nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-021-00310-5

Check for updates

Expanding functional protein sequence spaces using generative adversarial networks

Donatas Repecka[1,6], Vykintas Jauniskis[1,2,6], Laurynas Karpus[1,6], Elzbieta Rembeza[2], Irmantas Rokaitis[1], Jan Zrimec[2], Simona Poviloniene[3], Audrius Laurynenas[1,3], Sandra Viknander[2], Wissam Abuajwa[4], Otto Savolainen[4], Rolandas Meskys[3], Martin K. M. Engqvist[2] and Aleksej Zelezniak[2,5] ✉

De novo protein design for catalysis of any desired chemical reaction is a long-standing goal in protein engineering because of the broad spectrum of technological, scientific and medical applications. However, mapping protein sequence to protein function is currently neither computationally nor experimentally tangible. Here, we develop ProteinGAN, a self-attention-based variant of the generative adversarial network that is able to 'learn' natural protein sequence diversity and enables the generation of functional protein sequences. ProteinGAN learns the evolutionary relationships of protein sequences directly from the complex multidimensional amino-acid sequence space and creates new, highly diverse sequence variants with natural-like physical properties. Using malate dehydrogenase (MDH) as a template enzyme, we show that 24% (13 out of 55 tested) of the ProteinGAN-generated and experimentally tested sequences are soluble and display MDH catalytic activity in the tested conditions in vitro, including a highly mutated variant of 106 amino-acid substitutions. ProteinGAN therefore demonstrates the potential of artificial intelligence to rapidly generate highly diverse functional proteins within the allowed biological constraints of the sequence space.

A protein's three-dimensional (3D) structure, physicochemical properties and molecular function are defined by its amino-acid sequence. From the 20 commonly occurring proteinogenic amino acids, a small-sized protein comprising 100 amino acids can be made in 10^{130} unique ways. In this vast multidimensional space—often referred to as the protein fitness landscape[1]—as little as 1 in 10^{77} sequences are estimated to fold into the defined 3D structures to carry out specific functions[2–4]. This imposes a great burden on experimental approaches aiming to screen for novel sequences with enhanced properties, such as random mutagenesis[1] and recombination of naturally occurring homologous proteins[5,6], as up to 70% of random single amino-acid substitutions typically result in a decline of protein activity, out of which 50% are completely deleterious to protein function[1,7–13]. On the other hand, machine learning methods are not limited by the number of sequence variations they can process[14–16] and, instead of depending on a semi-random local search process, infer protein properties[15,17] and function[16,18] directly from the amino-acid sequence. Computational approaches enabling the generation of novel functional sequence variants, bypassing experimental screening of the enormous protein sequence space, are becoming increasingly important to meet the challenges and demands for novel protein diversity in the biomedical and biotechnology fields.

Conventional bioinformatics approaches, such as those based on hidden Markov models (HMMs) and, more recently, machine learning approaches, have demonstrated great potential for capturing both the structural and evolutionary information found in natural protein sequences[14,19–21]. Successful applications of deep learning to in silico directed evolution exemplify the complementarity of traditional protein engineering and machine learning-driven synthetic biology[22,23]. Nevertheless, the majority of existing machine learning models in protein research are discriminative[14,15,23]; that is, the model is trained, using readily available data, to predict the properties of a given protein sequence. A generative modelling approach, in contrast, is able to learn the underlying data distribution and generate new samples from it. Thus, in theory, these can generate new protein sequences from the learned portion of the functional protein sequence space, providing access to unexplored yet functional sequence diversity and minimizing the need for testing large amounts of non-functional protein sequence variants. Despite the number of theoretical approaches for biological sequence generation that have been developed in the past decade[24,25], including HMMs that, by nature, are generative models and have been used for decades in protein research[25,26] (recent examples are described in refs. [27–33]), the ability of these techniques to generate novel diverse functional proteins is questionable due to the limited experimental evidence.

Hence, here we present ProteinGAN (Fig. 1a), a generative adversarial network[34] enabling the generation of novel functional protein sequences with natural-like biochemical properties. We demonstrate the neural network's ability to generalize protein sequence spaces by learning complex evolutionary dependencies between amino acids. ProteinGAN enables the generation of highly diverse sequences by generating protein structural domains that do not exist in the training data. Using malate dehydrogenase (MDH) as an example, we experimentally show ProteinGAN's potential to generate fully functional diverse enzyme proteins, where generated sequences with over 100 mutations were as active as natural enzymes. Verified by

[1]Biomatter Designs, Vilnius, Lithuania. [2]Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden. [3]Institute of Biochemistry, Life Sciences Center, Vilnius University, Vilnius, Lithuania. [4]Chalmers Mass Spectrometry Infrastructure, Chalmers University of Technology, Gothenburg, Sweden. [5]Science for Life Laboratory, Stockholm, Sweden. [6]These authors contributed equally: Donatas Repecka, Vykintas Jauniskis, Laurynas Karpus. ✉e-mail: aleksej.zelezniak@chalmers.se

mass spectrometry, the generated enzymes were specific and displayed reaction yields and activity levels similar to those of their natural counterparts. We anticipate that ProteinGAN will serve as a generalized framework enabling active protein sequence generation for synthetic biology and protein engineering applications.

Results

Generative network's latent space encodes protein features. ProteinGAN is based on generative adversarial networks[34] that we tailored to learn patterns from long biological sequences (Methods and Supplementary Fig. 1), and it extends alignment-based methods that treat each amino acid independently. Specifically, the ProteinGAN architecture is a customized temporal convolutional network[35] designed with the purpose to simultaneously analyse local and global sequence features, that is, to capture the meaningful sequence motifs and long-distance relationships that are known to be critical for correct protein structural assemblies[36]. Also, to help ProteinGAN focus on functionally important areas, such as catalytic residues (Supplementary Methods and Supplementary Fig. 1), we additionally introduced a self-attention layer[37]. The final architecture of the network comprised 45 layers with over 60 million trainable parameters.

To evaluate the performance of ProteinGAN and to demonstrate that neural networks can generalize a protein family sequence space, thus generating diverse functional proteins, we trained the neural network on a family of bacterial MDH enzymes (EC 1.1.1.37). MDH is a tricarboxylic acid cycle enzyme catalysing the conversion of malate to oxaloacetate using NAD^+ as a cofactor (Supplementary Fig. 2). We chose MDH based on the following criteria: (1) it has a large number of diverse sequences (a total of 16,706 unique sequences were used for training), which were on average 319 ± 18.2 (s.d.) amino acids long with pairwise sequence identities as low as 10%; (2) it is a complex enzyme that must bind both its substrate and the NAD^+ cofactor for catalysis and (3) its activity can be readily monitored in vitro. We assessed the progress of training by quantifying the similarity of the generated sequences to natural ones, given expected differences between the training and validation sets. At every 1,200 learning steps, 64 sequences were generated and their identities to natural sequences in the training and validation datasets were computed (Fig. 1b and Supplementary Fig. 3). After 2.5 million learning steps, at which training was terminated (Supplementary Fig. 4), the sequence identities between the generated and natural sequence sets had reached a plateau, and the median sequence identity to the closest natural sequences was 64.6%. Similar differences were observed between the training dataset and a held-out validation dataset of natural sequences (64.9%, Supplementary Fig. 5), indicating that the model did not overfit to the training dataset.

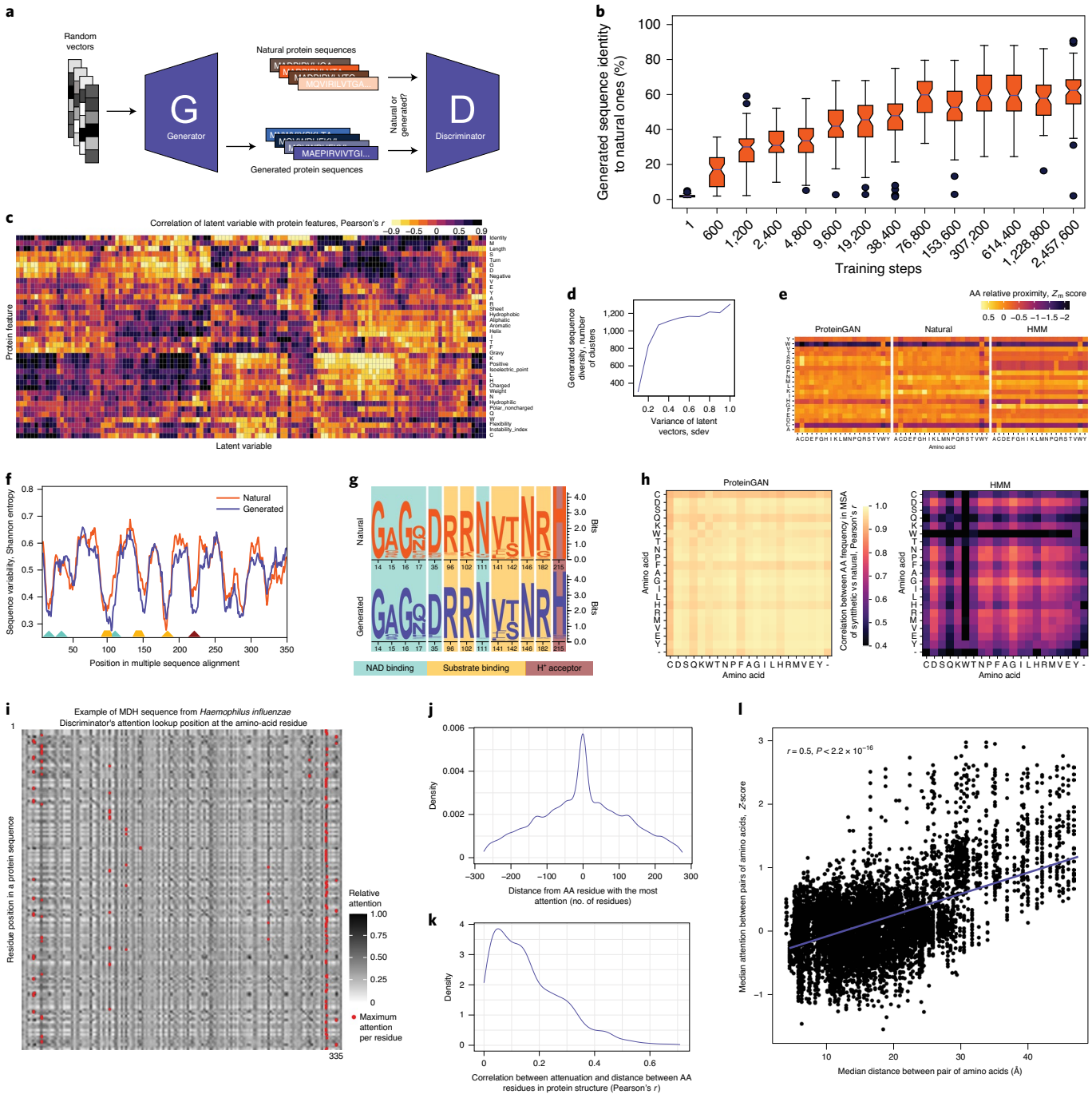
The key aspect for applying generative models to protein engineering is the understanding and control of biophysical sequence properties learned by the model. We first explored whether the model's latent space, which is a lower-dimensional representation of highly complex protein space, was reasonably mapped to primary and secondary sequence properties, by interpolating the latent space, one dimension (Supplementary Methods) at a time, and calculating the protein features from the sampled sequences (Supplementary Table 7). At least 76% of the model's latent space dimensions were highly correlated (absolute Pearson's $r > 0.8$) with corresponding primary or secondary sequence features (Fig. 1c,d and Supplementary Fig. 6) reflecting on sequence diversity that can be controlled by changing the variance of latent vectors (Fig. 1d). Following this initial quality assessment, by uniformly sampling the latent space we generated 20,000 sequences, which were further used to evaluate ProteinGAN performance.

ProteinGAN learns the intrinsic relationships of natural protein sequences. The main objective of generative models is to learn the underlying true distribution from representative samples, so we first evaluated the ability of ProteinGAN to capture biologically important information, such as evolutionary sequence properties reflected in the statistics of amino-acid variation in the natural and generated sequences. Shannon entropies were computed for each position in multiple sequence alignments of the generated and natural MDH sequences (Fig. 1f). The positional variability in generated sequences was highly similar to that of the natural sequences, with peaks (high entropy) and valleys (low entropy) appearing at nearly identical positions in the sequence alignment, demonstrating an overall high correlation (Pearson's $r = 0.89$, $P < 1 \times 10^{-16}$; m.s.e. $= 0.0039$) between the entropy values of generated and natural sequences. As a control we analogously trained (Methods) a conventional model based on

Fig. 1 | ProteinGAN learns the intrinsic relationships between natural protein sequences. a, ProteinGAN training scheme. Given a random input vector, the Generator network produces a protein sequence, which is scored by the Discriminator network by comparing it to the natural protein sequences. The generator tries to fool the discriminator by generating sequences that will eventually look like real ones (the generator never actually sees real enzyme sequences). **b**, Sequence identity of 64 generated sequences to the nearest natural sequence from training data at different training iterations. **c**, Latent space vectors correlate with protein properties as calculated by interpolating each variable dimension. **d**, Sequence diversity can be controlled by changing the variance of latent vectors. The *x* axis shows the standard deviation of interpolated latent vectors, and the *y* axis demonstrates the increase of the number of sequence clusters with 70% within the cluster sequence similarity threshold. **e**, ProteinGAN effectively captures the amino-acid (AA) distribution of natural MDH sequences. Sequence variability is expressed as Shannon entropies for generated and training sequences estimated from multiple sequence alignment (MSA). Low Shannon entropy values represent highly conserved and thus functionally relevant positions, whereas high entropy indicates high amino-acid diversity at a given position. **f**, A sequence logo of key conserved positions in the multiple sequence alignment. **g**, ProteinGAN learns the order of amino acids in natural MDH sequences as opposed to HMMs. Amino-acid pair association (Z_m positional score) matrices for natural and generated protein sequences are shown. Positive values indicate a larger distance than expected when comparing random sequences with the same amino-acid frequency. The numbers indicate by how many positions, on average, the amino acids in a pair are closer (negative values) or further apart (positive values) than in a random sequence. **h**, Amino-acid pair correlations of generated and natural sequences. Every point on the map represents the correlation of amino-acid pair frequencies between the MSA of natural MDH and synthetic sequences generated by ProteinGAN and HMMs. High correlation denotes that the same pairwise long-distance amino-acid interactions were found as in natural sequences. **i**, The hidden layer discriminator's self-attention representation for the 6AOO sequence taken from PDB[74], as an example. To discriminate between natural and synthetic sequences the discriminator attends at different positions in a given sequence. The vertical axis represents the running position in a sequence, while the horizontal axis represents the positions and attended scores for quantifying the residue importance to discriminate between fake and real sequences. The red dots represent residue positions that the discriminator paid maximum attention to when scoring the sequence. **j**, Distribution of positions where maximum attention was focused in real MDH sequences (training data); negative positions represent the lower left diagonal part of **i** and positive positions the upper right. **k**, Attention scores between pairs of amino acids correlate with amino-acid median Euclidean distances in the corresponding protein structures (Methods). **l**, Top 50 highly correlated examples of MDH sequence-to-structure matched representatives from training data (**k**) where each dot represents an amino-acid residue pair from one of the top sequences.

first-order HMM[38]. Although HMMs are powerful and a de facto golden standard for protein functional annotation[39,40], when used for sequence generation, the first-order HMM did not recapitulate the sequence conservation present in the training data, as the positional entropy error (m.s.e.) of HMM-generated sequences compared to natural ones exceeded the error with ProteinGAN by over 21-fold (Fig. 1e and Supplementary Fig. 7). At conserved positions, ProteinGAN-generated sequences preserved key substrate-binding and catalytic residues (Fig. 1f). Further comparative analysis of generated and natural sequences showed that, even in highly variable sequence regions, the frequencies of individual amino acids were perfectly correlated (Pearson's $r=0.96$, $P<1\times10^{-16}$; Supplementary Fig. 8). Moreover, for each individual sequence, ProteinGAN inferred the specific physicochemical amino-acid signatures present in the corresponding enzyme class. For example, despite the high sequence diversity among generated sequences, the fractions of hydrophobic, aromatic, charged and cysteine-containing residues were practically the same (Wilcoxon rank sum test, $P>0.05$) as in natural ones. Apart from the differences in hydrophilic and polar uncharged residues ($P=7\times10^{-5}$ and 1×10^{-28}, respectively), the network had learned the overall amino-acid composition corresponding to both the evolutionary and physicochemical constraints (Fig. 1e,g, Supplementary Table 1 and Supplementary Figs. 9 and 10).

Moving forward, we assessed whether ProteinGAN was able to learn the specific local positional order of amino acids across the full length of the MDH sequences. To investigate such local relationships, we calculated the amino-acid association measures for



natural and generated sequences using the minimal proximity function Z_m (ref. [41]). For each pairwise combination of the 20 amino acids, the function Z_m(A,B) counts the average distance between amino acid A and the next amino acid, B, occurring in the sequence. The calculated distances can be expressed as a matrix of all pairwise combinations (Fig. 1h, insets). The average similarity of the positional order in the natural and generated sequences is 82% (Fig. 1g), showing that ProteinGAN captures the local amino-acid relationships existing in natural sequences. By comparison, the average positional error of amino acids in HMM-generated sequences compared to natural sequences was 180% (Supplementary Fig. 11).

In proteins, amino-acid pairs that are remote on the primary sequence are often spatially close and interact in the 3D structure, ensuring appropriate protein stability and function[36]. To investigate global amino-acid relationships, we calculated the pairwise amino-acid frequency distributions for all combinations of positional pairs in all sequences in multiple sequence alignments. These frequency distributions were then used to calculate correlations between the training and generated sequences. Overall, we found strong correlations between the natural and generated sequences (averaged Pearson's $r=0.95$, $P<1\times10^{-6}$, Fig. 1h, left), which demonstrated that the amino-acid pairwise relationships were highly similar in both sets of sequences. As expected, the first-order HMMs are by design 'memoryless', meaning that the amino-acid residue of a particular position is independent of the long-distance residues of all others, and thus the HMMs do not capture global amino-acid relationships (Fig. 1h, right). In addition, we examined the self-attention layer of ProteinGAN to investigate which amino acids and their positions were important for distinguishing synthetic and natural sequences by the discriminator network (Fig. 1i). Although, on average, discriminator significantly (Z-score of attention scores > 3, Methods) attended to the local neighbouring residues, for over 66% of residues the maximum attention aimed from over ±50 residues away from the position of amino acid (Fig. 1j). Furthermore, the attention scores were moderately correlated (up to Pearson's $r=0.7$, $P<1\times10^{-6}$, Fig. 1k) with the physical distances between amino acids, as determined from the 3D structure of corresponding sequences (Fig. 1l), demonstrating the network's ability to recognize functionally relevant features directly from sequences. To expand on this, we inspected whether the generated MDH sequences possessed the two main Pfam[42] domains, 'Ldh_1_N' and 'Ldh_1_C', that were identified (E-value $<1\times10^{-10}$) in the natural MDH sequences and that are both over 100 amino acids long. We found that 98% of ProteinGAN-generated sequences contained both signatures, with the rest containing one of the two domains, preserving long-distance amino-acid relationships of the enzyme family. By contrast, only 1.7% of the HMM-generated sequences contained the required MDH domains. Collectively, ProteinGAN-generated sequences closely mimic natural MDH proteins, both in terms of amino-acid distributions at individual sites, as well as in terms of local and long-distance relationships between pairs of amino acids present throughout the primary sequence of the MDH family.

ProteinGAN expands the known MDH sequence space. Visualization of the sequence diversity of generated and natural sequences using t-distributed stochastic neighbour embedding (t-SNE) dimensionality reduction[43] showed that a majority of natural MDH sequences grouped into large clusters (Fig. 2a), as they were highly similar (median pairwise identity of 92%, Supplementary Fig. 12). By contrast, the generated sequences grouped into smaller clusters, interpolating between the natural sequence clusters, and resembled a learned manifold of the MDH sequence space (Fig. 2a). On the contrary, as confirmed with alignments (Fig. 1h and Supplementary Fig. 7), HMM-generated sequences were similar only to themselves (Supplementary Fig. 13), thus, failing to recapitulate the natural MDH sequence properties, and therefore were omitted from further analyses.

To investigate the diversity of the synthetic sequences, using clustering analysis we observed that, on average, over 95% of the generated sequences were not more than 10% similar to each other (90% sequence identity within the cluster, Fig. 2b), in contrast to only 17% of the natural sequences at the same sequence identity level. Similarly, at ~75% sequence identity, the generated sequence diversity exceeded the diversity of the training data by up to four times (Fig. 2b, inset). We then explored whether ProteinGAN was able to generalize the protein family beyond the training data, for instance, to be able to generate sequences that have novel structurally relevant sequence properties. For this, we first evaluated whether the network generated new structural domain diversity over the training period (Fig. 2c) by counting the presence of CATH[44] domains corresponding to all known protein 3D structural motifs (Methods). Although the number of identified structural domains plateaued at the early stage of training (after ~0.2 million steps), corresponding to 79% of all identified domains, additional structural CATH domains were discovered throughout the entire training process. We also evaluated whether the generated structural domain diversity was due to chance. As a control, we randomly introduced amino-acid substitutions into the natural MDH sequences, while preserving the natural amino-acid frequency distribution and the rate of mutations to mimic the natural sequence variability (Methods). ProteinGAN-generated sequences contained over 2.6 times more (ratio of area under the curve (AUC) in Fig. 2c) structural domains than expected by chance. Furthermore, to evaluate whether the fully trained model was able to reproduce the sequence diversity, we analysed the generated sequences after training. The trained ProteinGAN model introduced a total of 119 novel (not present in the training dataset) structural sequence motifs ($E<1\times10^{-6}$) in a random subset of 10,000 generated sequences (Fig. 2c, inset), demonstrating the network's potential to generalize relationships between residues by generating diverse sequences with novel structural properties. The total number of structural domains was reduced by 38.9% in the randomly mutated natural sequence controls, of which 97.4% of the mutated sequences were present in the training data, demonstrating it is highly unlikely that the observed structural domain diversity in the generated sequences is due to chance (Fig. 2c, inset; Fisher's exact test $P<8.2\times10^{-16}$).

Generated enzymes are functional in vitro. Finally, considering that random amino-acid substitutions typically result in a decline or even complete loss of protein activity[1,7–13], we experimentally tested whether the ProteinGAN-generated MDH sequences were catalytically active in vitro. We obtained pairs of the most similar generated and training sequences by searching the generated sequences against the training set. The results were filtered to include pairs with sequence identities ranging from 40 to 100% (Fig. 2a, inset) to discard extremely divergent sequences, which are likely to be inactive. From this set we picked up to 60 sequences for testing and checked for essential amino acids in functional positions (as shown in Fig. 1f) to avoid obvious true negatives. The resulting sequences fell in the range of 45–98% pairwise sequence identity to natural MDH, having 7–157 amino-acid mutations (including substitutions, insertions and deletions) compared to their closest MDH neighbour (Fig. 1b, Fig. 2a inset and Supplementary Table 2), of which 55 were successfully synthesized and cloned into an expression vector. As confirmed by homology modelling (Methods), the original sequence diversity was present throughout the entire structure of generated MDHs (Supplementary Fig. 14). This means that the amino acids of generated enzymes present in fewer than 5% of training sequences span the surface and core protein parts, showing that ProteinGAN generates distinctive sequences with no observed structural biases.

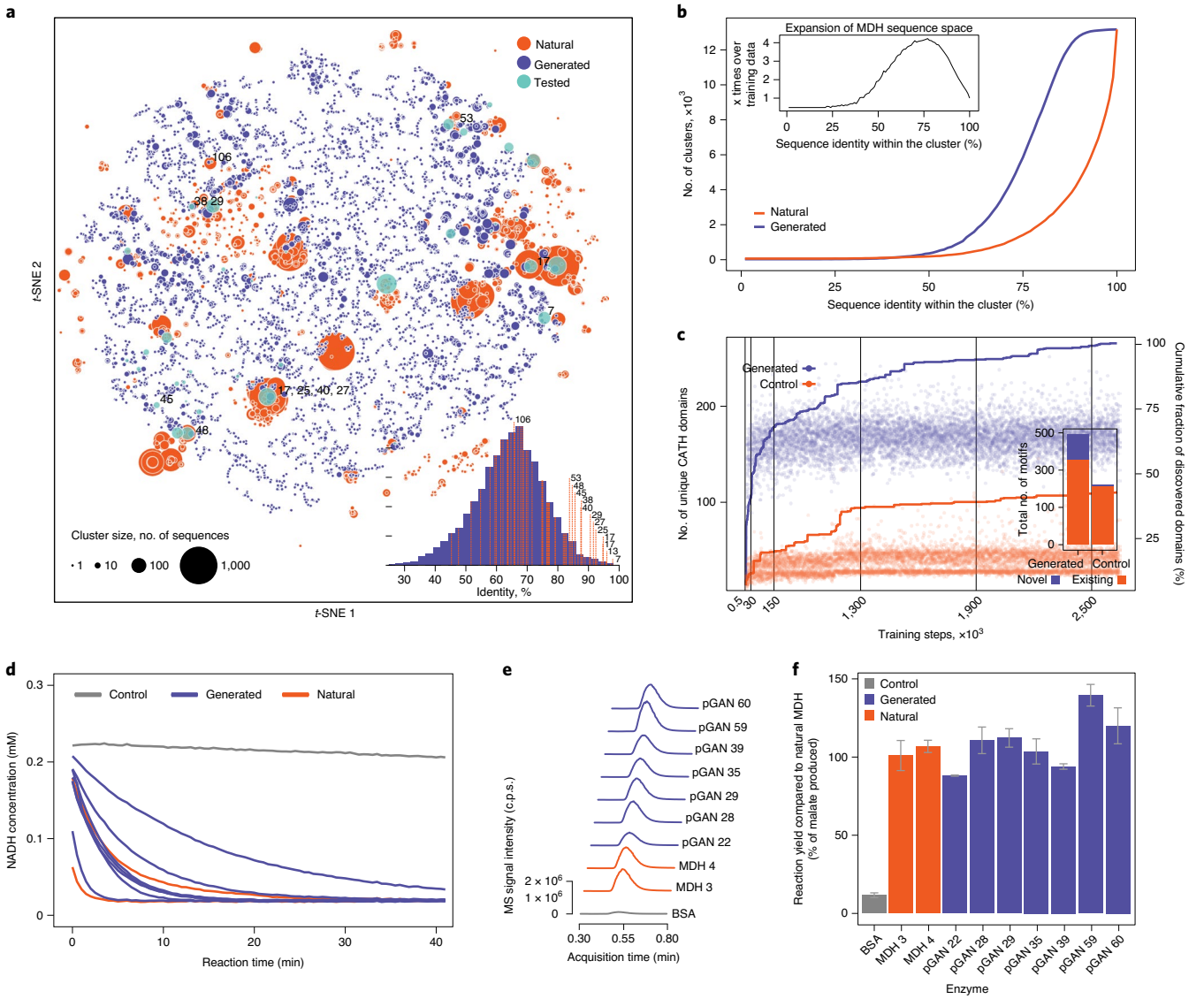


Fig. 2 | ProteinGAN expands the functional MDH sequence space. a, The protein sequence space was visualized by transforming a distance matrix derived from *k*-tuple measures of protein sequence alignment[80] into a *t*-SNE embedding. Dot sizes represent the 80% identity cluster size for each representative. Contrary to natural sequences, generated sequences formed disparate small clusters, indicating their diverse nature. Numbers indicate mutations (including substitutions, deletions and insertions) in experimentally active generated variants compared to their closest natural counterparts. Inset: global pairwise sequence identity of the generated sequences to the closest sequence in the training dataset. Dashed vertical lines indicate experimentally tested sequences, and numbers indicate the amount of amino-acid mutations, that is, the number of substitutions, insertions and deletions compared to the closest natural MDH neighbour observed in experimentally active synthetic sequences (Supplementary Fig. 17). **b**, Comparison of sequence diversity between the generated sequences and the training MDH dataset. By varying the sequence identity cutoffs, generated sequences group into up to four times more clusters than natural sequences, demonstrating an expanded sequence diversity. Inset: the ratio of the number of clusters (*y* axis) at different sequence identity cutoffs (*x* axis). **c**, CATH[44] domain diversity generated throughout the evolution of ProteinGAN training. At every 1,200 training iteration, 64 sequences were sampled. As controls, at every iteration, 64 sequences were sampled from the training data and randomly mutated to account for the average difference in the identity of generated and natural sequences. Both controls and generated sequences were then searched for representative CATH domains ($E < 1 \times 10^{-6}$). Inset: ProteinGAN-generated novel domains that are not present in the existing MDH family (left). In contrast to ProteinGAN-generated sequences, randomly introducing mutations does not expand MDH sequence diversity (Fisher's exact test $P < 8.2 \times 10^{-16}$), but rather decreases CATH domain diversity (right). For controls, 10,000 randomly mutated natural sequences were simultaneously searched for the presence of CATH domains (Methods). **d**, MDH activity measured by fluorescently monitoring NADH consumption (Methods, Protocol 1). **e**, Catalytic activity confirmed using liquid chromatography tandem mass spectrometry (LC-MS/MS; enzymes expressed using Methods, Protocol 1). **f**, Oxaloacetate-to-malate conversion yields are comparable to natural MDH enzymes as determined using mass spectrometry, confirming that the primary product of synthetic enzymes is analogous to the natural control.

One of the major challenges in recombinant protein expression is protein solubility[45], often as a consequence of the experimental set-up, where the expression systems (for example, codon optimality) and growth conditions can be inadequate. Also, natural protein sequences can even be misfolded and inactive. Therefore, to maximize the suitability and effectiveness of the experimental procedure, we used two different experimental set-ups and two different strains (Methods, Protocols 1 and 2).

The production of recombinant proteins in *Escherichia coli* and purification using affinity chromatography yielded 11 protein variants (Methods, Protocol 1) that could be purified from the cell lysate soluble fraction (Supplementary Table 3 and Source Data Fig. 1). With the aim of identifying additional soluble proteins, we repeated the experiment under growth conditions favouring protein folding and solubility using the ArcticExpress *Escherichia coli* strain (Methods, Protocol 2), expanding the number of purified soluble proteins to a total of 19 (Supplementary Table 3, 35% of all synthesized protein variants). This is comparable to other systematic studies, which typically obtain soluble fractions for 20–40% of all tested constructs[46–48]. The purified proteins were assessed for MDH activity by monitoring NADH consumption using a spectrophotometer (Supplementary Fig. 2). Thirteen of the 19 (16 generated + 3 natural controls) soluble enzymes, including a variant with 106 amino-acid substitutions (66% identity to the closest existing enzyme; Supplementary Fig. 15), showed MDH catalytic activity (Fig. 2d, Supplementary Table 3 and Supplementary Figs. 16 and 17). It is worth noting that most of the proteins that were insoluble or inactive fell in the lower sequence identity range from the nearest natural sequence (Fig. 2a, inset), whereas above 80% sequence identity (~60 mutations) the success rate reached 50% (12 active proteins out of 24). Furthermore, for the subset of eight purified enzymes for which the protein amount could be accurately quantified (Methods, Protocol 1), the generated MDH proteins displayed similar reaction rates as wild-type enzymes (Supplementary Fig. 16). These enzymes were also confirmed by LC-MS/MS to convert oxaloacetate specifically to malate, with reaction yields comparable to commercial MDH enzyme controls (Fig. 2e,f).

Discussion

A protein family is a group of evolutionarily related proteins descended from a common ancestor and generally regarded as having similar sequences, 3D structures and functions. By examining statistical patterns of amino-acid relationships in aligned protein sequences, one can gain insight into the diversity and physicochemical constraints that determine the structure and function of a particular protein domain or family. Based on these insights, however, it is exceptionally challenging to design a protein containing functionally relevant sequence motifs and the correct position-specific amino-acid composition that preserves long-range amino-acid interactions. Learning statistical models of protein sequences from multiple sequence alignments is often based on strong assumptions. For example, alignments and conventional evolutionary models[20] assume conditional independence between amino-acid residues, which is not necessarily biologically feasible, as mutations of non-adjacent residues generally have strong correlations[49–51]. Multiple sequence alignments are also often suboptimal due to the nature of greedy heuristics algorithms[52,53]. A successful protein generative model must therefore (1) cope with a high degree of conditional dependence between amino-acid residues, (2) at the same time represent well the underlying higher-dimensional sequence space to generate novel diverse sequences and (3) ideally not be dependent on alignments.

The ProteinGAN presented here, a generative adversarial network tailored explicitly for learning underlying amino-acid relationships directly from long biological sequences, enables the control of primary and secondary sequence properties by interpolating the latent space dimensions (Fig. 1c), by learning to map the high-dimensional sequence space to a lower-dimensional latent space representation. In contrast to the generation of images[54] or music[55], evaluating the quality of the generated samples—that is, if the generated sequence represents an actual protein—is challenging. We thus inspected whether ProteinGAN was able to recapitulate fundamental sequence properties existing in the MDH protein family. Indeed, by examining the variability of the expected amino-acid residues, conservation of the active site, as well as overall local and global amino-acid relationships, we showed that ProteinGAN-generated sequences contain the expected fundamental properties found in natural counterparts (Fig. 1e–h).

The goal of generative models is to learn generalized representations underlying the data distribution, enabling the generation of new samples with key properties of the target data. We show that ProteinGAN generated up to four times more diverse sequences than were present in the training data, with novel structural domains (Fig. 2b,c). To verify that the learned representation of the MDH sequence space retained MDH functionality, we experimentally tested 55 diverse sequence candidates (Fig. 2a and Supplementary Fig. 13). Through homology modelling, we also verified that the novel mutations were not localized in a particular protein domain, but spanned the entire protein structure (Supplementary Fig. 14) and primarily occurred close to the protein surface, where the most variation could usually be seen in natural proteins. In vitro experiments confirmed that 81% (13 of the 16) of all soluble generated enzymes display catalytic activities comparable to—or surpassing those of—the natural enzymes (Fig. 2d–f and Supplementary Figs. 16 and17). Indeed, the correct protein folding determines its solubility and is among the main bottlenecks in recombinant protein production, which, apart from the sequence, is highly dependent on experimental conditions[45]. For example, in the two different tested protein expression set-ups (Methods), only four common generated MDHs were active, and, out of three natural MDHs, despite all being soluble in both set-ups, not all were active in both (Supplementary Table 3, Source Data Fig. 1 and Supplementary Fig. 17). The generated functional enzymes contain up to 106 (34%) mutations (including insertions and deletions) compared to the closest natural MDH (Supplementary Figs. 13 and 15), which is a highly remarkable result given that typically up to 50% of single amino-acid substitutions result in a loss of protein function[1,7–13]. Overall, a quarter (13 out of 52) of all the tested generated enzymes were active (Supplementary Fig. 14), underlying ProteinGAN's ability to take large leaps to unexplored sections of the functional sequence space (Fig. 2a) and allowing biochemical exploration of highly diverse enzymes. Such enzymes may have catalytic properties that differ substantially from those found in natural enzymes, as they have not evolved with the constraint to carry out specific functions in living organisms, as natural enzymes have.

Navigating the fitness landscape using purely experimental methods, including random mutagenesis[1], is often highly laborious or may not even be feasible due to the exponential decline in protein fitness[8,56]. The recombination of homologous proteins allows larger leaps[5,6], but the achievable sequence space is fundamentally limited by the number of viable combinations of unique motifs found in the parent molecules used to generate the recombinant libraries[57,58] and also the experimental viability[57,58]. Instead, artificially generated sequences may provide suitable, non-natural and diverse starting points for protein engineering[59], with great potential for applications in biocatalysis[60].

Methods

Neural network architecture details. The GAN architecture consisted of two networks—a discriminator and a generator—each of which uses ResNet blocks[61] (Supplementary Fig. 1). Each block in the discriminator contained three 1D convolution layers with filter size of 3 (ref. [62]) and leaky rectified linear unit (ReLU) activations[63]. The generator residual blocks consisted of two transposed convolution layers, one convolution layer with the same filter size of 3 and leaky ReLU activations. Each network had one self-attention layer[37]. The transposed convolution technique was chosen for up-sampling, as it yielded the best results experimentally (Supplementary Fig. 18). For loss, non-saturating loss with R1 regularization[64] was used (Supplementary Fig. 19). To ensure training stability, spectral normalization[65] was implemented in all layers.

The input to the discriminator was one-hot encoded with a vocabulary size of 21 (20 canonical amino acids and a sign that denoted a space at the beginning or end of the sequence). The generator input was a vector of 128 values that were

drawn from a random distribution with mean 0 and standard deviation of 0.5, with the exception that values whose magnitudes were more than two standard deviations away from the mean were re-sampled. The dimensions of generated outputs were 512×21, where some of the positions denoted spaces.

Network training data. Bacterial MDH sequences were downloaded from UniProt on 10 January 2019[66]. Sequences longer than 512 amino acids or containing non-canonical amino acids were filtered out. The final dataset consisted of 16,898 sequences, which were clustered into 70% identity clusters using the MMseq2 tool[67] to balance the dataset during the training process. A total of 20% of the clusters with fewer than three sequences were randomly selected for validation (192 sequences) and the rest of the dataset was used for training (16,706 sequences).

Network training process. The ratio 1:1 between generator and discriminator training steps was selected (Supplementary Information and Supplementary Fig. 20). The Adam algorithm[68] was used to optimize both networks. Throughout the training, the learning rate was gradually decreased from 1×10^{-3} to 5×10^{-5} for both the generator and the discriminator. To avoid bias towards sequences with a large number of homologues, smaller clusters were dynamically up-sampled during training. To track the performance, along with GAN losses, the generated data were constantly evaluated. Without halting the training process, every 1,200 steps, generated sequences were automatically aligned with the training and validation datasets using BLAST[69]. Throughout the training, BLOSUM45 identity scores as well as the standard deviation of the discriminator layer were calculated and monitored (Supplementary Figs. 3 and 21–24). ProteinGAN was trained for 2.5 million steps with a batch size of 64 (one step consisted of 64 sequences). The training took 210 h (~9 days) on a NVIDIA Tesla P100 system (16 GB).

Bioinformatic analyses of generated sequences. *Multiple sequence alignments.* Multiple sequence alignments (MSAs; Fig. 1) were constructed using Clustal Omega[70] by merging natural and generated datasets in equal amounts. To calculate further Shanon entropies, after MSA we split the alignment corresponding to generated and training sequences. Columns with more than 75% of gaps in either dataset were removed from further analysis. For each column in MSA, the Shannon entropy was calculated as

$$\mathrm{SE} = -\sum_{i=1}^{20} p(x_i)\log_{20}p(x_i)$$

where $p(x_i)$ is the frequency of amino acid i occurring at a column of MSA.

HMM sequence generation. The profile HMM was created using *jackhmmer* (HMMER v3.3.1)[38]. An unbalanced training dataset was used as a target database and *E. coli* MDH was used as query sequence (UniProt ID P61889). The *jackhmmer* search was iterated to convergence. To prevent custom weighting of the HMM profile, the *--wnone* option was used; otherwise, the settings were not changed. The HMM profile from the last iteration of *jackhmmer* was used together with the *hmmemit* (HMMER v3.3.1) tool to generate 20,000 sequences, where default parameters were used. Generated sequences were aligned together with an unbalanced training dataset in the same way as described above.

Amino-acid pair association matrices. Amino-acid pair association matrices were calculated for every possible pair in a sequence and averaged over the whole dataset. The association score was used as reported in ref. [41], where Z_m is expressed as

$$Z_\mathrm{m}(a,b) = \frac{P_\mathrm{m}(a,b) - \underline{P_\mathrm{m}(a,\mathrm{Rand}(b))}}{\sigma_{P_\mathrm{m}(a,\mathrm{Rand}(b))}}$$

Here, $\underline{P_\mathrm{m}(a,\mathrm{Rand}(b))}$ and $\sigma_{P_\mathrm{m}(a,\mathrm{Rand}(b))}$ are the average and the standard deviation of the randomly shuffled sequence association score for the same pair. The association function for scoring was selected as the minimal proximity function:

$$P_\mathrm{m}(a,b) = \frac{1}{n}\sum_{i=1}^{n}\min_{j=1,\ldots,m}\{|x_i - y_i|\}$$

Here, for each position x_i of amino acid a, the closest occurrence of amino acid b at position y_i is identified and the average of the distances between the pairs is calculated. In our implementation, if a sequence does not contain a certain amino acid, a Null value is returned for the pairs containing the amino acid.

Sequence clustering. Sequence clusterings for dataset normalization, validation set picking and downstream analyses were performed using MMseqs2[67] with the easy-cluster option and required sequence identity cutoff.

Pfam/CATH domain search. All sequences generated by ProteinGAN were classified using a HMMER3 (HMMER v3.3.1)[38] search over the Pfam 32.0 database[42]. HMMs for each CATH representative domain from the sequence clusters at 35% sequence identity (v4.1) were downloaded from the CATH database repository[44]. To avoid biases in sequence scoring, generated sequences together with the natural sequences in equal quantities were appended to the same file and in all tests were searched simultaneously using the *hmmsearch* tool with default options[38] in MPI mode using 256 threads. To evaluate whether the generated domain diversity was not due to chance, we chose a random subset of 10,000 natural sequences and mutated them by randomly introducing, on average, 100 ± 30 (s.d.) substitutions of amino acids (corresponding to the median identity of generated sequences, Fig. 1b) that were uniformly sampled from the natural MDH amino-acid probability distribution. Generated, natural and mutated sequences (10,000 of each) were searched as one database and hits were considered significant when the $E<1\times10^{-6}$. The analysed sequences were generated using the latest checkpoint model (~2.5 million training steps). Analogous controls were used during ProteinGAN by sampling 64 sequences at every 1,200th training checkpoint. As controls, at every 1,200th checkpoint, 64 sequences were sampled from training data and randomly mutated to account for the average difference in identity of generated and natural sequences (Fig. 1b). Then, both controls and generated sequences were searched for representative CATH domains using the same database, and hits were considered significant with an E value of $<1\times10^{-6}$.

t-SNE plot generation. A distance matrix of cluster representatives was used as the *t*-SNE input. To obtain cluster representatives, the numbers of sequences in both datasets were first equalized by taking 13,272 sequences from natural and generated datasets. These sequences were independently clustered using MMseqs2[67] with 80% minimal sequence identity. This generated 926 clusters of natural sequences and 3,778 clusters of generated sequences. Representative sequences of these clusters were chosen based on the MMseqs2 output. From the representative sequences, a distance matrix was generated using Clustal Omega[70,71]. The distance matrix was used with the scikit-learn *t*-SNE module[72] with default settings (early exaggeration 12, learning rate 200, maximum number of iterations 1,000), except that the embedding generation perplexity was set to 7. Coordinates given by *t*-SNE were used for plotting and the size of a given dot was visualized based on the cluster size it represents.

Visualization of GAN training. Sequences generated during the training period were sampled at 14 different times. For each of the 14 checkpoints, 64 sequences were taken, every checkpoint was taken after $i(x)=2^x\times300$ (where x is the number of checkpoints) GAN steps, with the first checkpoint replaced with 1 instead of 300. For all the generated sequences, a global identity to the closest sequence in the training dataset was calculated. Identities of each checkpoint were plotted.

Correlation of distant dimer pairs. To calculate the correlation of close and distant dimer pairs between the datasets, the total number of individual dimers for every possible MSA position pair was calculated as $d_{n,m}=\sum_{z=1}^{s} a_i a_j$, where $m_{ai,aj}$ is a set of dimer $a_i a_j$ counts over each position pair of a multiple sequence alignment ($m_{ai,aj}=\{d_{1,1}, d_{1,2}, \ldots, d_{n,n}\}$). Each number of the $m_{ai,aj}$ set was calculated by summing $a_i a_j$ dimers over all sequences of MSA in positions n and m:

$$d_{n,m} = \sum_{z=1}^{s} a_i a_j$$

where s is the total number of sequences in MSA. *ds* was calculated for each dataset and for each member of the *ds* set, and Pearson's r was calculated between the datasets (natural and generated). These correlations were plotted as a heatmap (Supplementary Fig. 11). For *ds* calculations, only columns containing less than 75% of gaps in both natural or generated datasets were used.

Correlation of attention with distance in protein structures. To match the training set data, MDH sequences were blasted against target sequences stored in the PDB database (accessed 15 September 2020)[73]. For each query sequence, two best matching PDB structures were used for analysis with $E<1\times10^{-10}$ and match identify of >50%. The pairwise distances between each residue were calculated as the Euclidean distance between the coordinates of amino-acid β-carbons. For glycine residues, the coordinates of α-carbons were used. To obtain the attention scores, cluster representatives from the training data (section 'Network training data') were fed through the discriminator network to the self-attention layer where intermediate calculations of attention scores were extracted. The discriminator self-attention layer was up-sampled to match the input sequence positions. Then, for each amino-acid pair in a sequence-to-structure match, we computed the Pearson's correlation between the median attention score (standardized to zero mean and unit variance) and analogously the median Euclidean distance between the same amino-acid pair.

Homology modelling. Homology modelling (Supplementary Fig. 14) was carried out for all experimentally tested ProteinGAN sequences by first identifying suitable templates in the PDB database[74]. The template search was performed using BLAST[69] with default settings, and only the first best hit for each sequence was considered further. Sequence–template alignments obtained with BLAST were realigned using the Align2D routine in the Modeller package[75]. The templates

found using BLAST are listed in Supplementary Table 5 with corresponding *E* values, PDB IDs and identity percentages from the Align2D generated alignments. Models for each sequence were generated using Modeller and previously obtained sequence–template alignments. Tetrameric models were built from monomeric models by using the 3NEP structure as a template. For this task, the PyMOL routine 'align' was used. The originality of sequences was investigated using MSAs; that is, MSAs were constructed using BLAST with default settings (over the non-redundant protein database) to find the 100 nearest homologues and by realigning them with Clustal Omega[76]. From the obtained MSAs, probability distributions of amino acids in each MSA column were calculated. Here, originality was defined as the probability to find an amino acid in a position of interest. A small probability represented a very original substitution and a large probability represented a very conservative amino acid (Supplementary Fig. 14). The statistics of 100 sequences found for each ProteinGAN-generated sequence are listed in Supplementary Table 6. Residue depth calculations were carried out using structures of monomers modelled with the *biopython*[77] routine ResidueDepth, which first constructs the molecular surface with MSMS[78] and then calculates the distance of the α-carbon atoms to the surface.

Experimental validation of generated enzymes. The sequences generated by ProteinGAN were synthesized, cloned into the pET21a expression vector and sequence-verified by Twist Bioscience. In addition to the enzyme sequence, a C-terminal linker and four histidines (AAALEHHHH) were added, resulting in a deca-His-tag in the final construct (which includes six histidines derived from the expression vector), to enable downstream affinity purification. In Protocol 1, the constructs were transformed into the BL21(DE3) *E. coli* expression strain. From the resulting transformation mixture, 15 µl was used to inoculate 500 µl of LB broth supplemented with 100 µg ml^{-1} carbenicillin. Cells were grown overnight at 32 °C in a 96-deep-well plate with 700 r.p.m. orbital shaking. Protein expression was achieved by diluting the overnight cultures 1:30 into 1 ml autoinduction Terrific Broth (TB) medium including trace elements (Formedium) and supplemented with 100 µg ml^{-1} carbenicillin and grown for 4 h at 37 °C, followed by overnight growth at 18 °C and 700 r.p.m. shaking. Cells were collected by centrifugation and the cell pellets were frozen at −80 °C overnight. To purify the recombinant proteins, cells were thawed, resuspended in 200 µl lysis buffer (50 mM HEPES pH 7.4, 5% glycerol, 300 mM NaCl, 0.5 mM tris(2-carboxyethyl)phosphine (TCEP), 0.5 mg ml^{-1} lysozyme, 10 U ml^{-1} DNaseI, 2 mM $MgCl_2$), and incubated for 30 min at room temperature. To improve lysis, Triton-X-100 was added to a final concentration of 0.125% (vol/vol), and the cells were frozen at −80 °C for 30 min. After thawing in a room-temperature water bath, the lysates were spun down for 10 min at 3,000*g* to remove cell debris, and the supernatants were transferred to a new 96-well plate with 50 µl Talon resin in each well (Takara Bio). Unspecific binding of proteins to the resin was reduced by adding imidazole to a final concentration of 10 mM in each well. The plate was incubated at room temperature for 30 min with shaking at 400 r.p.m., after which the lysates with the beads were transferred to a 96-well filter plate (Thermo Scientific, Nunc 96-well filter plates), placed over a 96-well collection plate, and centrifuged for 1 min at 500*g* in a swing-out centrifuge. The resin was washed three times with 200 µl wash buffer (50 mM HEPES pH 7.4, 5% glycerol, 300 mM NaCl, 0.5 mM TCEP, 40 mM imidazole), and the proteins were eluted from the resin in two 50-µl fractions using elution buffer (50 mM HEPES pH 7.4, 5% glycerol, 300 mM NaCl, 0.5 mM TCEP, 250 mM imidazole). The two eluate fractions were combined and transferred to a 96-well desalting plate (Thermo Scientific, Zeba Spin Desalting Plate, 7K molecular weight cutoff) pre-equilibrated with sample buffer (50 mM HEPES pH 7.4, 5% glycerol, 300 mM NaCl, 0.5 mM TCEP). The plate was spun down at 1,000*g* for 1 min, and collected proteins were analysed by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS–PAGE) followed by Coomassie staining. The soluble proteins were carried forward for further characterization.

To test for MDH activity, an aliquot of purified protein was added to a reaction mixture containing 0.15 mM NADH, 0.2 mM oxaloacetic acid and 20 mM HEPES buffer (pH 7.4). The final reaction volume was 100 µl and the reaction was carried out at room temperature in a UV-transparent 96-well half-area plate (UV-Star Microplate, Greiner). Activity was measured in triplicates by following NADH oxidation to NAD+, with an absorbance reading at 340 nm performed every 30 s for 15 min in a BMG Labtech SPECTROstar Nano spectrophotometer. Unspecific oxidation of NADH was monitored in no-substrate controls, and these values were subtracted from the other samples. Conversion from absorption values to NADH concentration was carried out using an extinction coefficient of 6.22 mM. For calculation of kinetic parameters, 10 nM of each protein was assayed with a range of oxaloacetate concentrations.

LC-MS/MS quantification was performed for selected active enzymes where concentration could be measured accurately. The activity assay was performed as outlined above, in triplicates, with protein concentrations ranging between 10 and 250 nM. Reactions were terminated after 45 min by diluting the assay mixtures in water to 1 µg ml^{-1} starting concentration of oxaloacetate. For chromatographic separation, a Zorbax Eclipse Plus C18 50 mm × 2.1 mm × 1.8 µm column (Agilent) was used with a Nexera series high-performance liquid chromatography (HPLC) system (Shimadzu). Mobile phase A was composed of H_2O (MiliQ HPLC grade) with 0.1% formic acid (Sigma), and mobile phase B was methanol (Sigma) with 0.1% formic acid (Sigma). The oven temperature was 40 °C. The chromatographic gradient was set to consecutively increase from 0% to 100%, hold, decrease from 100% to 0% and hold, in 60 s, 30 s, 30 s and 30 s, respectively. The autosampler temperature was 15 °C and the injection volume was 0.5 µl with full loop injection. For MS quantification, a QTRAP 6500 system (Sciex) was used, operating in negative mode with multiple reaction monitoring parameters optimized for malic acid based on published parameters[79]. Electrospray ionization parameters were optimized for 0.8 ml min^{-1} flow rate, and were as follows: electrospray voltage of −4,500 V, temperature of 500 °C, curtain gas of 40, collisionally activated dissociation (CAD) gas set to medium, and gas 1 and 2 of 50 and 50 p.s.i., respectively. The instrument was mass-calibrated with a mixture of polypropylene glycol standards. Analyst 1.7 (Sciex) and MultiQuant 3 (Sciex) softwares were used for analysis and quantitation of results, respectively.

In Protocol 2, additionally, to increase protein solubility, MDH constructs were transformed into ArcticExpress competent cells (Agilent Technologies). The transformants were inoculated into 500 µl of Luria Broth (LB) medium with 15 µg ml^{-1} gentamicin and 50 µg ml^{-1} ampicillin and grown overnight at 30 °C in a Thermomixer Comfort Eppendorf thermomixer (Eppendorf). Volumes (250 µl) of overnight culture were transferred to 10 ml (dilution 1:40) of semi-synthetic medium (1% tryptone, 0.5% yeast extract, 0.268% $(NH_4)_2SO_4$, 0.15% NH_4Cl, 0.6% KH_2PO_4, 0.4% K_2HPO_4, 1% glycerol, pH 7.0) supplemented with 15 µg ml^{-1} gentamicin and 50 µg ml^{-1} ampicillin. The cells were cultivated at 37 °C for 2 h, until reaching an optical density at 600 nm of 0.6–0.8, then the medium was enriched with 0.5 M saccharose. Induction was carried out at 12 °C with 0.5 mM IPTG overnight. The cells were collected by centrifugation (4,000*g* for 10 min at 4 °C), resuspended in 0.1 M potassium phosphate buffer, pH 7.0 and then sonicated on ice in 2.0-ml tubes at 30% amplitude for 5 min of total on time (30 s on/30 s off) by using the Bandelin SonoPuls HD 2070 homogenizer.

To remove cell debris, the lysates were centrifuged at 16,000*g* and 4 °C. The soluble recombinant MDH mutants were purified using HisPur Ni-NTA spin columns (Thermo Fisher Scientific). The columns with loaded supernatants were washed with a wash buffer (0.1 M potassium phosphate buffer, pH 7.4, NaCl 250 mM, 40 mM imidazole). Proteins were eluted with an elution buffer (0.1 M potassium phosphate buffer, pH 7.4, NaCl 250 mM, 300 mM). The eluted fractions were dialysed against 0.1 M potassium phosphate buffer, pH 7.4. The concentration of the proteins was determined using a NanoDrop 2000 system (Thermo Fisher Scientific). The aliquots of total lysate, soluble lysate fraction and purified protein were loaded onto SDS–PAGE 15%.

The MDH activity was measured at 25 °C in a 96-well flat-bottom UV-transparent plate (UV-Star microplate, Greiner Bio-One). The reaction mixture (final volume of 200 µl) contained an aliquot of purified protein, freshly prepared 0.15 mM NADH and 0.2 mM oxaloacetic acid, and 0.1 M potassium phosphate buffer, pH 7.4. The absorbance reading was performed at 340 nm every 5 s for 3 min in a BioTek PowerWave XS microplate reader (Biotek). For NADH at 340 nm, an extinction coefficient of 6.22 mM cm^{-1} (εM) was used. The path length (l) in the microplate was calculated according to $A = c \times \varepsilon M \times l$.

Data availability

All training data files, including ProteinGAN running examples, have been deposited to the Zenodo repository and are available at https://doi.org/10.5281/zenodo.4068040. Source data are provided with this paper.

Code availability

The implementation of ProteinGAN can be accessed at https://github.com/Biomatter-Designs/ProteinGAN.

Received: 26 June 2020; Accepted: 1 February 2021;
Published online: 04 March 2021

References

1. Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **10**, 866–876 (2009).
2. Keefe, A. D. & Szostak, J. W. Functional proteins from a random-sequence library. *Nature* **410**, 715–718 (2001).
3. Taverna, D. M. & Goldstein, R. A. Why are proteins marginally stable? *Proteins* **46**, 105–109 (2002).
4. Axe, D. D. Estimating the prevalence of protein sequences adopting functional enzyme folds. *J. Mol. Biol.* **341**, 1295–1315 (2004).
5. Hansson, L. O., Bolton-Grob, R., Massoud, T. & Mannervik, B. Evolution of differential substrate specificities in Mu class glutathione transferases probed by DNA shuffling. *J. Mol. Biol.* **287**, 265–276 (1999).
6. Crameri, A., Raillard, S. A., Bermudez, E. & Stemmer, W. P. DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* **391**, 288–291 (1998).
7. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl Acad. Sci. USA* **103**, 5869–5874 (2006).
8. Guo, H. H., Choe, J. & Loeb, L. A. Protein tolerance to random amino acid change. *Proc. Natl Acad. Sci. USA* **101**, 9205–9210 (2004).

9. Rennell, D., Bouvier, S. E., Hardy, L. W. & Poteete, A. R. Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.* **222**, 67–88 (1991).
10. Axe, D. D., Foster, N. W. & Fersht, A. R. A search for single substitutions that eliminate enzymatic function in a bacterial ribonuclease. *Biochemistry.* **37**, 7157–7166 (1998).
11. Shafikhani, S., Siegel, R. A., Ferrari, E. & Schellenberger, V. Generation of large libraries of random mutants in *Bacillus subtilis* by PCR-based plasmid multimerization. *Biotechniques* **23**, 304–310 (1997).
12. Rockah-Shmuel, L., Tóth-Petróczy, Á. & Tawfik, D. S. Systematic mapping of protein mutational space by prolonged drift reveals the deleterious effects of seemingly neutral mutations. *PLoS Comput. Biol.* **11**, e1004421 (2015).
13. Sarkisyan, K. S. et al. Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).
14. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019).
15. AlQuraishi, M. End-to-end differentiable learning of protein structure. *Cell Syst.* **8**, 292–301 (2019).
16. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Preprint at *bioRxiv* https://doi.org/10.1101/622803 (2020).
17. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).
18. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-only deep representation learning. Preprint at *bioRxiv* https://doi.org/10.1101/589333 (2019).
19. Romero, P. A., Krause, A. & Arnold, F. H. Navigating the protein fitness landscape with Gaussian processes. *Proc. Natl Acad. Sci. USA* **110**, E193–E201 (2013).
20. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).
21. Söding, J. Protein homology detection by HMM–HMM comparison. *Bioinformatics* **21**, 951–960 (2005).
22. Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M. & Church, G. M. Low-N protein engineering with data-efficient deep learning. Preprint at *bioRxiv* https://doi.org/10.1101/2020.01.23.917682 (2020).
23. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322 (2019).
24. Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I. & Langmead, C. J. Learning generative models for protein fold families. *Proteins* **79**, 1061–1078 (2011).
25. Boomsma, W. et al. A generative, probabilistic model of local protein structure. *Proc. Natl Acad. Sci. USA* **105**, 8932–8937 (2008).
26. Krogh, A., Brown, M., Mian, I. S., Sjölander, K. & Haussler, D. Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235**, 1501–1531 (1994).
27. Tubiana, J., Cocco, S. & Monasson, R. Learning protein constitutive motifs from sequence data. *eLife* **8**, e39397 (2019).
28. Riesselman, A. J., Shin, J. E., Kollasch, A. W. & McMahon, C. Accelerating protein design using autoregressive generative models. Preprint at *bioRxiv* https://doi.org/10.1101/757252 (2019).
29. Greener, J. G., Moffat, L. & Jones, D. T. Design of metalloproteins and novel protein folds using variational autoencoders. *Sci. Rep.* **8**, 16189 (2018).
30. Anand, N. & Huang, P. Generative modeling for protein structures. In *Advances in Neural Information Processing Systems* Vol. 31 (eds Bengio, S. et al.) 7494–7505 (Curran Associates, 2018).
31. Killoran, N., Lee, L. J., Delong, A., Duvenaud, D. & Frey, B. J. Generating and designing DNA with deep generative models. Preprint at https://arxiv.org/pdf/1712.06148.pdf (2017).
32. Amimeur, T., Shaver, J. M., Ketchem, R. R. & Taylor, J. A. Designing feature-controlled humanoid antibody discovery libraries using generative adversarial networks. Preprint at *bioRxiv* https://doi.org/10.1101/2020.04.12.024844 (2020)
33. Gupta, A. & Zou, J. Feedback GAN for DNA optimizes protein functions. *Nat. Mach. Intell.* **1**, 105–111 (2019).
34. Goodfellow, I. et al. Generative adversarial nets. In *Advances in Neural Information Processing Systems* Vol. 27 (eds Ghahramani, Z. et al.) 2672–2680 (Curran Associates, 2014).
35. Bai, S., Kolter, J. Z. & Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. Preprint at https://arxiv.org/pdf/1803.01271.pdf (2018).
36. Marks, D. S. et al. Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6**, e28766 (2011).
37. Zhang, H., Goodfellow, I., Metaxas, D. & Odena, A. Self-attention generative adversarial networks. Preprint at https://arxiv.org/pdf/1805.08318.pdf (2018).
38. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
39. Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A. & Durbin, R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.* **26**, 320–322 (1998).
40. Bileschi, M. L. et al. Using deep learning to annotate the protein universe. Preprint at https://doi.org/10.1101/626507 (2019).
41. Santoni, D., Felici, G. & Vergni, D. Natural vs random protein sequences: discovering combinatorics properties on amino acid words. *J. Theor. Biol.* **391**, 13–20 (2016).
42. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
43. Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
44. Dawson, N. L. et al. CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.* **45**, D289–D295 (2017).
45. Rosano, G. L. & Ceccarelli, E. A. Recombinant protein expression in *Escherichia coli*: advances and challenges. *Front. Microbiol.* **5**, 172 (2014).
46. Huang, H. et al. Panoramic view of a superfamily of phosphatases through substrate profiling. *Proc. Natl Acad. Sci. USA* **112**, E1974–E1983 (2015).
47. Pertusi, D. A., Stine, A. E., Broadbelt, L. J. & Tyo, K. E. J. Efficient searching and annotation of metabolic networks using chemical similarity. *Bioinformatics* **31**, 1016–1024 (2015).
48. Mashiyama, S. T. et al. Large-scale determination of sequence, structure and function relationships in cytosolic glutathione transferases across the biosphere. *PLoS Biol.* https://doi.org/10.1371/journal.pbio.1001843 (2014).
49. Lockless, S. W. & Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295–299 (1999).
50. Socolich, M. et al. Evolutionary information for specifying a protein fold. *Nature* **437**, 512–518 (2005).
51. Russ, W. P., Lowery, D. M., Mishra, P., Yaffe, M. B. & Ranganathan, R. Natural-like function in artificial WW domains. *Nature* **437**, 579–583 (2005).
52. Pervez, M. T. et al. Evaluating the accuracy and efficiency of multiple sequence alignment methods. *Evol. Bioinform. Online* **10**, 205–217 (2014).
53. Nuin, P. A. S., Wang, Z. & Tillier, E. R. M. The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics* **7**, 471 (2006).
54. Karras, T., Laine, S. & Aila, T. A style-based generator architecture for generative adversarial networks. Preprint at https://arxiv.org/pdf/1812.04948.pdf (2018).
55. van den Oord, A. et al. WaveNet: a generative model for raw audio. Preprint at https://arxiv.org/pdf/1609.03499.pdf (2016).
56. Bloom, J. D. et al. Thermodynamic prediction of protein neutrality. *Proc. Natl Acad. Sci. USA* **102**, 606–611 (2005).
57. Neylon, C. Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution. *Nucleic Acids Res.* **32**, 1448–1459 (2004).
58. Voigt, C. A., Martinez, C., Wang, Z.-G., Mayo, S. L. & Arnold, F. H. Protein building blocks preserved by recombination. *Nat. Struct. Biol.* **9**, 553–558 (2002).
59. Chen, T. & Romesberg, F. E. Directed polymerase evolution. *FEBS Lett.* **588**, 219–229 (2014).
60. Truppo, M. D. Biocatalysis in the pharmaceutical industry: the need for speed. *ACS Med. Chem. Lett.* **8**, 476–480 (2017).
61. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. Preprint at https://arxiv.org/pdf/1512.03385.pdf (2015).
62. Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. Preprint at https://arxiv.org/pdf/1502.03167.pdf (2015).
63. Maas, A. L. Rectifier nonlinearities improve neural network acoustic models. In *Proc. 30th International Conference on Machine Learning* Vol. 30 (ACM, 2013).
64. Mescheder, L., Geiger, A. & Nowozin, S. Which training methods for GANs do actually converge? Preprint at https://arxiv.org/pdf/1801.04406.pdf (2018).
65. Miyato, T., Kataoka, T., Koyama, M. & Yoshida, Y. Spectral normalization for generative adversarial networks. Preprint at https://arxiv.org/pdf/1802.05957.pdf (2018).
66. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
67. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
68. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/pdf/1412.6980.pdf (2014).
69. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
70. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
71. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* **27**, 135–145 (2018).
72. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

73. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
74. Berman, H. M. et al. The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.* **58**, 899–907 (2002).
75. Eswar, N. et al. Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Sci.* **2**, 2.9 (2006).
76. Sievers, F., Wilm, A., Dineen, D. & Gibson, T. J. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
77. Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
78. Sanner, M. F., Olson, A. J. & Spehner, J. C. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **38**, 305–320 (1996).
79. McCloskey, D. & Ubhi, B. K. Quantitative and qualitative metabolomics for the investigation of intracellular metabolism. *SCIEX Tech Note* 1–11 (2014).
80. Wilbur, W. J. & Lipman, D. J. Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl Acad. Sci. USA* **80**, 726–730 (1983).

Acknowledgements

We thank G. Stonyte, J. Nainys and C. Correia-Melo for comments on the manuscript. We also thank A. Repecka and L. Petkevicius for their valuable and constructive suggestions for improving the model. L.K. and R.M. were supported by the Agency for Science, Innovation and Technology (Lithuania) grant no. 31V-59/(1.78)SU-1687. J.Z. and A.Z. were supported by SciLifeLab fellow programme funding. S.V. was supported by VR starting grant no. 2019-05356. The computations were enabled with resources provided by the Swedish National Infrastructure for Computing (SNIC) at C3SE, partially funded by the Swedish Research Council through grant agreement no. 2018-05973. M. Öhman and T. Svedberg at C3SE are acknowledged for technical assistance in making the code run on Vera C3SE resources.

Author contributions

D.R. implemented the method, contributed with principal analysis and wrote the first draft. V.J. contributed principal analysis, designed experiments and wrote the first draft. L.K. contributed principal analysis, designed experiments and wrote the first draft. E.R. performed laboratory experiments. I.R. contributed principal analysis and wrote the first draft. J.Z. contributed principal analysis, performed laboratory experiments and wrote the first draft. S.P. performed laboratory experiments. A.L. contributed principal analysis. S.V. contributed principal analysis. W.A. performed laboratory experiments. O.S. contributed principal analysis and supervised the mass spectrometry work. R.M. supervised the study and designed the experiments. M.K.M.E. supervised the study, designed experiments, contributed principal analysis and wrote the manuscript. A.Z. supervised the study, designed experiments, contributed principal analysis, financed the experiments and wrote the manuscript. All authors contributed to writing of the paper and read the final manuscript.

Competing interests

L.K., V.J., D.R., I.R. and R.M. are shareholders of the company Biomatter Designs. The company has submitted a patent application for the technology described in the Article. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-021-00310-5.

Correspondence and requests for materials should be addressed to A.Z.

Peer review information *Nature Machine Intelligence* thanks Frances Arnold and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.