TITLE

Towards AI-designed genomes using a variational autoencoder

AUTHOR LIST

Dudek, N.K.[1,2], Precup, D.[1,2]

AUTHOR AFFILIATIONS

1. School of Computer Science; McGill University; Montreal, QC, H3A 0G4; Canada
2. Mila - Québec Artificial Intelligence Institute; Montreal, QC, H2S 3H1; Canada

CORRESPONDING AUTHOR

Natasha Dudek

KEYWORDS

Microbial genomics, machine learning, VAE, generative AI, bioinformatics, synthetic biology

ABSTRACT

Genomes serve as the blueprints for life, encoding elaborate networks of genes whose products must seamlessly interact to support living organisms. Humans' capacity to understand biological systems from scratch is limited by their sheer size and complexity. In this work, we develop a framework for training a machine learning algorithm to learn the basic genetic principles that underlie genome composition. Our variational autoencoder model, DeepGenomeVector, was trained to take as input corrupted bacterial genetic blueprints (i.e. complete gene sets, henceforth "genome vectors") in which most genes had been "removed", and re-create the original. The resulting model effectively captures the complex dependencies in genomic networks, as evaluated by both qualitative and quantitative metrics. An in-depth functional analysis of a generated gene vector shows that its encoded pathways are interconnected and nearly complete. On the test set, where the model's ability to re-generate the original, uncorrupted genome vector was evaluated, an AUC score of 0.98 and an F1 score of 0.83 provide support for the model's ability to generate diverse, high-quality genome vectors. This work showcases the power of machine learning approaches for synthetic biology and highlights the possibility that just as humans can design an AI that animates a robot, AIs may one day be able to design genomic blueprints that animate carbon-based cells.

INTRODUCTION

What is life? A fundamental goal of biology is to understand the basic principles underlying living systems. This hypothetically involves mapping and understanding the function of the molecular processes in a cell, from which genes encode that process to how those genes interact with one another and with their environment. For perspective, it is estimated that 10^{10}–10^{13} genes across 10^{7}–10^{8} bacterial species have evolved since the dawn of life on Earth over four billion years ago [1,2].

Mycoplasmas have been proposed as an optimal model for understanding the basic principles of life, as they are the most minimal, and therefore the most “simple”, cells capable of self-sufficient growth [3]. Working with this model system, [4] used transposon mutagenesis to identify non-essential genes in the *Mycoplasma mycoides* genome (901 genes), and successfully reduced the genome to encode only the most basic gene set required for life in a laboratory setting (473 genes), including 149 of unknown function. This reduction of life to its bare essentials was a major breakthrough in our understanding of what constitutes life and was hailed as setting the stage for the design of life from scratch. Such efforts illustrate that while there is structure (i.e. sets of rules) to genome composition, elucidating that structure using laboratory techniques may be labour intensive, time consuming, and costly [5].

Machine learning (ML) excels at learning complex patterns in data. A generative ML model trained to design blueprints for life could potentially offer insights into basic scientific questions about the genomic building blocks of life and the nature of interactions between poorly understood or uncharacterized proteins and metabolic networks. Additional applications could include be extended to generating synthetic genome blueprints for augmenting training dataset size for ML models (akin to [6,7]), or inferring metabolic potential of partially reconstructed genomes of interest (e.g., somewhat akin to PICRUSt [8]). While the use of ML for *de novo* genome design remains largely unexplored, related tasks have benefited from the use of ML. Within synthetic biology, ML has shown promise for applications such as optimization of metabolite production levels [9–12], gene expression and metabolic flux [13–15], and fermentation [16–18]. Furthermore, generative artificial intelligence (AI) approaches have developed for the design of peptides (e.g., with anti-microbial, anti-cancer, or protein-protein interaction inhibition properties)[19–23], CRISPR guide RNA sequences [24], *de novo* antibodies [25–27], and synthetic promoter sequences [28].

Here, we undertake the design of blueprints for bacterial genomes using generative AI. Specifically, we employ a variational autoencoder (VAE) to model the basic tenets of bacterial genome composition (i.e., gene content). First, we source 2,584 high-quality, complete bacterial genome assemblies from the KEGG database, and transform each into a genetic blueprint or “genome vector” that is a binary representation of the genes encoded in each genome. We create 100 heavily corrupted copies of each genome vector by turning off bits representing genes, thereby augmenting dataset size while also preparing inputs for a reconstruction task. The VAE is then trained to generate the original genome vector from the corrupted input, such that the output of the model is sets of genes that theoretically encode for viable bacterial cells. An in-depth analysis of a generated genome vector suggests that the model learns to generate largely complete, interconnected pathways, while an AUC score of 0.98 and an F1 score of 0.83

suggest that our model has high discriminative ability, precision, and recall. This work highlights the potential of ML algorithms to learn the basic genetic principles that encode for life.

RESULTS

Overview of DeepGenomeVector

In this work we propose and implement DeepGenomeVector, an ML approach to model the gene composition of bacterial genomes. This model generates sets of genes that hypothetically encode for viable bacterial cells. Gene sets are represented in the form of "genome vectors", which we define as a binary representation of all genes encoded by a complete genome.

Annotated bacterial genomes were sourced from a set of 2,584 complete, high-quality, annotated genomes from 46 major bacterial lineages (approximately phyla) from the KEGG genomes database [29–31] (Supplemental Figure 1A, B). Annotations in the form KO identifiers were used to construct genome vectors consisting of 9,863 bits. While KO identifier annotations do not cover the scope of all genes that exist (a feat which is impossible as many genes do not even have known functions), or even likely that are essential for a cell, they do provide annotations for core metabolic and functional processes and are well curated, consistent, high-quality annotations. Genome vectors included in the full dataset for this work (train + test sets combined) encode a median of 1,885 genes, with a range of 528 - 4,536 genes per genome (Supplemental Figure 1C) Genes in the KEGG database are further organized into modules, which are functional units of gene sets. For example, KEGG modules encode reactions such as glycolysis, the pentose phosphate pathway, methanogenesis, denitrification, etc.

One of the biggest obstacles to using data-driven machine learning methods to generate genome vectors is that the total number of complete, high-quality genome assemblies is limited. After splitting genomes into training (~90%) and testing (~10%) datasets (see Methods), we created 100 corrupted versions of every genome vector. The reason 100 corruptions were performed per original genome vector was to augment dataset size. Each corruption was achieved by retaining genes encoding 10 randomly selected modules encoded by the real genome. These corrupted genome vectors served as input to an ML algorithm.

From the generative algorithms family, a VAE was selected for our task of modeling genome composition due to the strong performance of VAEs on generative modeling tasks in general and the ability of an end-user to evaluate a generated output against the input (or the uncorrupted input) in a high-throughput manner. In brief, VAEs learn to map their input to latent space, sample from the latent space, and use that sample to generate a novel instance of the data. VAEs are trained to compress and decompress data, with latent space representations consisting of probabilistic distributions characterized by a mean and variance. The VAE code is then generated by sampling the probability distribution, and that sample is then decoded. Sampling of the distributions in the latent space is what confers the VAE its generative, rather than discriminative, properties [32]. One of the greatest challenges in this project was how one could evaluate the realism and potential viability of a genome vector. Using a VAE permits one to evaluate the similarity of the output to the original, uncorrupted input. Note that methods such

as the Fréchet Inception Distance [33] require prohibitively large sample sizes for the limited dataset available here.

The encoder and decoder of our trained model, DeepGenomeVector, consist of three layers each, with a latent representation size of 100 neurons (Figure 1), as this architecture was determined to yield superior model performance during the training phase (see Methods). DeepGenomeVector was trained to optimize (i.e. minimize) the sum of binary cross-entropy (BCE) loss and Kullback-Leibler divergence (KLD). Quantitative performance during development was evaluated using the F1 score, such that the output of the VAE was compared against the original, uncorrupted genome vector from which the corrupted input to the model was derived.

Learning curves showing loss and F1 score as evaluation metrics show that the model was able to learn principles underlying genome composition (Figure 2A,B).

Quantitative evaluation of DeepGenomeVector's performance

In a scenario where generated outputs cannot be intuitively assessed by a human (like here, where our output is a list of thousands of gene identifiers), it is important to quantitatively evaluate whether the generated products have similarity to real ones. To gain insight into the model's tradeoff between sensitivity and specificity, we plotted an ROC curve (Figure 2C). A micro-average ROC-AUC score of 0.98 suggests the DeepGenomeVector model has high discriminative ability. To obtain a qualitative measure of performance, we evaluated how similar genomes generated from corrupted inputs were to the original, uncorrupted genome from which they derived. Model performance was evaluated using a test set of 27,900 corrupted genomes derived from 279 real genomes (~10% of all genomes). The DeepGenomeVector model obtained F1 scores ranging from 0.99-0.39, with a median of 0.83 (Figure 2D, Supplemental Figure 2). DeepGenomeVector's F1 score of 0.83 on the test set substantially outperformed those of five baseline models (Table 1). Adjusting the degree of corruption (i.e. the number of modules input to the model) substantially affected model performance on the training dataset (Supplemental Table 1). We further evaluated the fidelity of the model with respect to the presence of user-input genes in the generated genome (Supplemental Figure 3). In 85% of cases, 90% or more of the original input genes were also present in the output generated genome.

High-throughput, qualitative evaluation of DeepGenomeVector's performance

As quantitative evaluation metrics are limited in their ability to evaluate the novelty, ecological realism, and diversity of generated genome vectors, we next sought to gain a deeper qualitative understanding of how the real vs. generated genomes compare. To do so, we used DeepGenomeVector to generate a set of 100 genome vectors that could be used for high-throughput quality assessment of generated genome vector novelty, realism, and diversity.

First, we investigated whether generated genome vectors were indeed different from / novel compared to the genome vectors upon which the model was trained. We calculated the minimum Hamming distance from each generated genome vector to its closest neighbour in the training set (Figure 3A). For context, we repeated the same exercise comparing test set genome vectors against training set genome vectors. The distribution of minimum Hamming distances

for generated genome vectors was greater than that for test set genome vectors. This suggests that generated genome vectors are more novel compared to those in the training set than are real genome vectors from the test set.

Next, we performed a high-level evaluation of the realism of generated genome vectors. We calculated and plotted the Jaccard similarity between test genomes and a set of 100 generated genomes, using their Hamming distance, which is the number of bits that differ between two vectors (Figure 3B). While the distribution of generated genomes does appear to be slightly shifted from that of the real genomes, there is nonetheless a significant overlap in the distribution of real and generated genomes. Also noteworthy is that the distance between any given generated genome to its nearest real genome neighbour is smaller than the distance between the two most distantly related real genomes. The fact that we do not see two distinct groups of generated vs real genome vectors supports that the generated genome vectors are qualitatively similar to real ones.

Another way of estimating the similarity of the generated genome vectors to real ones is to consider how similar the lifestyle of organisms they could encode for would be. This is arguably a richer comparison metric than one based on pure distance, as distance does not capture which genes are different and what the significance of such differences may be. Therefore to give a qualitative perspective on whether the composition of generated genome vectors is realistic, and whether the genome vectors being generated are diverse and span multiple phyla across the bacterial tree of life, we placed the real and generated genomes in a parsimony-based dendrogram, built from a character matrix of gene presence/absence. We hypothesized that all the generated genomes clustered in a single branch of the dendrogram, while the real genomes clustered within another, it would suggest that the composition of genes was seriously flawed and unrealistic. Conversely, if the generated genome vectors cluster within the real ones, it supports that they are qualitatively similar. The dendrogram shows that real and generated genome vectors are both distributed across the dendrogram, with generated genomes often bounded within well-defined clades of real genomes from a taxonomic group (e.g.: Alphaproteobacteria) (Figure 3C). This suggests that the composition of generated genome vectors is not deeply aberrant from that of real genomes and that DeepGenomeVector is able to generate genome vectors representative of organisms from across the bacterial tree of life (albeit skewed towards certain groups, likely reflective of the data imbalance in the training set and in nature itself).

Functional profile of an AI-designed genome vector

Metrics such as F1 score give an estimate of whether the general gene content of a generated genome vector is correct. Missing, however, is information on critical parameters such as whether metabolic pathways are complete, whether they are properly linked together, whether they encode plausible networks / lifestyles, and whether core cellular machinery is present. To gain insight into these questions, we performed an in-depth functional analysis of an example AI-generated genome vector, dubbed “g13” (generated #13), using techniques common to genome-resolved metagenomics [34–37]. This genome vector was selected quasi-randomly, with the sole criterion being that its position in the PCA ordination in Figure 3B (see red arrow) overlap with the distribution of real genomes.

To qualitatively address whether metabolic networks are complete and properly linked together, we created a metabolic reconstruction diagram depicting genome vector g13's central carbohydrate metabolism, energy metabolism, and the 10 original input modules to the VAE and how they connect to the former (Figure 4; Supplemental Data 1). This is a small, meaningful subset of the full metabolic network that can be assessed by a human and correlates with one of the study goals which is to design genomes encoding user-defined functional properties. Of 30 modules, only two are partial: guanine ribonucleotide biosynthesis (missing one gene) and anaerobic photosynthesis (missing light harvesting complexes 1 and 2, and cytochrome C2). This is striking when considering, for example, that all enzymes required for a complex process such as oxidative phosphorylation (n=41) are present. The genome does lack the ability to synthesize ubiquinol, however, which is a component of the electron transport chain. Remarkably, modules were appropriately connected together. For example, the original input module C1-unit interconversion was connected to the genomes's central carbohydrate and energy metabolism via the PRPP => IMP and guanine ribonucleotide biosynthesis that the model encoded.

We next sought to gain insight into the ecological plausibility and theoretical viability of an organism hypothetically encoded for by genome vector g13. Generated genome g13 encodes 2,015 genes, including 50/51 universal bacterial single-copy genes often used to assess bacterial genome completeness in metagenomics studies [38] (missing aspartyl tRNA synthetase). Core cellular machinery was near complete. For example, all genes required for an RNA degradosome were all present, while 9/10 required for the RecFOR homologous recombination pathway were encoded. The genome vector encodes genes that would allow an organism to utilize compounds such as acetate and trans-cinnamate as carbon and energy sources. The Calvin cycle is also encoded, suggestive of a capacity for carbon fixation. The genome vector encodes glycolysis, the Entner-Doudoroff pathway, oxidative decarboxylation of pyruvate, the TCA cycle (forward direction), and oxidative phosphorylation, and would thus theoretically support aerobic respiration. The genetic potential to generate energy from inorganic compounds is present via both the assimilatory sulfate reduction and nitrogen fixation pathways. Anaerobic photosynthesis is partially encoded in g13. The lifestyle encoded by genome vector g13 is reminiscent of a bacterium capable of both heterotrophic and lithoautotroph growth, and would theoretically display a high metabolic diversity.

From this functional analysis, we conclude that AI generated genome vector 13 has metabolic networks that are largely complete and properly connected together, and encodes a metabolic potential that is ecologically plausible.

Error analysis of DeepGenomeVector

To gain insight into where the model over- vs under- performed, we examined biological and data-related factors that could potentially influence model performance. From a biological perspective, we queried whether the model had differential performance based on taxonomic group (approximately phylum) and lifestyle (metabolic functions). We found that both phylum and metabolic process affected model performance (Kruskal-Wallis test, $p = 0$ and $p = 1.01e^{-154}$, respectively), with Chlorobi test genomes and photosynthetic genes having the highest median F1 scores (Figure 5A,B). Notably, genes pertaining to processes that are often acquired via

horizontal gene transfer, such as drug resistance, tended to score poorly, as one might expect from modules not inherently tied to any given genome composition or lifestyle.

We next assessed whether DeepGenomeVector's performance may be mechanistically correlated with properties of the data upon which was trained. We found that taxonomic composition of the training set at the genus level affected performance on the test set, but only when none to one same-genus representatives were present in the training set (ANOVA + Tukey test, p=5.98e-16). Similarly, genes that were rare in the training set were more likely to be mis-assigned during model evaluation (linear regression analysis, p=0.0). In both of these cases, it is likely that composition of the training set reflects human sampling bias in genome coverage as well as naturally occuring discrepancies in the evolutionary diversity of certain taxonomic groups / ubiquity of certain genes and lifestyles across the bacterial domain. Post-hoc analysis of the model outputs suggests that DeepGenomeVector has learned a strategy whereby it uses more genes and complete genetic modules per genome vector than a typical bacterial genome encodes (Supplemental Figure 4), despite the fact that it was provided with no *a priori* information about which genes function together in any given module.

DISCUSSION

In this work, we presented a proof-of-concept that ML algorithms can learn the basic genetic principles underlying genome composition. In-depth functional analysis of a generated genome vector suggests that it is near-complete and encodes largely intact pathways that are interconnected. When evaluating whether the model was able to generate genome vectors similar in nature to those used as input seeds, quantitative evaluation metrics support strong model performance, with DeepGenomeVector achieving an AUC score of 0.98 and an F1 score of 0.83 on the test set.

While the model presented here serves as a proof of concept that genomic gene repertoires can be modeled by ML algorithms, there are limitations to this study. From the biological perspective, a major limitation is that our features only included genes annotated with a KO number from the KEGG dataset. This means our study excludes many genes that have evolved in nature but for which humans have not yet studied and characterized their function. Genes of unknown function can play essential roles in cellular biology; for example, when [4] engineered a minimal bacterial genome, they found that the minimal viable gene set included 149 genes of unknown function that were absolutely essential for producing a viable organism. Another important contextual consideration for this study is that genomes are more than just "bags of genes". They include intergenic spacing, promoters, transcription factor binding sites, and more. Furthermore, genes often have organization structure within genomes, e.g., into operons. Future extensions to this work could consist of including spatial features into genome vectors. Finally, if one want to extend such work to generating genomes rather than genome vectors, a major piece of the problem that would need to be addressed is figuring out how to convert binary vectors of 0's and 1's to gene sequences. The field of synthetic biology is ripe with challenges that ML may help to address, with the potential to provide both valuable scientific and engineering insights.

From an ML perspective, one area for future improvement to the DeepGenomeVector approach could be to impose sparsity constraints that reduce the metabolic diversity of output genomes. Quantitative analysis showed that generated genomes encoded more modules than would be expected from real bacterial genomes. In line with this, the metabolic analysis of generated genome g13 was suggestive of a highly metabolically diverse bacterium, for example as demonstrated by the capacity for chemoautolithotrophy using both nitrogen and sulfate as energy sources. From an evolutionary perspective, a sparsity constraint would be akin to imposing the microevolutionary forces in nature that shape bacterial genome architecture [39]. For example, it has been theorized that resource specialization is driven by competition through natural selection [40,41]. The realities of living in real environments, such as the intense competition that exists, are not reflected in our model.

Another tantalizing possibility would be to employ GANs or even language-like models that generate sequences for the purposes of generating artificial genome blueprints. A major barrier to doing so, however, is the limited amount of data available. Here, we circumvented this issue by creating corrupted versions of genomes for input to the model, which allowed us to train and evaluate the VAE. Critically, using a VAE allowed us to compare the original, uncorrupted genome vector against the genome vector generated from the corrupted input. This provided a rough evaluation of the quality of the generated product, albeit an imperfect one as many combinations of genes can be viable but may not be represented in the genome set available. Evaluation methods that are used to evaluate the realism of outputs of generative models in other fields, such as using the Frechet Inception Distance [33] for evaluating photo realism, require prohibitive amounts of data for this application. Methods for evaluating genome realism that are minimally data-hungry would broaden the scope of future approaches (e.g.: GANs) to genome vector design. Similarly, ever increasing dataset sizes will provide greater flexibility and breadth to ML applications in microbial genomics.

DeepGenomeVector serves as proof-of-concept that an ML algorithm can learn and model the basic tenets of how genomes, and thus life forms, are encoded. The development of such technologies may eventually lead to their use in synthetic biology, for example via their incorporation into synthetic genomic CAD software [42]. From a more philosophical perspective, this work highlights the fascinating possibility that AI agents may one day be able to design viable carbon-based lifeforms.

MATERIALS AND METHODS

Genome selection, annotation, and representation

Genome annotations were acquired from the KEGG knowledge base [29–31] which can also be used to link sequence information to high-level cellular functions (e.g.: pathways). The KEGG genomes database includes a set of 6542 complete reference and representative genomes from NCBI that have been annotated using KEGG's internal pipeline (accessed May 14, 2020, includes Eukaryotic, Archaeal, and Bacterial genomes). The KEGG bacterial genomes are strongly phylogenetically skewed. As such, we performed thinning to reduce the representation of over-represented taxa. For each genome, we used the taxid to recover the NCBI taxonomic identification. First, any genome not identified to the phylum level was discarded. Next we randomly selected only one genome to represent each species, and no more than five species to represent each genus. Per phylum we allowed for up to 50 genomes unclassified at the species level. Finally, we removed genomes with incomplete KEGG annotations, fewer than 500 annotated genes / genome (presumed symbionts with highly atypical lifestyles compared to the "average" bacterium), and genomes encoding fewer than 10 KEGG modules (same rational re: symbionts). The final dataset consisted of 2,584 bacterial genomes spanning 46 major lineages (approximately phyla) and 9,863 features. For greater detail into the rationale behind these decisions, see the Supplemental Methods.

Genomes can be thought of as blueprints for life that either encode a gene or don't. Thus we represented each genome as a vector of 1's and 0's representing whether, for each gene in the dataset's 9,863 genes, a given gene is encoded.

Train-test split and corruption process

Genomes were split into a training set (~90%) and testing set (~10%) prior to undergoing corruption. As the goal of this study was to generate genome vectors representative of bacteria from across the bacterial domain, we included representatives from each bacterial phylum into each of the training and test sets where possible. This is important for ensuring the training and test sets are from the same distribution. A VAE trained only on generating genome vectors from phylum A will never generate high-quality genome vectors from a distantly related phylum B, much as a VAE trained to generate images of human faces (phylum Chordata) will not generate images of creatures from other eukaryotic phyla, such as worms (phylum Annelida), scorpions (phylum Arthropoda), or jellyfish (phylum Cnidaria). For major lineages (approximately phyla) represented in our dataset by ≥10 genomes, 88% were assigned to the training set and 12% to the test set. For those represented by 2 - 9 genomes, 50% were assigned to each of the training and test sets. Finally, for lineages represented by a single genome, that genome was assigned to the training set.

Corrupted genomes were used as input to the VAE. Corruptions were achieved by turning on the genes from 10 randomly selected KEGG modules that naturally co-occur within real genomes, and turning all other bits off. This was done 100X per genome in the training set and 100X per genome in the test set. The extent to which a genome was corrupted had a significant impact on the quality of generated genomes, as measured by the training set F1 score (Supplemental Table 1). We decided to use the genes from 10 modules as input in an

effort to minimize the number of genes a user would theoretically need to specify while simultaneously avoiding the rapid drop in performance that comes with using genes from fewer than ~10 modules.

Machine learning model

The DeepGenomeVector model was built in PyTorch v1.5.0 [43]**.** Hyperparameter tuning was implemented in Tune v0.8.2 [44] and performed using 10-fold cross-validation with AdamW as the optimizer [45,46]. Leaky ReLU activation functions [47] were applied to all layers except the output, for which a sigmoid activation was applied. Correspondingly, He initialization [48] was used for all layers except the last, for which Xavier initialization [49] was used. Sweeps over the following hyperparameters were conducted: number of layers (1, 2, 3, 4), batch size (32, 64, 128, 256), learning rate (log uniform selection between $1e^{-4}$ - $1e^{-1}$), and weight decay (log uniform selection between $1e^{-5}$ - $1e^{-2}$). While the number of layers had a strong impact on model performance, the effect of other hyperparameters was negligible. The final model used 3 layers for the encoder and decoder each, batch size 128, learning rate 0.001, and weight decay 0.1. At every iteration, loss was calculated by summing together binary cross-entropy (BCE) loss and Kullback-Leibler divergence (KLD). Training was performed over 10 epochs. When the model is used to generate a genome, the probabilities for each bit being on are converted to 0 or 1 using a 0.5 replacement threshold.

Parsimony-based dendrogram

The 279 test genome vectors, 100 generated genomes, and an archaeal genome were used to create a parsimony-based dendrogram with the Wagner method [50,51]. This was achieved using the mix program in Phylip v3.695 [52]. The input order of species was randomized 10 times using seed=99 and using the archaeal genome as an outgroup root. All other settings were set to default. The dendrogram was visualized using iTOL v5.7 [53].

Metabolic reconstruction of a generated genome

Identification of metabolic pathways was achieved using the KEGG Mapper “reconstruct” tool, available at https://www.genome.jp/kegg/mapper/reconstruct.html [54,55].

CODE AVAILABILITY STATEMENT

Code for this project is available on Github at https://github.com/natasha-dudek/DeepGenomeVector.

ACKNOWLEDGMENTS AND FUNDING SOURCES

This work was supported by NSERC Discovery and Canada-CIFAR AI Chair funding awarded to Doina Precup.

WORKS CITED

1. Koonin E, Galperin MY. Sequence — Evolution — Function: Computational Approaches in Comparative Genomics. Springer Science & Business Media; 2002.

2. Betts HC, Puttick MN, Clark JW, Williams TA, Donoghue PCJ, Pisani D. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. Nat Ecol Evol.

3. Morowitz HJ. The completeness of molecular biology. Isr J Med Sci. 1984;20: 750–753.

4. Hutchison CA 3rd, Chuang R-Y, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, et al. Design and synthesis of a minimal bacterial genome. Science. 2016;351: aad6253.

5. Pennisi E. Synthetic genome brings new life to bacterium. American Association for the Advancement of Science; 2010.

6. Reiman D, Dai Y. Using conditional generative adversarial networks to boost the performance of machine learning in microbiome datasets. bioRxiv. 2020. doi:10.1101/2020.05.18.102814

7. Mulenga M, Kareem SA, Sabri AQM, Seera M, Govind S, Samudi C, et al. Feature extension of gut microbiome data for deep neural network-based colorectal cancer classification. IEEE Access.

8. Douglas GM, Beiko RG, Langille MGI. Predicting the Functional Potential of the Microbiome from Marker Genes Using PICRUSt. Methods Mol Biol. 2018;1849: 169–177.

9. Radivojević T, Costello Z, Workman K, Garcia Martin H. A machine learning Automated Recommendation Tool for synthetic biology. Nat Commun. 2020;11: 4879.

10. Long B, Fischer B, Zeng Y, Amerigian Z, Li Q, Bryant H, et al. Machine learning-informed and synthetic biology-enabled semi-continuous algal cultivation to unleash renewable fuel productivity. Nat Commun. 2022;13: 541.

11. Zhang J, Petersen SD, Radivojevic T, Ramirez A, Pérez-Manríquez A, Abeliuk E, et al. Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism. Nat Commun. 2020;11: 4880.

12. Carbonell P, Jervis AJ, Robinson CJ, Yan C, Dunstan M, Swainston N, et al. An automated Design-Build-Test-Learn pipeline for enhanced microbial production of fine chemicals. Commun Biol. 2018;1: 66.

13. Jervis AJ, Carbonell P, Taylor S, Sung R, Dunstan MS, Robinson CJ, et al. SelProm: A Queryable and Predictive Expression Vector Selection Tool for Escherichia coli. ACS Synth Biol. 2019;8: 1478–1483.

14. Meng H, Wang J, Xiong Z, Xu F, Zhao G, Wang Y. Quantitative design of regulatory elements based on high-precision strength prediction using artificial neural network. PLoS One. 2013;8: e60288.

15. Groher A-C, Jager S, Schneider C, Groher F, Hamacher K, Suess B. Tuning the Performance of Synthetic Riboswitches using Machine Learning. ACS Synth Biol. 2019;8: 34–44.

16. Li W, Huang J, Shi Z, Han W, Lü T, Lin Y, et al. Machine learning enabled prediction and process optimization of VFA production from riboflavin-mediated sludge fermentation. Front Environ Sci Eng China. 2023;17: 135.

17. Packiam KAR, Ooi CW, Li F, Mei S, Tey BT, Ong HF, et al. PERISCOPE-Opt: Machine learning-based prediction of optimal fermentation conditions and yields of recombinant periplasmic

protein expressed in Escherichia coli. Comput Struct Biotechnol J. 2022;20: 2909–2920.

18. Roell GW, Sathish A, Wan N, Cheng Q, Wen Z, Tang YJ, et al. A comparative evaluation of machine learning algorithms for predicting syngas fermentation outcomes. Biochem Eng J. 2022. doi:10.1016/j.bej.2022.108578

19. Dean SN, Walper SA. Variational Autoencoder for Generation of Antimicrobial Peptides. ACS Omega. 2020;5: 20746–20754.

20. Kilimci ZH, Yalcin M. ACP-ESM: A novel framework for classification of anticancer peptides using protein-oriented transformer approach. arXiv [q-bio.BM]. 2024. Available: http://arxiv.org/abs/2401.02124

21. Wang X-F, Tang J-Y, Liang H, Sun J, Dorje S, Peng B, et al. ProT-Diff: A Modularized and Efficient Approach to De Novo Generation of Antimicrobial Peptide Sequences through Integration of Protein Language Model and Diffusion Model. bioRxiv. 2024; 2024–2002.

22. Chen S, Kim HU. Designing Novel Functional Peptides by Manipulating a Temperature in the Softmax Function Coupled with Variational Autoencoder. 2019 IEEE International Conference on Big Data (Big Data). IEEE; 2019. pp. 6010–6012.

23. Chen S, Lin T, Basu R, Ritchey J, Wang S, Luo Y, et al. Design of target specific peptide inhibitors using generative deep learning and molecular dynamics simulations. Nat Commun. 2024;15. doi:10.1038/s41467-024-45766-2

24. Mantena S, Pillai PP, Petros BA, Welch NL, Myhrvold C, Sabeti PC, et al. Model-directed generation of CRISPR-Cas13a guide RNAs designs artificial sequences that improve nucleic acid detection. bioRxiv. 2023. doi:10.1101/2023.09.20.557569

25. Shanehsazzadeh A, McPartlon M, Kasun G, Steiger AK, Sutton JM, Yassine E, et al. Unlocking de novo antibody design with generative artificial intelligence. bioRxiv. 2024. p. 2023.01.08.523187. doi:10.1101/2023.01.08.523187

26. Luo S, Su Y, Peng X, Wang S, Peng J, Ma J. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. Adv Neural Inf Process Syst. 2022;35: 9754–9767.

27. Lim YW, Adler AS, Johnson DS. Predicting antibody binders and generating synthetic antibodies using deep learning. MAbs. Taylor & Francis; 2022. p. 2069075.

28. Seo E, Choi Y-N, Shin YR, Kim D, Lee JW. Design of synthetic promoters for cyanobacteria with generative deep-learning model. Nucleic Acids Res. 2023;51: 7071–7082.

29. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28: 27–30.

30. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44: D457–62.

31. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45: D353–D361.

32. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv preprint arXiv:1312 6114.

33. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Adv Neural Inf Process Syst.

34. Dudek NK, Sun CL, Burstein D, Kantor RS, Aliaga Goltsman DS, Bik EM, et al. Novel Microbial Diversity and Functional Potential in the Marine Mammal Oral Microbiome. Curr Biol. 2017;27: 3752–3762.e6.

35. Valentin-Alvarado LE, Fakra SC, Probst AJ, Giska JR, Jaffe AL, Oltrogge LM, et al. Autotrophic biofilms sustained by deeply sourced groundwater host diverse bacteria implicated in sulfur and hydrogen metabolism. Microbiome. 2024;12: 15.

36. Qiu X, Zhang Y, Hong H. Classification of acetic acid bacteria and their acid resistant mechanism. AMB Express. 2021;11: 29.

37. Zhou Z, Tran PQ, Breister AM, Liu Y, Kieft K, Cowley ES, et al. METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. Microbiome. 2022;10: 33.

38. Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P. Prediction of effective genome size in metagenomic samples. Genome Biol. 2007;8: R10.

39. Bobay L-M, Ochman H. The Evolution of Bacterial Genome Architecture. Front Genet. 2017;8: 72.

40. Fiegna F, Scheuerl T, Moreno-Letelier A, Bell T, Barraclough TG. Saturating effects of species diversity on life-history evolution in bacteria. Proc Biol Sci. 2015;282. doi:10.1098/rspb.2015.1794

41. Lawrence D, Fiegna F, Behrends V, Bundy JG, Phillimore AB, Bell T, et al. Species interactions alter evolutionary responses to a novel environment. PLoS Biol. 2012;10: e1001330.

42. Richardson SM, Mitchell LA, Stracquadanio G, Yang K, Dymond JS, DiCarlo JE, et al. Design of a synthetic yeast genome. Science. 2017;355: 1040–1044.

43. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: An imperative style, high-performance deep learning library. Adv Neural Inf Process Syst.

44. Liaw R, Liang E, Nishihara R, Moritz P, Gonzalez JE, Stoica I. Tune: A Research Platform for Distributed Model Selection and Training. arXiv [cs.LG]. 2018. Available: http://arxiv.org/abs/1807.05118

45. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412 6980.

46. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv preprint arXiv:1711 05101.

47. Maas AL, Hannun AY, Ng AY, Others. Rectifier nonlinearities improve neural network acoustic models. Proc icml. Atlanta, GA; 2013. p. 3.

48. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. 2015 IEEE International Conference on Computer Vision (ICCV). IEEE; 2015. doi:10.1109/iccv.2015.123

49. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Teh YW, Titterington M, editors. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Chia Laguna Resort, Sardinia, Italy: PMLR; 13--15 May 2010. pp. 249–256.

50. Chang MA, Dayhoff MO, Eck RV, Sochard MR. Atlas of protein sequence and structure. 1965.

51. Kluge AG, Farris JS. Quantitative Phyletics and the Evolution of Anurans. Syst Biol. 1969;18: 1–32.

52. Felsenstein J. PHYLIP (phylogeny inference package), version 3.5 c. Joseph Felsenstein.; 1993.

53. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47: W256–W259.

54. Kanehisa M, Sato Y. KEGG Mapper for inferring cellular functions from protein sequences. Protein Sci. 2020;29: 28–35.

55. Kanehisa M, Sato Y, Kawashima M. KEGG mapping tools for uncovering hidden features in biological data. Protein Sci. 2022;31: 47–53.

FIGURES AND TABLES

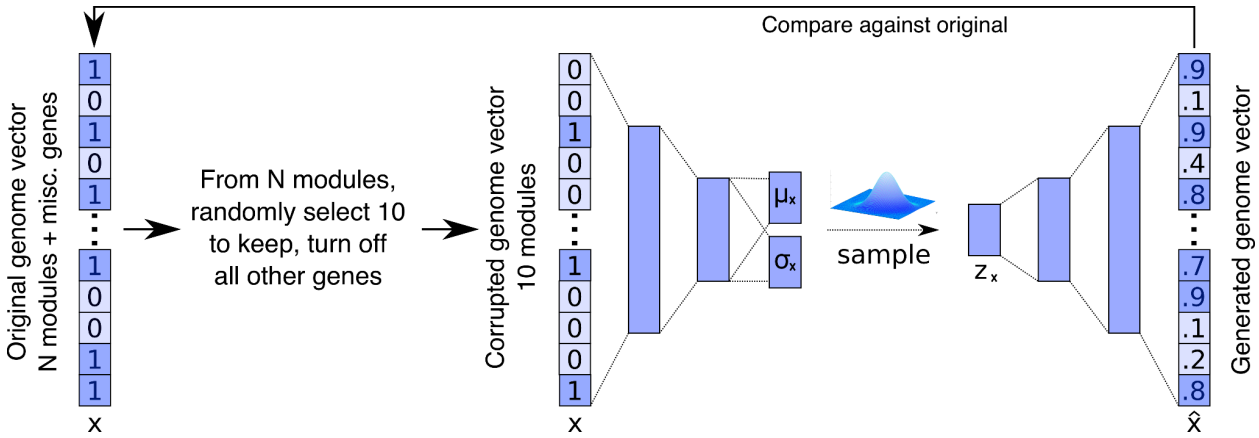


Figure 1: Overview of DeepGenomeVector. Data corruption process and the DeepGenomeVector architecture. Genomes are represented as binary vectors (genome vectors) that denote which genes are encoded. To create input for the model, 100 corrupted versions of each genome vector are created by randomly selecting 10 functional modules it encodes, and creating a new corrupted genome vector that only encodes the genes from those 10 modules. The input to the model is a corrupted genome vector (x), where the presence (1) / absence (0) of genes is encoded. The model consists of three encoder layers (9,863, 500, 250, and 2 x 100 neurons) and three decoder layers (100, 250, 500, 9,863 neurons). The latent representation (z_x) is sampled from the multivariate Gaussian distribution parameterized by μ_x and σ_x. The output from the model is a probabilistic reconstruction of the corrupted input genome vector ($\hat{x}$). Per gene probabilities output by the model are binarized using a threshold of 0.5. The performance of the VAE is measured by comparing the generated genome vector against the original, uncorrupted one from which it was derived.

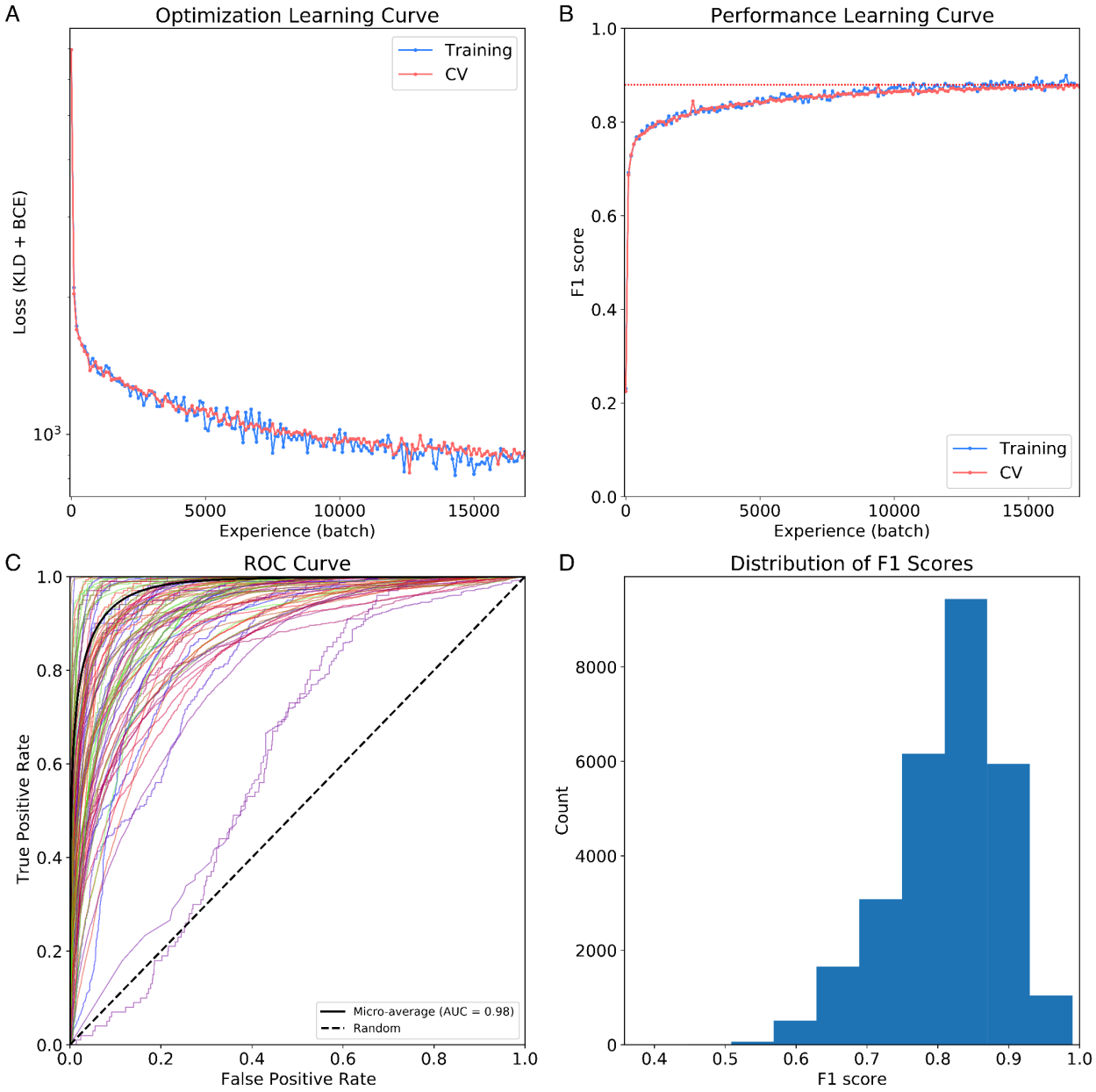


Figure 2: DeepGenomeVector learns and generates genome vectors that are quantitatively similar to real ones. (A) Optimization learning curve, with batch loss (KLD + BCE) on the y-axis. Training was performed over 10 epochs. **(B)** Performance learning curve obtained using the test set, with F1 score on the y-axis. The maximum cross-validation fold F1 score of 0.88 is indicated with a red dotted line. **(C)** Receiver operating characteristic (ROC) curve obtained using the test set. Each coloured line represents one of 100 randomly selected genes (for clarity only a subset of the full 9,863 genes is visualized). The micro-average is plotted in solid black and corresponds to a micro-average AUC score of 0.98. **(D)** Histogram of F1 scores obtained using the test set.

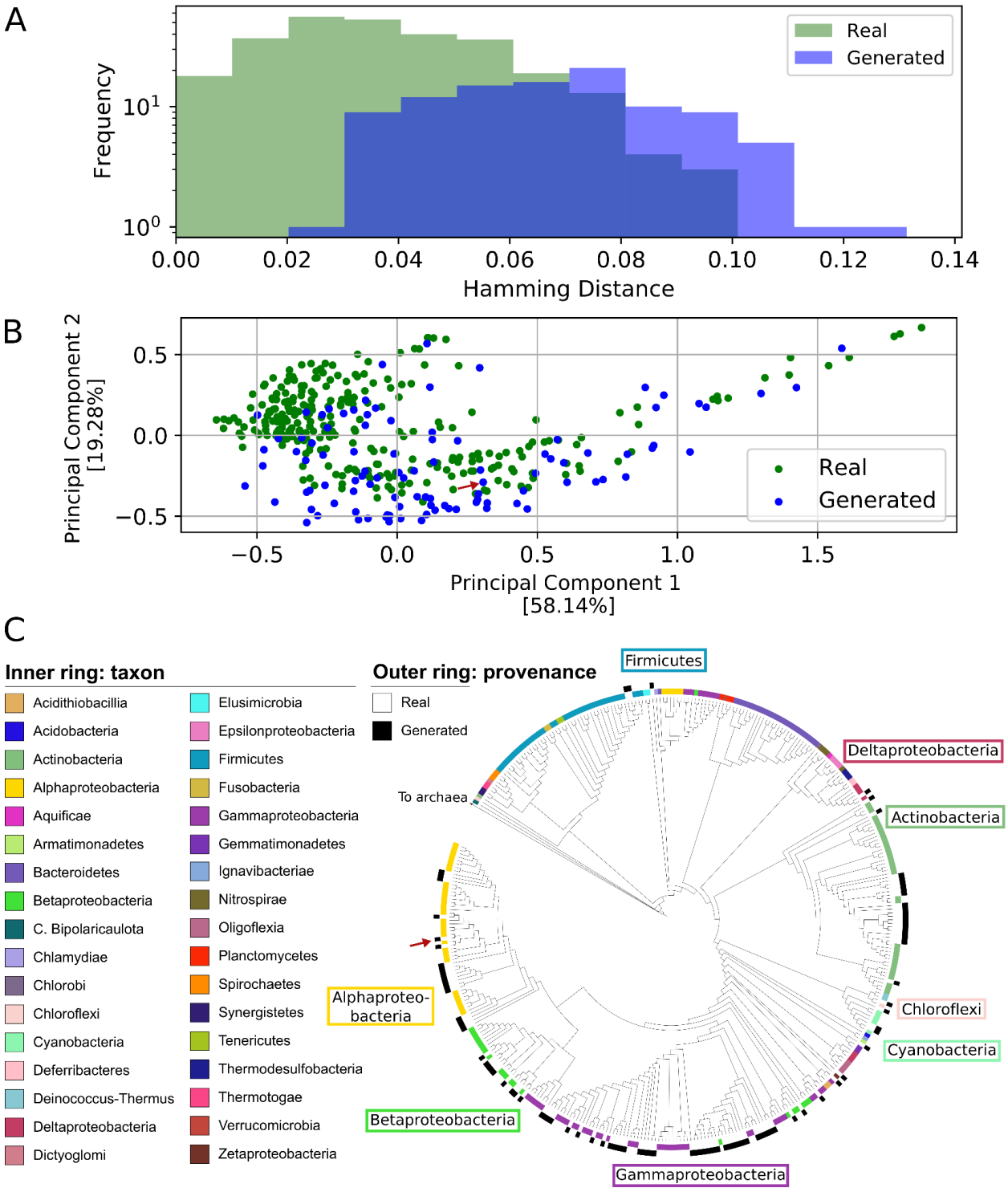

A
Real
Generated
Frequency
10^1
10^0
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Hamming Distance
B
Principal Component 2
[19.28%]
0.5
0.0
−0.5
Real
Generated
−0.5
0.0
0.5
1.0
1.5
Principal Component 1
[58.14%]
C
Inner ring: taxon
Acidithiobacillia
Acidobacteria
Actinobacteria
Alphaproteobacteria
Aquificae
Armatimonadetes
Bacteroidetes
Betaproteobacteria
C. Bipolaricaulota
Chlamydiae
Chlorobi
Chloroflexi
Cyanobacteria
Deferribacteres
Deinococcus-Thermus
Deltaproteobacteria
Dictyoglomi
Elusimicrobia
Epsilonproteobacteria
Firmicutes
Fusobacteria
Gammaproteobacteria
Gemmatimonadetes
Ignavibacteriae
Nitrospirae
Oligoflexia
Planctomycetes
Spirochaetes
Synergistetes
Tenericutes
Thermodesulfobacteria
Thermotogae
Verrucomicrobia
Zetaproteobacteria
Outer ring: provenance
Real
Generated
Firmicutes
Deltaproteobacteria
Actinobacteria
To archaea
Alphaproteo-
bacteria
Chloroflexi
Cyanobacteria
Betaproteobacteria
Gammaproteobacteria

Figure 3: DeepGenomeVector's generated genome vectors are qualitatively novel, realistic, and diverse**.** The 279 real genome vectors from the test set were compared against 100 generated genome vectors with randomly selected input modules. **(A)** Histogram showing the minimum Hamming distance of real genome vectors to the most similar training set genome vector vs generated genome vectors to the most similar training set genome vector. **(B)** Principal component analysis (PCA) Jaccard similarity ordination of real (green) vs generated (blue) genome vectors. Jaccard similarity was calculated using the Hamming distance metric. Axes are scaled in proportion to the variance explained by each principal component. **(C)** Parsimony-based dendrogram of real vs generated genome vectors. The dendrogram was generated from a character matrix where each gene's presence/absence in a genome vector is a character. The inner ring around the dendrogram is a colour-coded representation of genome vector phylum (real genome vectors only), while the outer ring indicates whether a genome vector was real (white) vs generated (black). Names of major lineages (approximately phyla) whose clades in the dendrogram neighbour or include generated genome vectors are indicated for reference. The generated genome vector selected for functional analysis (Figure 4) is denoted by a red arrow in panels B and C.

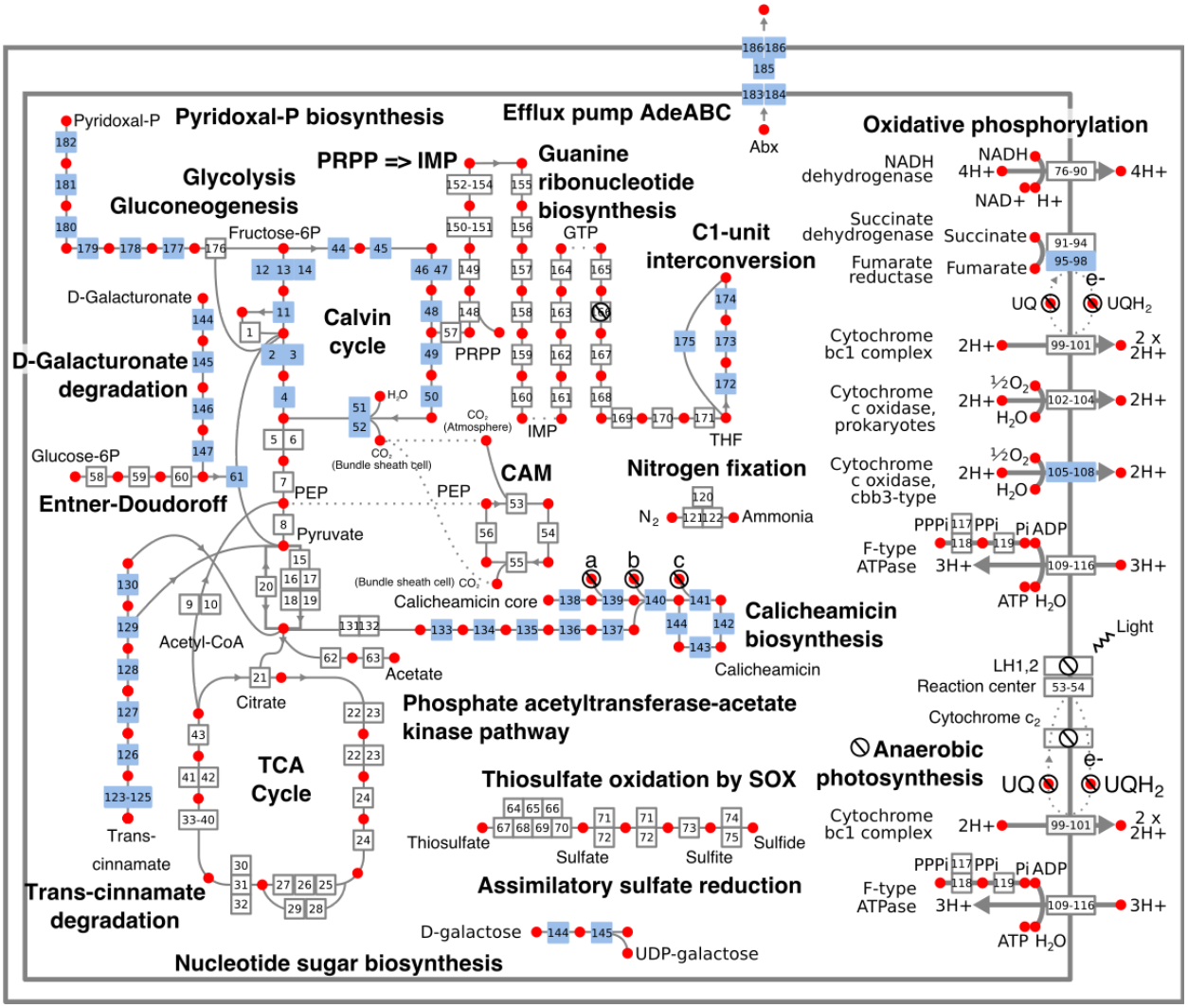


Figure 4: Cell diagram for a generated genome, depicting central carbohydrate metabolism, energy metabolism, and the 10 modules used as input. Boxes represent reactions catalyzed by genes encoded in the generated genome vector g13, while red dots represent key compounds involved in reactions. The 10 modules used as input to the model are colour-coded blue. Cross icons indicate genes that are not encoded or compounds whose synthesis is not encoded for. Metabolic pathways that are encoded by generated genome vector g13 but not directly related to central carbohydrate metabolism, energy metabolism, or an input module are not shown. Numbers in boxes allow for the identification of the reaction/gene via **Supplemental Data 1**.

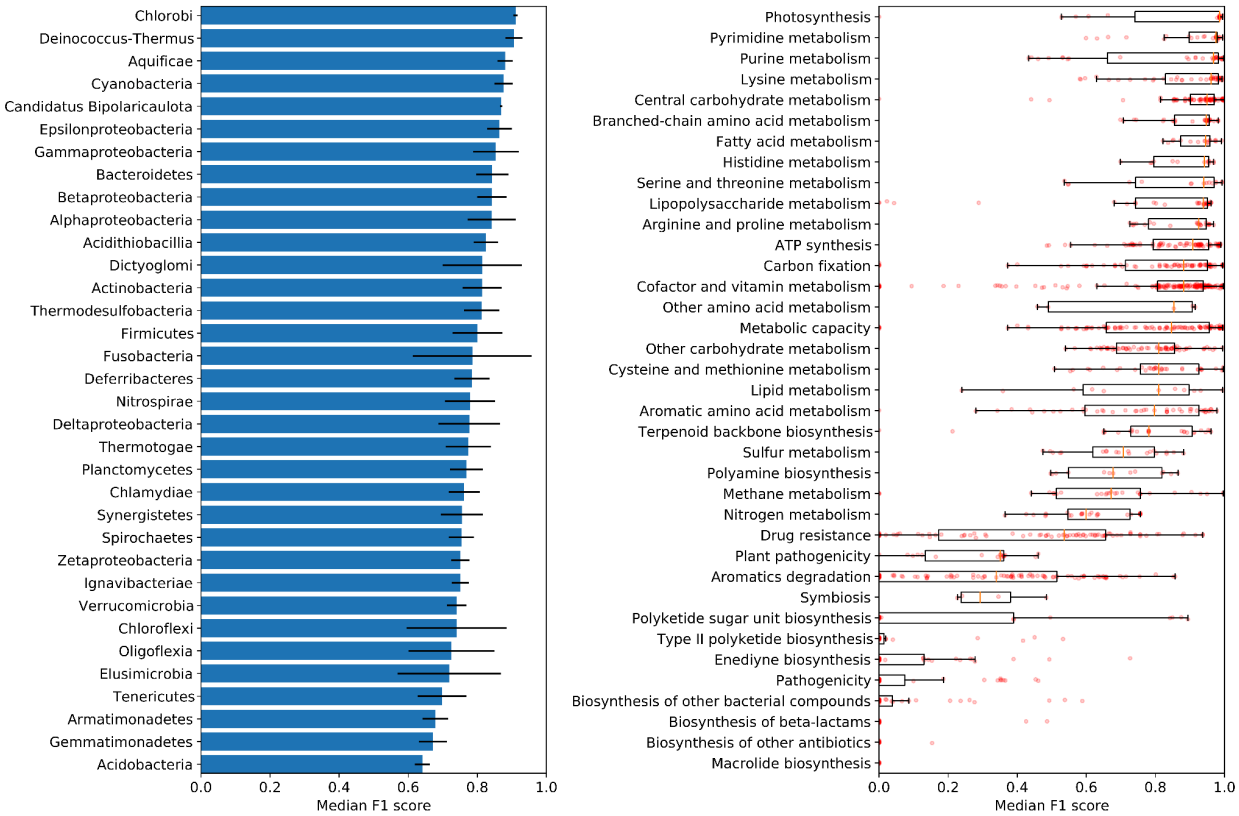


Figure 5: Biological covariates of DeepGenomeVector performance. (A) Model performance on genomes by taxonomic group (approximately phylum). Error bars represent median absolute deviation. **(B)** Model performance on genes by metabolic group. Individual data points represent modules. On the boxplots, the orange bars represent the median, boxes extend to the lower and upper quartile values, whiskers represent the range of the data, and data points beyond whiskers are outliers.

TABLES

	AUROC	F1	TN	FP	FN	TP
Baseline 1	0.52	0.19	68.49%	16.77%	11.39%	3.36%
Baseline 2	0.49	0.16	67.95%	17.3%	11.91%	2.83%
Baseline 3	0.50	0.23	42.54%	42.72%	7.32%	7.42%
Baseline 4	0.62	0.38	**83.89%**	**1.37%**	10.98%	3.76%
Baseline 5	0.76	0.52	70.15%	15.11%	4.34%	10.40%
DeepGenome Vector	**0.98**	**0.83**	82.92%	2.34%	**2.91%**	**11.83%**

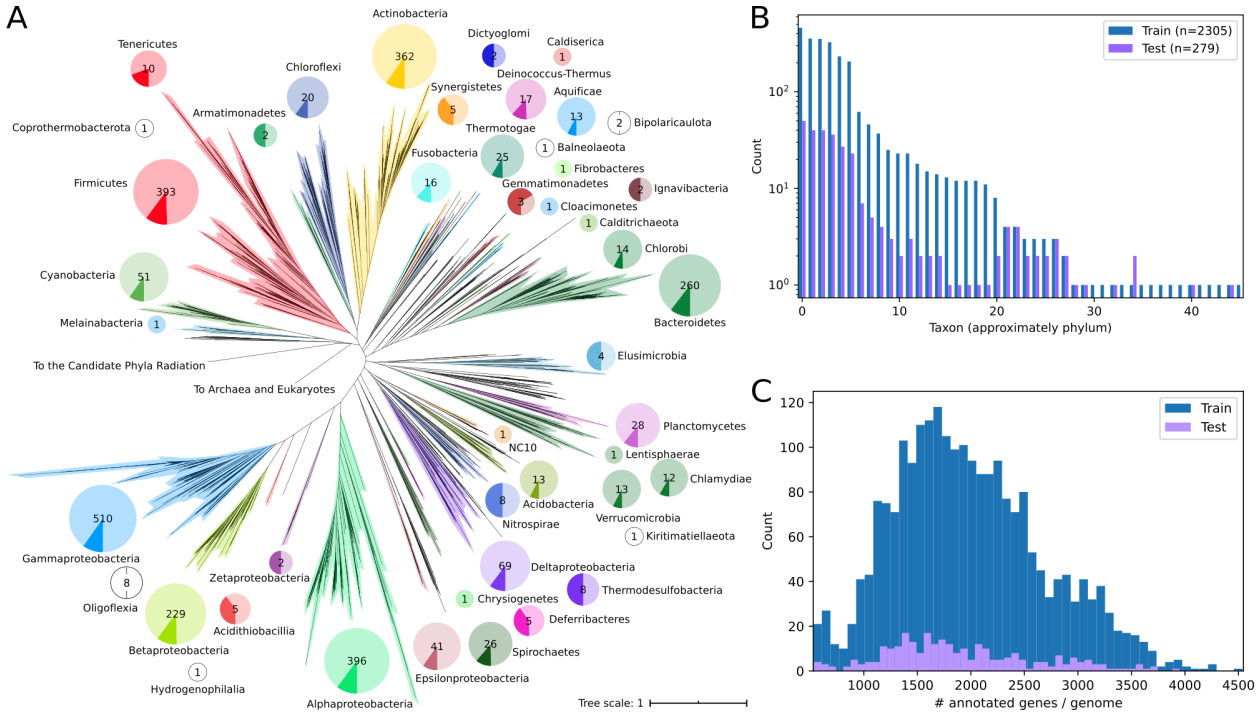
Table 1: DeepGenomeVector's performance on the test set compared to baseline models. AUROC and F1 scores are micro-averages. True negatives (TN), false positives (FP), false negatives (FN), and true positives (TP) are the percentage of each, respectively, across all genomes in the test dataset. For baselines 1 & 2, let n = a randomly selected number of genes encoded by a real genome from the training set. Baseline 1 consisted of randomly turning on n bits in a corrupted genome vector from the test set. Baseline 2 consisted of randomly turning on the n bits with the highest probability of being on across the entire training set. Baseline 3 consisted of making predictions using an untrained VAE model. Baseline 4 consisted of always predicting the smallest real bacterial genome vector in the training set (*Hoaglandella endobia*, Gammaproteobacteria). Baseline 5 consisted of always predicting the largest bacterial genome vector in the training set (*Paraburkholderia caribensis*, Betaproteobacteria).

Supplemental information for
"Towards AI-designed genomes using a variational autoencoder"

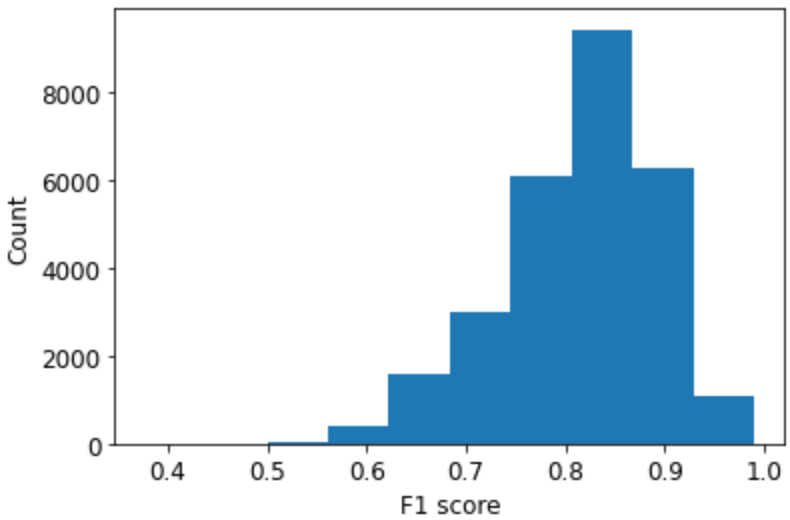
Dudek, N.K.[1,2], Precup, D.[1,2]

AUTHOR AFFILIATIONS

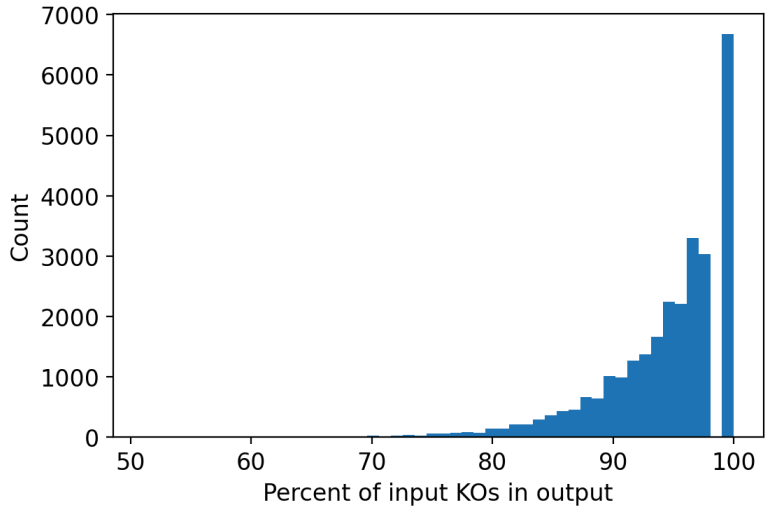
1. School of Computer Science; McGill University; Montreal, QC, H3A 0G4; Canada
2. Mila - Québec Artificial Intelligence Institute; Montreal, QC, H2S 3H1; Canada



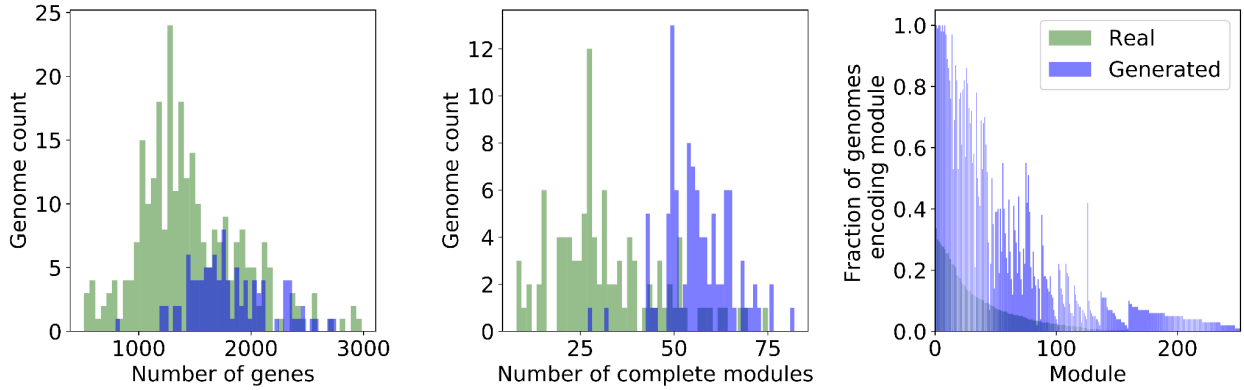
Supplemental Figure 1: Overview of the genome dataset used to develop DeepGenomeVector. (A) Phylogenetic distribution of genomes incorporated in the training and test sets. The bacterial tree of life is shown (modified from Hug *et al*., 2016), with the exclusion of the Candidate Phyla Radiation as no genomes from this supergroup were included. Major lineages (approximately phyla) from which genomes were included here are highlighted, with a colour-matching pie chart showing the portion of genomes from that lineage used in the training set (light) vs test set (dark). The total number of genomes from each lineage is indicated in the center of each pie chart. Major lineages not included in the Hug *et al*., 2016 tree are denoted by pie charts with no fill, and are placed in the proximity of the supergroup with which the given phylum affiliates. **(B)** The distribution of genomes per taxon (approximately phylum) in the training (blue) and test (purple) sets (see Methods). **(C)** Distribution of genes per genome. The number of genes annotated by KEGG per genome in the training (blue) and test (purple) sets are shown. Genomes with fewer than 500 annotated genes were not included in this study.



Supplemental Figure 2: Distribution of DeepGenomeVector F1 scores across test set genome vectors. F1 scores range from 0.38 - 0.99, with a median of 0.83.



Supplemental Figure 3: Frequency with which user-defined DeepGenomeVector inputs are present in the DeepGenomeVector output genome vector. The percentage of input genes present in the output genome vector for a given genome is shown on the x-axis and occurrence counts are on the y-axis.



Supplemental Figure 4: Distribution of gene and module counts for real vs. DeepGenomeVector generated genome vectors. (**Left**) Distribution of the number of genes per genome. (**Center**) Distribution of the number of complete modules per genome. (**Right**) For each module, the fraction of genomes that encodes it.

SUPPLEMENTAL TABLES

# modules input	F1 score	# genomes	# features
0	0.61	2332	9866
1	0.66	2330	9866
5	0.80	2325	9865
10	0.88	2305	9863
20	0.94	2200	9837
30	0.96	1955	9741
40	0.98	1580	9574

Supplemental Table 1: Relationship between the number of modules retained during the corruption process and DeepGenomeVector F1 score. Using the training dataset, the model was trained three times for each possible number of modules input. The best F1 score on any cross-validation fold across the three trials is reported. Because genomes with fewer than the threshold number of modules were discarded from the dataset during the corruption process, the number of genomes and features changes.

SUPPLEMENTAL METHODS

Genome selection, annotation, and representation

For a machine learning algorithm to be able to learn what is encoded by a genome, it is essential that genomes used for training be complete, high-quality genomes that are consistently annotated such that homologues (i.e. the same gene in different genomes) are identifiable with a high degree of confidence. To acquire such annotated genomes, we turned to the KEGG knowledge base (1–3), which can also be used to link sequence information to high-level cellular functions. The KEGG genomes database includes a set of 6542 complete reference and representative genomes from NCBI that have been annotated using KEGG's internal pipeline (accessed May 14, 2020, includes Eukaryotic, Archaeal, and Bacterial genomes). Where possible, KEGG predicted proteins are assigned not only a name (e.g.: hexokinase) but also a KEGG Orthology (KO) number (e.g.: K00844). This is important because it allows for consistent identification and comparison of orthologous genes between genomes because each gene has a unique KO number, if not a unique name (e.g.: K00844 may be referred to as glucose ATP phosphotransferase, hexokinase type IV glucokinase, hexokinase (phosphorylating), etc). KEGG annotations are also of high quality as each KO number is assigned to protein sequence data in KEGG for which a protein function has been experimentally validated.

The KEGG bacterial genomes are strongly phylogenetically skewed. As such, we performed thinning to reduce the representation of over-represented taxa. For each genome, we

used the taxid to recover the NCBI taxonomic identification. First, any genome not identified to the phylum level was discarded. Next we randomly selected only one genome to represent each species, and no more than five species to represent each genus. This worked for thinning well studied, well characterized lineages of bacteria, but there are many taxa for which this level of taxonomic classification is not available. For example, many candidate phyla do not robustly have defined classes, let alone genera or species. Thus per phylum we allowed for up to 50 genomes unclassified at the species level. Those lacking species classification were randomly selected to fill this quota. Thinning yielded 2,760 genomes, 250 of which were unclassified at the species level, and substantially helped even out the phylogenetic distribution of genomes, although it should be noted that the dataset is still phylogenetically skewed. This reflects that a) diversity is skewed across the tree of life (some lineages are more diverse than others), and b) genomic representation of diversity across the tree of life is skewed towards lineages that lend themselves to being cultured easily in the lab (Hug *et al*., 2016). Finally, we removed genomes with incomplete KEGG annotations, fewer than 500 annotated genes / genome (presumed symbionts with highly atypical lifestyles compared to the “average” bacterium), and genomes encoding fewer than 10 KEGG modules (same rational re: symbionts). The final dataset consisted of 2,584 bacterial genomes spanning 46 major lineages (approximately phyla) and 9,863 features.

WORKS CITED

1. M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

2. M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–62 (2016).

3. M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).