is reflected by a growth rate of the host cell.
30

24. The method according to claim 23, wherein the growth rate is determined from growing the host cell from step b(I.) under selection pressure favoring enzyme variants with increased fitness or activity.

25. The method according to any of the preceding claims, wherein the DNA nucleotide sequence coding for enzyme variants comprise a promoter such as SEQ ID NO. 15 or SEQ ID NO.16.

5 26. The method according to any of the preceding claims, wherein the DNA variant library comprises 100.000-5 million DNA variants, such as 300.000-5 million DNA variants, such as 500.000-4 million DNA variants, preferably 800.000-4 million DNA variants, more preferably 1-3 million DNA variants, most preferably 3 million DNA variants.

10

27. The method according to any of the preceding claims, wherein the host cell is grown under selection pressure at an optical density (OD) of 0.01-0.5, such as 0.3-0.5, preferably 0.05-0.5.

15 28. The method according to any of the preceding claims, wherein the host cell is grown in a colorless media.

29. Use of at least one barcoded DNA sequence encoding an enzyme variant for optimizing said DNA sequence encoding said enzyme variant, wherein the at least
20 one barcoded DNA sequence is subjected to long-read sequencing and short-read sequencing.

30. The use according to claim 29, wherein the at least one barcoded DNA sequence for short-read sequencing has been obtained from a host cell subjected
25 to selection pressure favoring enzyme variants with increased fitness.

31. A polypeptide according to SEQ ID NO. 3 comprising one or more of the following mutations P72S, K132R, D147E, I103N, P286L, R77C, G105S, K187R, M215K, P231S, E46D, S84I, and/or P302L.

32. The polypeptide according to claim 31 comprising one or more of the following
30 combinations of mutations:

- P72S, K132R, and D147E,
- P72S, I103N, and P286L,
- R77C, G105S, and K187R,

- R77C, M215K, and P231S, and/or
- E46D, S84I, and P302L.

33. The polypeptide according to any of claims 31 or 32, the polypeptide is a polypeptide according to any of SEQ ID NOs 10-14.

5

34. A processor system programmed to operate according to a machine learning (ML) algorithm for optimizing a nucleotide sequence encoding an enzyme variant, the machine learning (ML) algorithm being trained, and/or being trainable on data comprising a correlation of a specific nucleotide sequence encoding an enzyme
10 variant to frequency information.

35. A processor system programmed to operate according to a machine learning (ML) algorithm for optimizing a nucleotide sequence encoding an enzyme variant, the machine learning (ML) algorithm being trained, and/or being trainable, on
15 data obtained by a method according to any of claims 1-28.

36. Use of a machine learning (ML) algorithm trained on data comprising a correlation of a specific nucleotide sequence encoding an enzyme variant to frequency information for optimizing a nucleotide sequence encoding an enzyme
20 variant, or to guide the generation of an improved nucleotide sequence encoding an enzyme variant with given fitness properties.

37. Use of a machine learning (ML) algorithm trained on data obtained by a method according to any of claims 1-28 for optimizing a nucleotide sequence
25 encoding an enzyme variant, or to guide the generation of an improved nucleotide sequence encoding an enzyme variant with given fitness properties.

38. A system suitable for executing an algorithm, such as machine learning (ML) algorithm, for optimizing a nucleotide sequence encoding an enzyme variant, the
30 system being trained, and/or being trainable, on data obtained by a method according to any of claims 1-28, the system being arranged for:

- receiving a first set of information (1SI), such as a first database, said first of information being obtained, or being obtainable, from d) subjecting the isolated

DNA variant library grown under condition I.) to long-read sequencing, thereby obtaining said first set of information, such as a first database, linking a specific DNA barcode to a specific DNA variant coding for a specific enzyme variant,

- 5 - receiving a second set of information (2SI), such as a second database, said second of information being obtained, or being obtainable, from e) subjecting the isolated DNA variant library grown under condition II.) to short-read deep sequencing, thereby obtaining said second set of information, such as a second database, linking a specific DNA barcode to a frequency of the specific DNA
10 variant coding for a specific enzyme variant or a rate of change of the specific DNA variant coding for a specific enzyme variant;

wherein the system comprises:

- a first comparison module (1CM) for comparing said first set of information (1SI)
15 to said second set of information (2SI), thereby correlating specific enzyme variants to the frequency information, such as frequency of the variant, using the DNA barcode as an identifier; and
- an optional second comparison module (2CM) for correlating determined frequency information to the enzyme fitness of the specific enzyme variant.

20

39. The system suitable for executing an algorithm according to claim 38, wherein the algorithm is MoCHI.

40. A method for training a machine learning (ML) system for optimizing a
25 nucleotide sequence encoding an enzyme variant according to any of claims 38 or 39, the method comprises the steps of:

- receiving training data comprising a first set of information (1SI), such as a first database, and a second set of information (2SI), such as a second database,
-training the ML system for optimizing a nucleotide sequence encoding an enzyme
30 variant using said training data, and
- validating the ML system using correlated specific enzyme variants to the frequency information, such as frequency of the variant.

41. The processor system according to any of claims 34 or 35, the use according
35 to any of claim 36 or 37, the system according to any of claims 38 or 39, and/or

the method according to claim 40, being implemented on a computer, such as being computer-implemented.

42. A computer-readable storage medium comprising data obtained or obtainable
5 by a method according to any of claims 1-28.

43. A neural network model obtained or obtainable by a method according to claim 40.