Pharmacological Research 227 (2026) 108159

Contents lists available at ScienceDirect

Pharmacological Research

journal homepage: www.elsevier.com/locate/yphrs

Self-organizing neural network-based generative AI with embedded error inflation control enhances effective knowledge extraction from preclinical studies with reduced sample size

Jörn Lötsch [a,b,c,*], Benjamin Mayer [a], Natasja de Bruin [b], Alfred Ultsch [d,1]

[a] Institute of Clinical Pharmacology, Goethe - University, Theodor Stern Kai 7, Frankfurt am Main 60590, Germany
[b] Fraunhofer Institute for Translational Medicine and Pharmacology ITMP, Theodor-Stern-Kai 7, Frankfurt am Main 60596, Germany
[c] University of Helsinki, Faculty of Medicine, Helsinki 00014, Finland
[d] DataBionics Research Group, University of Marburg, Hans-Meerwein-Straße 22, Marburg 35032, Germany

ARTICLE INFO

Keywords:
Preclinical research
Data science
Generative AI
And animal models and ethics

ABSTRACT

Small sample sizes in preclinical research limit the extraction of reliable knowledge and hinder translational progress. We propose genESOM, a generative artificial intelligence method based on emergent self-organizing maps. genESOM is designed to augment small biomedical datasets while controlling α-error inflation. It separates structure learning from data synthesis and integrates error propagation mitigation through dimensionality modulation, enabling safe and interpretable data augmentation. Using lipid signaling data from a preclinical multiple sclerosis study employing the experimental autoimmune encephalomyelitis (EAE) model (26 female SJL/J mice, three treatment groups, and 62 lipid mediators), we intentionally reduced the sample size from 26 to 18 animals. This reduction abolished detectable group differences by both statistical and machine learning analyses. Augmenting the reduced dataset with AI-generated cases restored treatment-specific segregation and recovered the original key lipid mediators. genESOM achieved consistent fidelity without introducing false positives. In contrast, Gaussian mixture and conditional GAN models failed under comparable constraints. These results demonstrate that genESOM provides a robust, error-controlled framework for enhancing knowledge extraction from limited preclinical samples. While synthetic augmentation cannot substitute for biological replication, it can support exploratory analyses and help reduce the need for additional animal experimentation.

1. Introduction

The translation of advances in basic science knowledge gained through animal models into new clinically effective compounds often falls short of expectations [1,2]. Initiatives to improve translation have promoted transparent reporting or rigorous statistical methodology, but they do not resolve the central challenge of extracting valid knowledge while minimizing animal use. Meta-analyses and pooled experiments can increase statistical power, yet they require many comparable studies and harmonized procedures across laboratories, conditions that are often not met.

Generative artificial intelligence (genAI) offers an alternative strategy [3]. By generating synthetic data to expand existing datasets, genAI can mitigate underrepresentation and small cohort sizes in biomedical research [4]. Typically, generative learning proceeds in two stages: a structure learner infers latent features and data structure, and a generator uses these learned structures to produce synthetic observations. Formally, its objectives are (i) to estimate the joint distribution $p(x|c)$ to generate new data points x and predict class labels c, and (ii) to sample additional instances and labels consistent with the inferred structure [3]. In small datasets, genAI can support protocol optimization, facilitate recruitment, and reveal hidden patterns through advanced data mining [5–7]. However, genAI is prone to propagate and amplify random errors in the data, to increase stochastic noise and inflate type I error rates [8–10].This restricts its use and requires genAI types for which this issue has been addressed.

The present study examines the hypothesis that a recently introduced generative method based on emergent self-organizing maps (genESOM)

* Correspondence to: Goethe - University, Theodor - Stern - Kai 7, Frankfurt am Main, 60590, Germany.
E-mail address: j.loetsch@em.uni-frankfurt.de (J. Lötsch).
[1] www.ulweb.de

https://doi.org/10.1016/j.phrs.2026.108159
Received 15 January 2026; Received in revised form 12 February 2026; Accepted 11 March 2026
Available online 13 March 2026

can reliably augment small preclinical datasets and extract valid knowledge from reduced data sets. genESOM is a special type of generative AI (genAI) that uniquely includes error propagation mitigation via dimensionality modulation as a method for safely augmenting biomedical datasets. Specifically, genESOM separates structure learning from data synthesis. This allows for dimensionality modulation and the injection of engineered diagnostic features (permuted counterparts of real variables) that serve as negative controls to monitor the stability of feature importance during data generation.

We use a robust preclinical lipidomics dataset in multiple sclerosis (MS) research [11], which has already been shown not to exhibit error inflation under genESOM-based augmentation [12]. We deliberately reduced the dataset until the main findings were no longer statistically significant and then augmented it with synthetic data. The objective was to determine the extent to which this augmentation can restore the ability to extract valid, reproducible knowledge. The here reported evaluations are based on the previous development of this specific type of generative AI, including comparative evaluations of its data generation abilities [4] and the implementation of error inflation stopping criteria [12].

2. Methods

2.1. Preclinical data set

The dataset has been described in detail elsewhere [11] and is freely available at https://data.mendeley.com/datasets/m2p6rr9v36/1 (DOI: 10.17632/m2p6rr9v36.1). It was derived from a comprehensive preclinical investigation exploring the regulation and functional significance of various lipid signaling mediators in multiple sclerosis in a translational context [13,14].

In brief, the dataset originates from a preclinical study in SJL/J mice using the relapsing-remitting experimental autoimmune encephalomyelitis (EAE) model to investigate neuroinflammation and the effects of fingolimod (FTY720), a sphingosine-1-phosphate (S1P) receptor modulator approved for multiple sclerosis [13]. Fingolimod targets S1P, a lipid mediator derived from sphingomyelin metabolism that plays a central role in G protein-coupled S1P receptor signaling. Quantitative lipidomic profiles were obtained from plasma and central nervous system tissues (cerebellum, hippocampus, and prefrontal cortex) and are described in detail elsewhere [15].

The study comprised 26 female SJL/J mice (8 weeks old), assigned to three groups: (i) controls without EAE ($n = 10$), (ii) EAE (n = 8), and (iii) EAE plus fingolimod treatment (n = 8; 0.5 mg/kg/day in drinking water, treatment initiation > 18 days post-immunization) [13]. Outcome measures included clinical scores assessing motor function, coordination, and social behavior, along with a lipidomics matrix of d = 62 lipid mediators quantified by targeted LC-MS/MS [15]. The 62 variables comprised lysophosphatidic acids (LPA: 16:0, 18:0, 18:1, 18:2, 18:3, 20:4), ceramides (C16, C18, C20, C24, C24:1), sphingolipids (C16 sphinganine, C18 sphinganine, C24 sphinganine, C24:1 sphinganine, sphingosine, sphinganine, sphingosine-1P, sphinganine-1P), and endocannabinoids (anandamide (AEA), palmitoylethanolamide (PEA), 1-AG, 2-AG, oleoyl-ethanolamide (OEA)). Lipid concentrations (ng/ML) were measured in plasma (24 variables), cerebellum (18), hippocampus (7), and prefrontal cortex (13), yielding an input data matrix of $n = 26$ mice × d = 62 lipid mediator concentrations across the three experimental groups.

2.2. Data analysis

2.2.1. Computational setup

Coding and data analysis were primarily conducted using the R language [16], version 4.3.1 for Linux, obtained from the Comprehensive R Archive Network (CRAN, https://CRAN.R-project.org/ [17]). Additional routines were implemented in MATLAB (version 9.13.0.2049777 (R2022b), MathWorks, Natick, MS, USA) and in Python [18], versions 3.10.12/3.11.6 for Linux, available free of charge at https://www.python.org. All computations were executed on an AMD Ryzen Threadripper PRO 7985WX 64-Core processor (Advanced Micro Devices, Inc., Santa Clara, CA, USA) running Ubuntu Linux 24.04.3 LTS (Canonical, London, UK).

2.2.2. Data preprocessing

2.2.2.1. Data transformation. Pharmacokinetic concentration data in plasma, blood, or tissue generally follow a log-normal distribution due to multiplicative biological processes, resulting in positively skewed values. There is clear, formally stated guidance from both regulatory agencies and expert consensus supporting log-transformation of PK variables unless strong evidence indicates otherwise [19–21]. Accordingly, log transformation was applied to all concentration data in this dataset.

2.2.2.2. Missing value imputation. The original mouse lipidomics data set contained 5.3% missing values. Imputation of missing values was performed as previously described [22]. The imputation method was selected after comparative testing [12], i.e., univariate and multivariate methods were compared using mean, median and mode imputation for the former and, among others, regression tree bagging as implemented in the R library "caret" (https://cran.r-project.org/package=caret [23]) and random forests [24,25] for the latter. Random forests was suggest form method comparison and used form the R library "missForest" (https://cran.r-project.org/package=missForest [26]).

2.2.3. Initial statistical analyses in the complete data set

Statistical analyses were performed to reproduce the key findings of the original study [14] as a starting point for data set reduction to the point of disappearance of reported results due to insufficient sample size. This included data transformation and statistical group comparisons as follows.

2.2.3.1. Statistical approach to treatment group differentiation. Group comparisons consisted of subjecting each of the d = 62 lipid marker variables to univariate analysis of variance with the factor "group" at three levels, i.e., no EAE, EAE, and EAE plus fingolimod. The statistical procedures used in the original publication [14] included additional grouping of variables by tissue and lipid class, which were then analyzed with separate two-way analyses of variance with the factors "group" as above and "lipid" comprising the actual lipid mediators analyzed in each tissue. However, this analysis mainly addressed the role of data completion and augmentation. However, the α-correction for multiple testing according to Šidák [27] was carried over from the original publication.

2.2.3.2. Machine learning-based feature selection. Machine learning approaches to subgroup differentiation emphasize generalizability. The primary objective of supervised learning is accurate prediction of subgroup membership for new, unseen samples, contrasting with traditional statistical methods that focus on characterizing whether group members arise from the same underlying distribution. Consequently, ML requires an independent validation dataset not overlapping with training data. In the absence of external data, it is common practice to split available samples into training and validation subsets, training algorithms on the former and evaluating predictive performance on the latter. While ML methods incorporate aspects of classical statistics and share the goal of detecting group differences, they differ fundamentally by focusing on predictive accuracy rather than solely testing distributional differences. The multivariate nature of supervised learning and emphasis on prediction underscore risks inherent to univariate predictor selection, which can introduce bias and is not consistently correlated

with predictive power [28,29]. Thus, ML aligns well with translational goals in preclinical research, although partitioning limited datasets into training and test subsets poses challenges when group sizes are small.

Treatment group discrimination within the lipidomics dataset was evaluated using three widely used classification methods: random forests, which are considered as robust tree-based bagging classifiers [24,25]; support vector machines (SVM), which utilize hyperplane separation [30]; and k-nearest neighbors (k-NN), a distance-based classifier [31,32]. Hyperparameter tuning, such as kernel choice for SVM and number of trees for random forests, was performed via grid search, with classifier training and validation implemented through nested cross-validation. Random forests enable estimation of feature importance through permutation of out-of-bag (OOB) cases [25]. This was extended by applying the "Boruta" method from the R package "Boruta" (https://cran.r-project.org/package=Boruta) [33], which employs cross-validation and Bonferroni-corrected p-values to unambiguously determine the importance of each variable.

2.2.4. Reduction of the group sizes to the level of non-significant results

The data set had been acquired in an in-house preclinical study in which sample sizes were selected according to rigorous case number estimation, resulting in group sizes of $n_{original} = [8, 10]$ exceeding the common practice of group sizes of six mice [34]. As noted by others [35], six animals per group is often considered an adequate sample size by some researchers, although the scientific and statistical bases of this perception are weak. Systematic reviews correct this perception by finding many preclinical studies with larger sample sizes [36,37]. Nevertheless, the lower boundary for valid generative augmentation in small-sample biomedical datasets remains insufficiently characterized. The resource equation method [38] suggests that error degrees of freedom (E = total animals minus total groups) should lie between 10 and 20, which for typical experimental designs with multiple treatment groups yields minimum group sizes in the range of $n = 5$ to 6 [35]. Furthermore, group-sequential designs with interim analyses at n = 6 and multiples thereof have demonstrated average savings of 20% in animal use without decreasing statistical power [39]. For the present evaluations, an iterative reduction approach was employed to empirically identify the threshold at which statistical significance disappeared. This threshold was found to be n = 6 per group, which served as the intentional design point for testing genESOM-based augmentation and aligns with the resource equation recommendations and group-sequential design practices described above.

To generate an appropriately reduced data set, in an iterative experiment, the data set was reduced stepwise to the first $n = [10, ..., 2]$ mice from each group, as if the actual experiments had been stopped with this number of mice. Two groups had only $n = 8$ mice, so the first data set was the original transformed data set, the second contained $n = 9$ controls and all EAE and EAE + fingolimod mice, the third contained $n = 8$ controls and all experimental autoimmune encephalomyelitis (EAE) and EAE + fingolimod mice, and from the fourth iteration on, the data set contained the first $n = [7, ..., 2]$ mice from each group. Missing value imputation was repeated on each iteration of the reduced datasets as described above. Statistical analyses in the simplified form described above were performed on the transformed and imputed data at each step to determine when statistical significance was lost.

2.2.5. Evaluation of key information extraction from biomedical data using genESOM

2.2.5.1. Generative AI driven data augmentation. Generative AI was applied to increase group size by leveraging the intrinsic structural properties of the dataset using emergent self-organizing maps (ESOM) to identify true structures in high-dimensional biomedical data [40,41]. The methods have been comprehensively described previously [4,12] and are therefore only briefly summarized here.

ESOM extends classical self-organizing maps (SOM) by arranging thousands of neurons on a two-dimensional toroidal grid and adding a third dimension encoding distances between subgroups or projection errors onto a 2D plane [42–44]. This preserves neighborhood relationships and detects cluster structures without imposing specific cluster shapes, outperforming many traditional methods [45]. The transformed, imputed, and scaled lipidomics data were projected onto the ESOM, where the learning update follows

$$\Delta w_i = \eta(t)\, h(bmu_i, r, t)(x_i - w_i) \quad (1)$$

where x_i is a data point, bmu_i is the closest neuron for x_i in the SOM (best matching unit, BMU), w_i denotes the weight vector of neuron n_i, $h(...)$ describes the neighborhood and $\eta(t) \in [0, 1]$ the learning rate, both of which decrease during learning [42,46]. The ESOM map consists of many neurons (e.g., ≥4000) arranged on the toroidal grid [43,45].

After training, the U-matrix encodes average distances between neuron prototypes, indicating cluster boundaries [44,47]. The P-matrix quantifies point density by counting points within hyperspheres of radius *r*around each neuron. It displays local densities, estimated as the number of data points in a hypersphere of radius *r* around each weight vector (w_i) of a neuron on the output grid of the ESOM.

$$P(n_{i,j}) = \left|\{x | d(w_{i,i} \leq r\}\right| \quad (2)$$

where $n_{i,j}$ is the neuron of the U-matrix at row *i* and column *j* with weight vector $w_{i,i}$.

For multivariate density estimation, the radius r is the critical parameter. Using ESOM, a suitable radius *r* can be found as follows: the abstract U-matrix heights (AU_heights) [44] are the subsets of data distances

$$\{AU\_heights\} \subseteq D = d(x_j, x_k) \text{ for all pairs of BMUs} \quad (3)$$

on edges of the Gabriel graph constructed from the BMUs. The distribution of AU_heights is modeled by a bimodal Gaussian Mixture Model (GMM) optimized by expectation maximization [44,48]. The critical radius *r*corresponds to the Bayesian decision boundary between the two Gaussians:

$$r = \arg\min_x \left[\pi_1 \mathcal{N}(x; \mu_1, \sigma_1^2) = \pi_2 \mathcal{N}(x; \mu_2, \sigma_2^2)\right] \quad (4)$$

Synthetic data generation occurs within neighborhoods defined by *r*, leveraging Bayesian modeling based on P-matrix densities to ensure validity and preservation of data structure [49]

The third step took advantage of the property of the ESOM projection that it is neighborhood preserving, i.e., data points that are close to each other in high-dimensional space are also close to each other in the projection. New data was generated in the neighborhood of a data point (seed) with respect to the distance of the generated point to the seed, which is well defined [49]. The generation used Bayesian statistics to model the decision of whether a new data point is to be expected, obtaining the probability of the existence of such a data point from the P-matrix, which shows the density of the data of the projection of the data set onto the ESOM. The bandwidth of the density estimate for the P-matrix can be estimated from the distribution of the distances in the U-matrix and is verified in the P-matrix visualization. This was used to generate valid new cases, based on the U-matrix/P-matrix analysis of the observed data.

Unlike many alternative approaches, genESOM separates structure learning from data synthesis and permits modulation of data dimensionality. This allows the injection of engineered diagnostic features, such as permuted counterparts of real variables, that act as negative controls to monitor feature importance stability during data generation. A data-driven stopping criterion terminates augmentation when error inflation emerges, thereby limiting overfitting. Thus, to control error inflation in data augmentation, genESOM integrates a stopping criterion based on dimensionality modulation and engineered negative control

features [4,12]. Permuted variables serve as diagnostics to monitor feature importance stability during bootstrap feature selection, establishing an error threshold $L_\Delta$on the difference in selection frequencies between original and permuted features. Synthetic data are generated incrementally within radius *r*, while the importance difference

$$\Delta I_j = I(X_j) - I(X_j^{\text{perm}}) \tag{5}$$

is evaluated for each variable *j*. Augmentation halts if $\Delta I_j > L_\Delta$, signaling overfitting and preserving the statistical validity of the expanded datasets [12,50]. In the "mouse_lipidomics_data" dataset in [12], which is exactly the data set used in the present report, this stopping criterion has resulted in halting the augmentation after one new data point per original. This was therefore followed in the present analyses and no repeated here. Reassessment of key informative variables for study group differentiation

2.2.5.1.1. Statistical approach to treatment group differentiation on augmented data. To evaluate the impact of data augmentation on key variable identification, we repeated the statistical analyses originally performed on the full dataset, now applied to the augmented dataset. Group comparisons involved univariate analyses of variance (ANOVA) conducted on each of the d= 62 lipid marker variables with the factor "group" at three levels: no EAE, EAE, and EAE plus fingolimod. Again, the Šidák correction for multiple testing was retained from the original study to control family-wise error rates [27].

2.2.5.1.2. Machine learning-based feature selection on augmented data. This expansion motivated complementing traditional statistical methods with machine learning (ML)-based feature selection applied to the augmented dataset. Supervised feature selection techniques, including regularization methods and tree-based models, were used to rank and select variables with the highest predictive relevance for distinguishing among groups. This approach provided an independent, data-driven confirmation of key informative features after augmentation. It was performed as described above for the complete data set.

2.2.6. Comparative evaluation of alternative data generative approaches

The generative emergent self-organizing map (genESOM) AI is unique in its ability to allow dimensionality modulation between the structural detection phase and the data generation phase, thereby enabling control of alpha error inflation [12]. This capability sets it apart from other generative AI methods that lack integrated error control mechanisms. Consequently, a direct quantitative comparison of alternative generative methods in terms of error inflation was not feasible. However, prior work has demonstrated that Gaussian mixture models (GMMs), a commonly used generative approach, are also prone to error inflation [12]. Similar risks are likely to affect other generative AI methods as well. Moreover, the structure-preserving data generation capabilities of genESOM have been extensively compared with alternative generative AI types previously and will therefore not be repeated in this report.

However, within the scope of this report, we compared alternative generative AI methods by evaluating their ability to recover statistical results originally obtained from the full, original dataset. We employed genESOM as a reference method due to its error control features, assessing how well each method reproduced key statistical findings post-augmentation. This comparative approach provides insight into the reliability and validity of different data augmentation techniques concerning preservation of original dataset characteristics and inferential integrity.

A selection, not intended as exhaustive, of alternative methods included Gaussian mixture models (GMM), generative adversarial networks (GAN), and autoencoders. They were applied to the same reduced dataset. The results, including false positives and false negatives as defined by the originally published hits [22], were compared with those obtained from ESOM-based data generation, which served as the primary method in this report.

Gaussian mixture models (GMM) are probabilistic models that assume that all data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. The expectation maximization (EM) algorithm is typically used in GMM to optimize the maximum likelihood; however, implementations of alternatives such as Markov chain Monte Carlo (MCMC) are available (summarized in [51]). The simplest form of GMM is the Bayesian approach [52], which is well established. Two variants of data generation algorithms based on Gaussian distributions have been used, i.e., independent Gaussian (IG) and multivariate Gaussian (MG). The independent Gaussian (IG) variant is also used in the so-called “DataBoost” algorithm that uses data generation to increase classification performance [53]. For each variable *i* in the data set and class *k*, the distribution is modeled as a Gaussian.

$$G(\mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(\nu-\mu_k)^2}{2\sigma_k^2}} \tag{6}$$

Data generation in IG is performed for each variable and class separately by drawing data from the model distribution. By contrast, multivariate Gaussian (MG) uses a multivariate normal distribution given as

$$N(\mu, \Sigma) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)}{\sqrt{(2\pi)^d|\Sigma|}} \tag{7}$$

with mean $\mu = (\mu_1,\ldots,\mu_d)$ representing a *d*-dimensional mean vector, where d denotes the number of variables in the data set and Σ denotes a d x d sized covariance matrix

$$\Sigma_{i,j} = E[(x_i - \mu_i) * (x_j - \mu_j)] \tag{8}$$

The D-dimensional normal distribution is constructed for each class using the EM method [48], using a MATLAB implementation described in [54]. The generated data was then drawn from the multivariate distribution for each class.

Generative adversarial networks (GANs) were considered as another alternative method, as they are currently widely used in the machine learning field. However, their conventional application to image datasets raised skepticism about their efficacy for augmenting an exceptionally small preclinical dataset that is organized in tabular form. In essence, a GAN consists of a generator function that produces synthetic data that is identified as true or false by a discriminator function until the distinctions between real and synthetic data become indistinguishable [55]. To specifically address the GAN model to a tabular data format, a Conditional Tabular GAN (CT-GAN) was applied to a reduced mouse dataset (n = 6 per condition/group) running on an NVIDIA GeForce GTX 1050 GPU (NVIDIA Corporation, Santa Clara, CA, USA) using the Compute Unified Device Architecture (CUDA) version 12.0. Scripting of the CT-GAN workflow was established based on the CTGAN-package basic tutorial for Python [56]. The model was trained based on 20 epochs, sampling for generated 1000 rows. The generated data was sampled at a size of n = 12 per group, consistent with the parameters used for other generative algorithms. The results of this generation are shown in Supplementary Figure 1.

3. Results

The main hypothesis addressed in this investigation was that key group-difference drivers (relevant "hits") identified in a well-powered preclinical dataset may become undetectable when the sample size is reduced, but can be recovered through data augmentation using genESOM. To test this, we assessed whether variables significant in the full original dataset would disappear after reduction and subsequently re-emerge following generative augmentation.

Additionally, to explore the potential benefits of machine learning approaches, particularly their ability to incorporate generalizability

assessment into the analysis pipeline, we applied supervised machine learning models to both feature selection and classification tasks on the augmented dataset. This allowed for a complementary evaluation of variable informativeness and the robustness of group discrimination beyond traditional statistical analysis.

3.1. Identification of key modulators of group differences in the original dataset

The statistically significant lipid mediators identified in the full dataset are consistent with previously reported dysregulation of lysophosphatidic acids and sphingolipids in experimental autoimmune encephalomyelitis and multiple sclerosis. These lipid classes are known

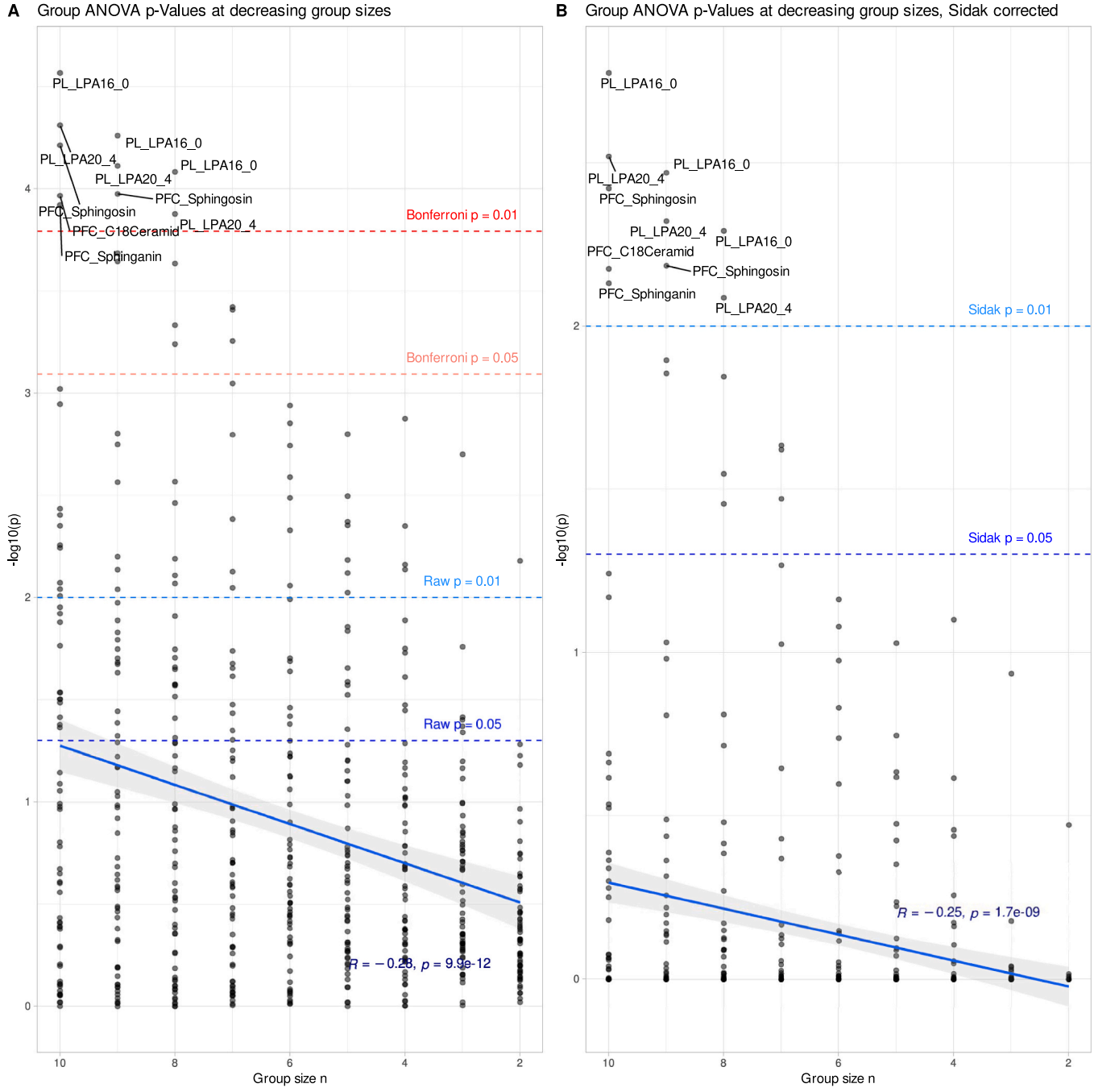


Fig. 1. Significant lipid mediators observed with different group sizes after analysis of variance of d = 62 log-transformed lipid mediator concentrations using the factor "group" with three levels, i.e., no EAE, EAE, and EAE plus fingolimod. In an iterative experiment, the data set was reduced stepwise to the first $n = [10, \ldots, 2]$ mice from each group. Two groups had only $n = 8$ mice, so the first data set was the original transformed data set, the second contained $n = 9$ controls and all EAE and EAE + fingolimod mice, the third contained $n = 8$ controls and all EAE and EAE + fingolimod mice, and from the fourth iteration on, the data set contained the first [7,…,2] mice from each group. **A:** Significant effects observed in successively smaller sample sizes per group (see above). P-value thresholds, either raw or α-corrected according to Bonferroni [57], are drawn as dashed horizontal lines. The names of significant lipids are given as text annotations to the respective p-values. **B:** Similar analysis as in panel A, but with p-value correction according to Šidák [27] to be consistent with the original publication where this was applied [14]. The figure was generated using the R software package. (version 4.3.1 for Linux; https://CRAN.R-project.org/ [17]) and the R library "ggplot2" (https://cran.r-project.org/package=ggplot2 [88]).

to participate in immune cell trafficking, neuroinflammation, and sphingolipid signaling pathways relevant to disease progression and therapeutic response. Thus, the present analysis serves as a biological reference point, confirming that the dataset captures established disease-relevant lipid signatures prior to any reduction or augmentation procedures.

3.1.1. Statistical analysis of group differences

In a straight-forward statistical group comparison consisting of 62 univariate analyses of variance using the three treatment groups as factor levels, significant treatment group effects were detected in d = 27 lipids at $p < 0.05$, and in $d = 15$ lipid wen setting the a-threshold to $p < 0.01$, corrected for multiple testing, either according to Bonferroni [57] or to Šidák [27]. Furthermore, the most significant group differences, if the lowest p-values are taken as criterion, were observed in $d = 5$ lipids ($p < 0.01$ Bonferroni corrected), two of which were lysophosphatidic acids, repeating the main finding of the original analysis [14] (Fig. 1).

3.1.2. Supervised machine learning-based feature selection

On the original dataset with group sizes of $n = [8, 10]$, random forests were successful in classifying the cases of the respective test data subsets with a balanced accuracy (BA) [58] better than chance, i.e., BA > 0.5 or 50% with the 95% confidence interval not including the guessing level values. This was obtained in a 1000 cross-validation scenario using class-proportional splits of the original data set into training and test subsets (Fig. 2 A).

3.2. Impact of sample size reduction on detection of group differences

3.2.1. Loss of statistical significance after reduction

Iterative reduction of group sizes to the first n mice indicated that at $n_{reduced} = 6$ mice per group, no variables remained that showed group differences at p values with $p < 0.01$ or $p < 0.05$ that passed the Bonferroni or Šidák α-correction procedures (Fig. 1). The ANOVA significances expressed as -log10(p) values showed weak but statistically significant product moment correlations [59] with the number of mice per group, with values of r = -0.28 and −0.25 for raw and Šidák corrected p-values, respectively, which can be interpreted as a "medium" correlation according to a recently proposed rule of thumb for effect size interpretation [60].

3.2.2. Impaired performance of supervised algorithms

When the dataset was reduced to the first n = 6 mice per group, the task could no longer be successfully accomplished, despite retuning the random forest algorithm for the now smaller dataset.

The loss of statistical significance and classification accuracy observed after reducing group sizes does not indicate the disappearance of underlying biological effects, but rather reflects reduced power to detect them. Importantly, the affected lipid mediators remain biologically plausible and mechanistically linked to EAE pathology, suggesting that their non-significance at smaller sample sizes represents a limitation of inference rather than an absence of disease-related regulation. This mirrors common challenges in preclinical research, where biologically meaningful effects may be obscured by practical constraints on animal numbers.

3.3. Recovery of informative modulators in augmented datasets

3.3.1. Structure detection using ESOM neural networks at reduced group sizes

The data from the first $n_{reduced} = 6$ mice per group, i.e., the number of cases in which the statistical results for the entire cohort fell below the alpha-corrected p-value threshold of 0.05 as determined above, were projected onto the $\mathbb{R}^2$ plane using an ESOM after scaling by percentage transformation (Fig. 2 A). The plane consisted of 50 × 80 artificial neurons on a toroidal grid. After training the artificial network in 20 epochs with decreasing learning rates from 0.3 to 0.05 and using a Gaussian neighborhood function, the distances between neurons representing a prototype were computed. The projection showed a tendency for the three treatment groups to separate, consistent with the structure of the data groups, but not perfectly, as also consistent with the reduced statistical significance of the group differences in this reduced data set. The P-matrix displays local densities, estimated as the number of data points in a hypersphere of radius r around each weight vector (w_i) of a neuron on the output grid of the ESOM (Fig. 2 B). It provided a two-dimensional representation of the multivariate density function of the data space.

3.3.2. Generative data augmentation based on detected structures

The EM algorithm suggested a radius of 230 as best suited for data density estimation (Fig. 2 C), which was converted to the untransformed data space using linear regression of the data sets' distances on a Shepard plot (Fig. 2 D) [61,62]. The radius of r = .3559 obtained during the ESOM analyses of the reduced preclinical data set was then used to generate new data from the lipidomics data of the first $n_{reduced} = 6$ mice per group. The generated data set size was $2 \bullet n_{reduced} = 12$ per group, i.e., for each original mouse of the reduced data set, an additional data point was created, doubling the size of the reduced dataset to $n_{generated} = 36$ mice with $k = 3$ groups.

The partial separation of treatment groups observed in the reduced dataset indicates that biologically relevant structure persists even when classical statistical significance is lost. This suggests that disease- and treatment-associated lipid patterns remain embedded in the data, albeit in a form that is difficult to extract using standard univariate analyses. From a biological perspective, this indicates that potentially meaningful molecular differences are not necessarily absent in small cohorts, but may be more difficult to identify reliably using standard univariate approaches, motivating cautious exploratory approaches aimed at stabilizing their detection rather than dismissing them outright.

3.3.3. Reappearance of statistical significance after augmentation

The ESOM U-Matrix/P-Matrix-based data generation method, genESOM, was then applied to the transformed and imputed lipidomics data of the first $n = 6$ mice of each group according to the above results. This succeeded in recovering the main results observed in the whole cohort in the generated data, i.e., the five lipid mediators LPA 16:0 and LPA 20:4 in plasma and sphingosine, sphinganine and C18 ceramide in prefrontal cortex. These were the most significant variables in the full data set that did not pass the α-corrected p-value threshold of $p < 0.01$ in the reduced data set, but passed again in the generated data (Fig. 2). However, a few additional hits were generated, all of which were in the upper range of statistical significance among the mediators in the original full data set.

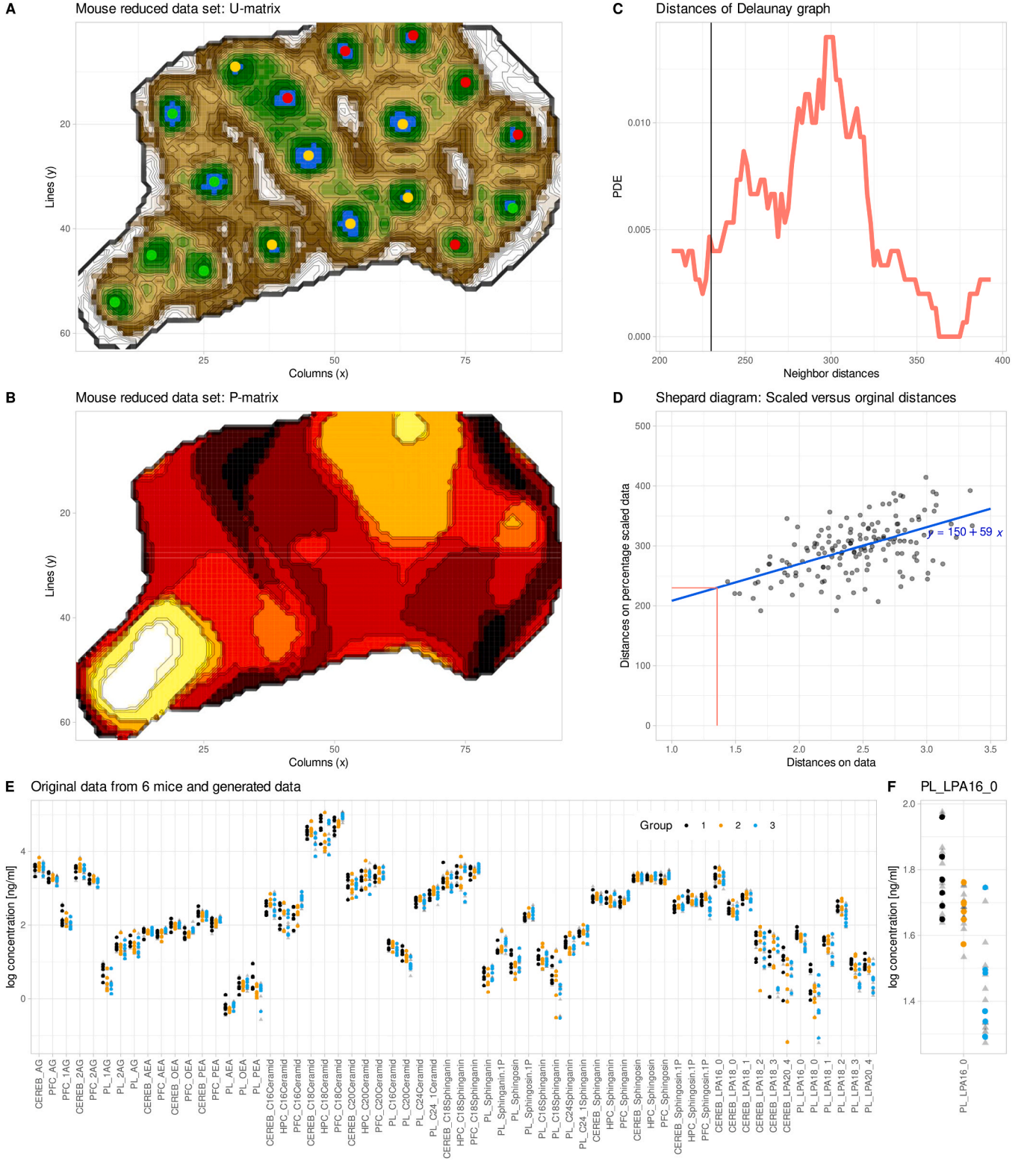
The re-emergence of the same lipid mediators that were significant in the full dataset, particularly lysophosphatidic acids and sphingolipid species, indicates that generative augmentation stabilized detection of known disease-associated signals rather than introducing novel or biologically implausible findings. Importantly, no entirely new lipid classes emerged as dominant drivers, supporting the interpretation that augmentation reinforces existing biological structure rather than creating artificial effects. These results therefore suggest recovery of attenuated biological signals, not the discovery of new mechanisms.

3.3.4. Restoration of successful machine learning group assignment

On the original dataset with group sizes of $n = [8, 10]$, random forests were successful in classifying the cases of the respective test data subsets with a balanced accuracy (BA) [58] better than chance, i.e., BA > 0.5 or 50% with the 95% confidence interval (CI) not including the guessing level values. This was obtained in a 1000 cross-validation scenario using 50/50 random class-proportional splits of the original

data set into training and test subsets (Fig. 2).

When the dataset was reduced to the first n = 6 mice per group, the task could no longer be successfully accomplished, despite retuning the random forest algorithm for the now smaller dataset. Running the algorithms on ESOM-based generated data with group sizes of n = 12 each, as used in the statistical analysis reported above, restored the ability of random forests to classify the cases of the test data subsets with class-wise accuracy better than chance. This made the dataset accessible for machine learning based feature selection. The retained variables included all the final hits from the statistical analysis, but several more,



(caption on next page)

Fig. 2. Emergent self-organizing maps (ESOM) based generation of valid data from the lipidomics dataset with reduced sample size of $n = 6$ mice per group (total: $n = 18$ mice). **A:** ESOM projection of the data set showing a 3-dimensional U-matrix visualization of the distance-based structures of the serum concentrations of d = 62 lipid mediators after projection of the data points onto a toroidal grid of 4000 neurons, where opposite edges are connected. The subgroups are separated by "snowy mountain ranges". The dots represent the so-called "best matching units" (BMU), i.e. neurons on the grid that, after ESOM learning, carried a data vector that primarily resembled a data vector of a sample in the data set. They are colored according to the prior group structure into no EAE, EAE, and EAE plus fingolimod. The U-matrix projection was performed with percentage scaled log-transformed imputed data. **B:** P-matrix associated with the U-matrix shown in panel A, representing local densities estimated as the number of data points in a hypersphere of *r*adius *r* around each weight vector (w_i) of a neuron on the output grid of the ESOM. **C:** Selection of the radius of the hyperspheres for density estimation based on a proposal obtained from the probability density distribution of the distances between data points. The panel shows the distribution of the distance between neighboring BMUs, measured as the length of the edges in the Delaunay graph. The vertical line marks the distance limit at which the class assignment of a BMU does not interfere with a neighboring class, as obtained by fitting a Gaussian mixture model to the density function using the expectation maximization (EM) algorithm [48]. **D:** Shephard plot [61] showing the distances in the original data versus the percentage scaled data to transform the radius obtained in panel C back to the original logarithmic data space (red lines). **E:** Generation of data from reduced mouse group sizes ($n = 6$ per group) and restoration of statistical significance of top hits in lipid mediators. Original (full color dots) and generated (semitransparent color dots) values of d = 62 log-transformed lipid mediator concentrations measured in four different tissues including plasma ($d = 24$ lipids), cerebellum ($d = 18$ lipids), prefrontal cortex ($d = 13$ lipids), and hippocampus ($d = 7$ lipids). **F:** Enlarged example showing the original and generated data of the concentration of the top hit of the original analysis [14], i.e., lysophosphatidic acid 16:0. The figure was generated using the R software package.(version 4.3.1 for Linux; https://CRAN.R-project.org/ [17]), the R library "ggplot2" (https://cran.r-project.org/package=ggplot2 [88]), and our R library and "Umatrix" (https://cran.r-project.org/package=Umatrix [65]).

mainly from the LPA class of lipids, consistent with the original findings.

Improved classification performance following augmentation reflects enhanced consistency of biologically meaningful lipid patterns across samples, rather than the introduction of independent biological variability. From a preclinical standpoint, this indicates that augmentation may help clarify whether an observed molecular trend is coherent and reproducible within the confines of the original experimental system. However, this should not be interpreted as evidence of increased biological diversity or replacement of biological replication.

3.4. Comparative effectiveness of alternative generative approaches

Of the alternative methods examined (Fig. 2), the multivariate Gaussian mixture-based generation yielded fewer statistically significant results compared to the original data set, achieving significance only at the α-corrected p-value threshold of $p < 0.05$. The independent Gaussian mixture-based generation restored original significance at the more stringent α-corrected p-value threshold of $p < 0.01$. In contrast, the independent Gaussian mixture-based generation restored original significance at the more stringent α-corrected p-value threshold of $p < 0.01$ but introduced additional variables with increased significance. Notably, some of these newfound significances occurred in variables that originally had higher p-values, suggesting a potential occurrence of false positives from the GMM-based generation. For comparison, both GMM-based methods introduced variables with lower original significances than those generated by the ESOM-based approach. Conversely, the CT-GAN-based generation did not produce statistically significant results at the α-corrected p-value threshold of $p < 0.01$, failing to reproduce the original significant findings observed in the full dataset. Due to the higher rates of false positives or false negatives compared to the primary ESOM-based method in the case of the GMM based generative models or the inability to restore original results in the case of the CT-GAN, the alternative methods were not further explored in the same depth as the main method.

When comparing different data augmentation methods, the ESOM-based approach most reliably reproduced the biologically meaningful lipid signals observed in the full dataset without introducing implausible new findings. In contrast, Gaussian mixture-based methods either recovered fewer significant results or generated additional signals that were less consistent with the original data, raising concerns about potential false positives. The CT-GAN approach failed to restore the key findings altogether.

Collectively, these findings indicate that generative augmentation can stabilize detection of established biological signals under constrained sample sizes, while not creating new biological effects or substituting for experimental replication.

4. Discussion

4.1. Generative learning for enhancing knowledge in small preclinical datasets

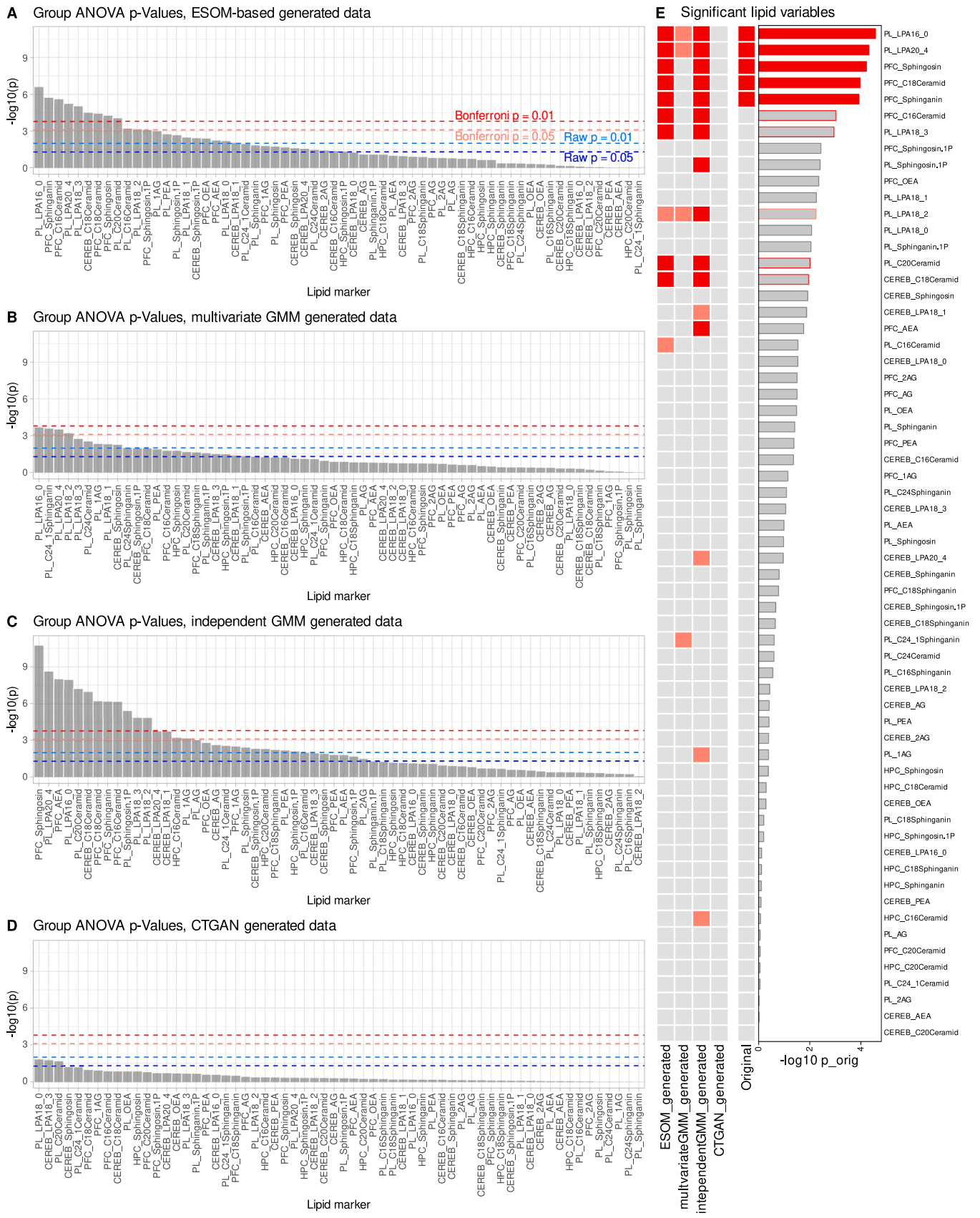
The core hypothesis, i.e., that generative learning can improve knowledge discovery from small preclinical datasets, is supported, with the caveat that applications must avoid overfitting. Generative data augmentation emerged as a viable strategy to recover signals lost due to limited sample sizes. A neural network approach based on ESOM has shown promise in improving the extraction of valid knowledge from small-sample preclinical studies. Applying this to a robust mouse model dataset (group sizes n = 8… 10) revealed statistically significant regulation of lysophosphatidic acid and other lipids [14]. Simulating smaller samples by reducing the dataset to $n = 6$ led to a loss of statistical significance. However, applying the generative algorithm restored these key findings, confirmed by appropriate statistical tests and machine learning-based assessments of treatment group differences.

4.2. Contribution to animal use reduction and integration with statistical methods

Generative AI-based data augmentation aligns with the 3Rs principles, replacement, reduction, and refinement, by potentially reducing animal usage without compromising scientific insight. In constrained settings where raising sample sizes is not feasible, these methods expand available research options validly.

The proposed workflow deploys machine learning throughout the entire process: starting with data imputation, progressing to unsupervised generative data augmentation via emergent self-organizing maps (genESOM), and concluding with supervised machine learning-based feature selection to identify variables relevant for treatment group segregation. This multistep process first uses neural network-based AI to analyze dataset structure, extracts essential information for data generation, augments the dataset accordingly, then tests classifier performance on held-out cases. Importantly, it focuses on ensuring identified informative features do not reflect importance inflation caused by the generative process itself.

The increasing use of machine learning models like random forests in small-sample biomedical analyses complements generative augmentation workflows. These algorithms effectively capture nonlinear complexity and provide reliable feature importance estimates to monitor overfitting [63,64]. Nevertheless, traditional statistical approaches, such as regularization, model simplification, cross-validation, and newer algorithmic techniques, remain essential alternatives for controlling overfitting [9].



(caption on next page)

Fig. 3. Significant effects observed in generated data with group sizes of $n = 12$ mice per group, obtained in statistical analyses of d = 62 log-transformed lipid mediator concentrations of four classes measured in four different tissues, including plasma ($d = 24$ lipids), cerebellum ($d = 18$ lipids), prefrontal cortex ($d = 13$ lipids), and hippocampus (d = 7 lipids). **A:** P-values obtained in the ESOM-based generated data (decreasing order of significance) of the group comparison in the generated data ($n = 12$ per group) performed as an analysis of variance (ANOVA) with the factor "group" at three levels, i.e., no EAE, EAE, and EAE plus fingolimod. P-value thresholds, either raw or α-corrected according to Bonferroni [57], are shown as dashed horizontal lines. **B:** Similar to panel A, but on data generated by a multivariate Gaussian mixture model. **C:** Similar to panel A, but on independent Gaussian mixture model generated data. **D:** Similar to panel A, but on data generated using a Conditional Tabular GAN (CTGAN). **E:** Summary of significant hits in lipid mediators using a significance criterion of $p < 0.01$ (red) or $p < 0.05$ (light red) Bonferroni α-corrected. The original top hits are shown in red in the right column of the matrix. The bar graph on the right shows the significance level (-log p-value) obtained in the original complete data set, i.e. [8,10] group size. The columns are filled according to the original significances. The borders of the columns are colored according to the majority vote of significance across the four generative methods. The figure has been created using the R software package (version 4.3.1 for Linux; https://CRAN.R-project.org/ [17]) and the R libraries "ggplot2" (https://cran.r-project.org/package=ggplot2 [88]) and "ComplexHeatmap" (https://www.bioconductor.org/packages/ComplexHeatmap/ [89]).

4.3. Cautions and limitations in data augmentation

4.3.1. Expert oversight and workflow requirements

The proposed workflow is not fully automated and requires expert oversight at critical decision points. The core computational components, ESOM/U-Matrix and genESOM, are implemented in our R package "Umatrix" (https://cran.r-project.org/package=Umatrix [65]), which provides fundamental functionality for structure detection and generative augmentation. Successful application requires careful calibration and expert supervision at multiple analytical stages and rigorous preprocessing of raw data before dimensionality reduction and structure detection, both of which fall outside the scope of the core algorithm. Since the purpose of data augmentation is to reduce animal usage while maintaining scientific validity, and failure to preserve biological signal integrity would render even reduced animal numbers ethically unjustifiable, this approach should be applied exclusively by researchers with expertise in the biological system under investigation and training in data science methodology to perform the preprocessing, validation of emergent data structures, and critical interpretation of augmented results.

Whether a fully automated solution is viable remains to be determined. For more constrained analytical scenarios, visually guided preprocessing platforms, such as our R library "pguIMP" (https://cran.r-project.org/package=pguIMP [22]) for cross-sectional tabular lipidomics data or the "MetaboAnalystR" package for metabolomics workflows (https://github.com/xia-lab/MetaboAnalystR [66]), demonstrate that user-friendly environments can provide comprehensive data handling. Development of similarly robust frameworks specifically designed for generative augmentation in the context of reduced animal experimentation may therefore be feasible but exceeds the scope of the present report.

4.3.2. Statistical considerations and error control

Synthetic observations generated by genESOM are not statistically independent biological replicates and should not be interpreted as such. They represent structure-preserving interpolations within the observed data manifold that can stabilize downstream analyses by increasing the effective sample density in regions where the original data are sparse. Statistical inference on augmented datasets must be interpreted conditionally, recognizing that synthetic data do not increase the true degrees of freedom available for hypothesis testing.

Generative models can only learn from observed data structure; extrapolation beyond the sampled distribution leads to hallucination [67]. As discussed previously [12], critical evaluation of dataset representativeness is essential when planning experiments that combine genESOM-based augmentation with reduced sample sizes. If the true biological population comprises distinct subpopulations in a 70:25:5 ratio but experimental sampling captures only 10 observations, low-frequency subpopulations may remain unobserved. Augmentation cannot recover unsampled biological variation and will instead interpolate only within the structure present in the observed data. While this sampling limitation is not unique to generative AI, researchers employing this framework for sample size reduction must explicitly consider whether their experimental design adequately samples the expected biological heterogeneity.

Alpha error inflation is a major concern in generative augmentation, as it may yield false positive results [68]. These risks are amplified in preclinical research with small sample sizes where animal reduction is a goal. Effective augmentation improves both model fairness and predictive power, but noise amplification and false discovery must be actively prevented [10]. Prior work with the "mouse_lipidomics_data" dataset, as the present data set was named in [12], demonstrates that the extent of error inflation scales with effect size and that careful calibration can keep augmented data within validity thresholds using random forest out-of-bag analysis [12]. In that study, safe augmentation was identified as adding one synthetic data point per original observation, which was the ratio applied in the present assessments. Excessive augmentation increases the risk of overfitting and false discoveries, necessitating strict stopping criteria, conservative augmentation ratios, and the error inflation control mechanisms unique to genESOM [12]. Naive or excessive augmentation, such as increasing sample size tenfold, substantially increases the likelihood of overfitting and false discoveries. Overly simplistic strategies such as adding random noise or indiscriminate oversampling can worsen the curse of dimensionality, fueling spurious findings as an issue to which biomedical datasets are particularly prone [69]. genESOM mitigates these risks by organizing synthetic data via emergent self-organizing maps, ensuring new data are placed within the correct biological topology and minimizing class boundary distortion [4].

Empirical evidence from artificial and biomedical datasets demonstrates that a one-to-one augmentation ratio (one synthetic data point per original observation) preserves variable ranking and statistical stability. This ratio has been systematically validated using the "mouse_lipidomics_data" dataset [12] and others, where it maintained strong negative correlations between statistical significance and feature importance (Kendall's tau range: −0.53 to −0.85) without signs of error inflation. This aligns with independent recommendations to limit synthetic data addition to approximately one-to-one to ensure reproducibility [68]. The optimal ratio of synthetic to original data remains uncertain beyond conservative one-to-one augmentation. Modest augmentation may still yield stability, whereas higher ratios markedly increase the risk of overfitting and false discoveries, as detected by genESOM's diagnostic features.

4.4. Scope and limitations

Despite its advantages, generative AI cannot replace essential experimental design steps, subgroup analysis, or thorough outlier detection. Minimum practical sample sizes for valid augmentation are underexplored. In the present study, n = 6 per group represented the empirically determined threshold at which statistical significance was lost in the original dataset, and successful augmentation at this sample size suggests it may represent a practical lower boundary for genESOM application, though smaller datasets may occasionally be adequate with clear statistical power justification. Augmentation of small datasets cannot recapitulate full biological diversity or synthesize novel

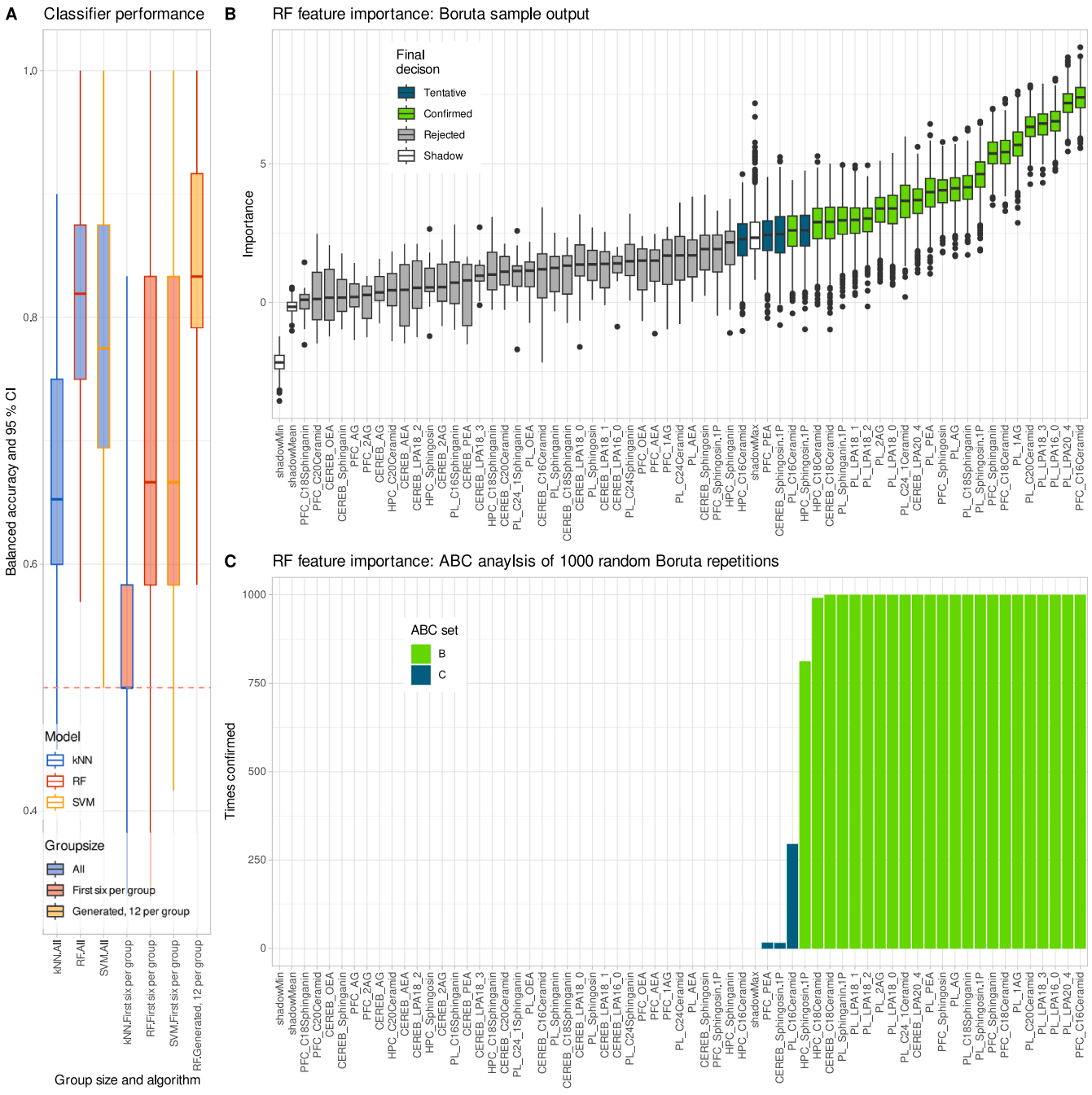


Fig. 4. Supervised learning approach for lipid mediators predicting group assignment to no EAE, EAE, or EAE plus fingolimod. **A:** Classifier performance expressed as balanced accuracy of three different models (kNN: k-nearest neighbors, RF: random forests, SVM: support vector machines) obtained in a 1,000-fold cross-validation scenario on 50/50% class-proportional random splits of the dataset into training and test subsets. The blue boxes show the performance on the original dataset with group sizes [8,10]. Only random forests could be trained successfully, so the 95% confidence interval (CI) does not include the 0.5 (=50%) confidence level. The red boxes show the same when applied only to the 1st $n = 6$ mice per group. The yellow box shows the random forest classifier when trained and tested on the generated data ($n = 12$ mice per group). The 95% CI is above the 0.5 guessing level. The boxes show the 25th, 50th, and 75th percentiles of the ranks across datasets and cluster quality measures obtained in the 1000 replicate runs. The whiskers span the 95% confidence interval from the 2.5th to the 97.5th percentile. **B:** Results of the variable selection procedure performed as a random forest based Boruta approach to identify the most informative features for class assignment [33]. This uses the random forests out of back (OOB) permutation importance and evaluates the importance measure of each variable based on the decrease in classification accuracy due to random permutation of values in a 1000-fold cross-validated approach. The importance measure is calculated separately for all trees in the forest that use that feature for classification. The mean and standard deviation of the loss of accuracy are then calculated and the z-score is used to compare against an external reference, the so-called "shadow" features (empty boxes), obtained by permuting the values of the original feature. Light green boxes represent "confirmed" significant features, while dark green and gray boxes represent "tentatively" significant and confirmed non-significant features, respectively. The boxes follow the standard construction rules for box plots, i.e., they were constructed using the 25th, 50th, and 75th percentiles of the values, but except in panel A, the whiskers add 1.5 times the interquartile range (IQR) to the 75th percentile or subtract 1.5 times the IQR from the 25th percentile. **C:** Summary of the 1000-fold cross-validated Boruta-based feature selection. The bars show the number of identifications of each variable as confirmed significant for class assignment. The coloring is according to an item categorization implemented as a computed ABC analysis (cABC), which places the irrelevant variables in subset "C" (dark green bars), while those that are relevant are placed in ABC subset "B". The figure has been created using the R software package (version 4.3.1 for Linux; https://CRAN.R-project.org/ [17]) and the R libraries "ggplot2" (https://cran.r-project.org/package=ggplot2 [88]) and "Boruta" (https://cran.r-project.org/package=Boruta [33]).

biological mechanisms. Generative AI expands the observed data distribution rather than creating entirely new phenomena, contrasting with biological simulators (e.g., brain models) [70], and emphasizing the need for cautious interpretation of augmented results.

The generalizability of the genESOM framework to additional data modalities remains an open question. To date, validation has been restricted to lipidomics and selected biomedical datasets, and performance in other domains such as transcriptomics, proteomics, imaging, or behavioral phenotyping may require further dedicated evaluation. Nevertheless, in the experiments with genESOM published so far [4,12,71], a range of artificial and biomedical data problems has already been covered, and genESOM has always been successful. Consequently, the present findings should not be extrapolated beyond comparable data structures without further empirical validation. Generative augmentation does not replace fundamental principles of experimental design, including appropriate subgroup stratification, outlier detection, and biological replication. The effectiveness of the approach in highly heterogeneous or noisy datasets has not been systematically evaluated. Traditional experimental safeguards remain essential, and augmented analyses should be interpreted as complementary rather than substitutive.

4.5. Comparison with alternative generative methods

The generative emergent self-organizing map (genESOM) AI is notable for allowing dimensionality modulation between structure detection and data generation phases, enabling direct control of alpha error inflation [12]. This feature distinguishes genESOM from other generative AI approaches that lack built-in error control.

A couple of alternative generative AI approaches were evaluated. Gaussian mixture models (GMMs), including independent GMMs and "DataBoost" [53], exhibited mixed performance: ESOM recovered all original significant hits and some extras; multivariate GMM captured only the top two hits and missed others, occasionally resulting in false positives; independent GMM inflated the number of significant variables beyond what was originally observed. Generative adversarial networks (GANs), although effective for image augmentation tasks [72], failed to capture group structure from very small samples. While GANs generated data within the original variable ranges (see Supplementary Figure 1), group differences were faint and statistical significance was not restored in reduced datasets. Top hits from the ceramide class were somewhat captured, but GAN results remained unconvincing for restoring significance and are generally unsuitable for small-sample learning. This failure likely reflects the insufficient sample size for GAN architectures rather than algorithmic limitations, as the same CTGAN implementation produced satisfactory results on larger datasets in prior work [4]. In that study, “the newer version of the tabular GAN-based data generator, called CTGAN, produced similar results in less time” compared to the TGAN implementation. Therefore, the current performance shortfall most plausibly stems from limited sample size rather than issues with the CTGAN algorithm itself.

More sophisticated generative neural architectures such as Boltzmann machines and restricted Boltzmann machines (RBMs) were not systematically evaluated in the present study [73–77]. These models typically require substantially larger datasets to achieve stable training and are known to perform poorly in very small-sample regimes. The scikit-learn documentation [78], the Python framework for machine learning used for some of the present experiments such as GAN, explicitly states that "The parameter learning algorithm used (Stochastic Maximum Likelihood) prevents the representations from straying far from the input data, which makes them capture interesting regularities, but makes the model less useful for small datasets, and usually not useful for density estimation" (https://scikit-learn.org/stable/modules/neural_networks_unsupervised.html). As such, their application to datasets of the size examined here (n = 6 per group) would likely require prior augmentation or strong regularization, limiting their suitability as baseline comparators for evaluating genESOM's performance in small-sample scenarios.

Among further generative approaches not addressed in the present experiments, hidden Markov models (HMMs) possess generative capacity [79] but are commonly used for temporal or sequential data [80]. While we have successfully employed HMMs for sequential data analysis in previous work [81], their application to the static, cross-sectional lipidomics measurements in the present study was considered less appropriate given the absence of temporal dependencies. Therefore, genESOM is specifically suited to the constraints and goals of small-sample biomedical research involving cross-sectional tabular numerical data.

Finally, genESOM possesses a unique advantage through its integrated error inflation control via dimensionality modulation between structure learning and data generation [12]. This embedded alpha error signaling mechanism, achieved through engineered diagnostic features that serve as negative controls, provides a data-driven stopping criterion that halts augmentation when error inflation is detected. To the authors' knowledge, none of the alternative generative models evaluated provides comparable built-in safeguards. Given the critical need to reduce animal sample sizes without compromising statistical validity, where failure would render preclinical experiments worthless and even the reduced number of animal lives wasted, the structural advantage of genESOM lies in its ability to actively signal when augmentation boundaries are approached. This distinguishes it from alternative methods, for which such error control mechanisms remain to be demonstrated.

5. Conclusions

This study addresses the persistent challenge of small sample sizes in preclinical research by demonstrating that generative artificial intelligence can improve analytical stability and sensitivity in exploratory analyses without compromising biological interpretability. Using a validated preclinical lipidomics dataset from an experimental autoimmune encephalomyelitis mouse model, we show that genESOM-based augmentation can help recover treatment-specific signals that become difficult to detect when sample sizes are reduced from n = 8–10 to n = 6 per group.

The proposed approach integrates machine learning across the analytical workflow: emergent self-organizing maps (ESOM) for structure detection, genESOM for topology-preserving data augmentation, and supervised feature selection for identifying treatment-discriminating variables. Critically, our genESOM [4] includes built-in safeguards against statistical inflation through its self-developed dimensional modulation procedure and engineered negative controls [12], providing a data-driven stopping criterion that mitigates overfitting. This error-control mechanism distinguishes genESOM from alternative generative methods lacking comparable constraints.

Application to the intentionally reduced dataset restored the principal patterns of statistically supported lipid mediators, including lysophosphatidic acid and ceramide species, yielding group segregation comparable to the full dataset. This was achieved without inflating false positives, as verified by both conventional statistics and machine learning validation. In contrast, Gaussian mixture models and conditional GANs failed to recover meaningful structure under identical conditions, underscoring the specific suitability of genESOM for small-sample data contexts. In line with the growing trend of applying machine learning to modestly sized datasets [64] and independent advances integrating generative methods into biomedical workflows [82–87], the presented genESOM framework constitutes a promising, empirically validated supporting tool to extract valid results from small-sample preclinical studies. It provides a means to extend analytical reach rather than to replace biological replication.

Our findings indicate that genESOM-based augmentation can, under appropriate conditions, assist in achieving reliable analyses with smaller

cohorts, i.e., potentially reducing the number of required animal subjects by up to about 30–50% in exploratory contexts, while maintaining reproducibility and scientific validity. Importantly, we emphasize that synthetic data cannot substitute for independent biological samples or adequately powered confirmatory experiments [68]. Instead, this framework should be viewed as a complementary analytical tool to guide early-stage discovery and inform efficient experimental design, consistent with the 3Rs principles of animal research. Overall, genESOM offers a scientific method to stabilize analyses of limited datasets. By embedding synthetic observations within the true biological data manifold and monitoring for error inflation, this framework advances both the methodological robustness and ethical standards of preclinical research, i.e., promoting exploratory knowledge extraction while respecting the boundaries of biological inference. While synthetic augmentation cannot substitute for biological replication, it can support exploratory analyses and help reduce the need for additional animal experimentation, provided the necessary caution is observed. That under these conditions valid results can be obtained is shown in this report.

Declarations

None

Ethics approval

Only previously published data were used in this report. The original experiments were performed according to the ethical guidelines and the details about ethics approvals are reported in the respective original publications cited in the present paper.

Funding

JL was supported by the Deutsche Forschungsgemeinschaft (DFG LO 612/16-1) for the project entitled “Generative artificial intelligence-based algorithm to increase the predictivity of preclinical studies while keeping sample sizes small”).

CRediT authorship contribution statement

Lotsch Jorn: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Benjamin Mayer:** Formal analysis, Data curation. **Natasja de Bruin:** Writing – original draft, Validation, Investigation. **Alfred Ultsch:** Writing – original draft, Validation, Methodology, Formal analysis.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

A preprint containing an earlier compilation of this study, together with other materials that have since been published separately, is available at https://ssrn.com/abstract=5099319 or https://doi.org/10.2139/ssrn.5099319. This preprint is not under review or submitted anywhere and will not be further pursued.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.phrs.2026.108159.

Data availability

available at https://data.mendeley.com/datasets/m2p6rr9v36/1 (DOI: 10.17632/m2p6rr9v36.1)

References

[1] R. Martić-Kehl Mi Fau - Schibli, P.A. Schibli R. Fau - Schubiger, P.A. Schubiger, Can animal data predict human outcome? Problems and pitfalls of translational animal research, (1619-7089 (Electronic)).
[2] J.S. Mogil, Animal models of pain: progress and challenges, Nat. Rev. Neurosci. 10 (4) (2009) 283–294.
[3] J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, M. West, Generative or discriminative? getting the best of both worlds, Bayesian Stat. 8 (3) (2007) 3–24.
[4] A. Ultsch, J. Lötsch, Augmenting small biomedical datasets using generative AI methods based on self-organizing neural networks, Brief. Bioinform 26 (1) (2024).
[5] H.P. Foote, C. Hong, M. Anwar, M. Borentain, K. Bugin, N. Dreyer, J. Fessel, N. Goyal, M. Hanger, A.F. Hernandez, C.P. Hornik, J.G. Jackman, A.C. Lindsay, M. E. Matheny, K. Ozer, J. Seidel, N. Stockbridge, P.J. Embi, C.J. Lindsell, Embracing Generative Artificial Intelligence in Clinical Research and Beyond: Opportunities, Challenges, and Solutions, JACC Adv. 4 (3) (2025) 101593.
[6] V. Mahajan, S. Konar, A. Ray, T. Samra, Applications of generative artificial intelligence to augment clinician's capability for medical data analysis in RStudio, Indian J. Anaesth. 68 (9) (2024) 836–838.
[7] N. Fahad, R.I. Rabbi, S. Benta Hasan, F. Sultana Prity, R. Ahmed, F. Ahmed, M. J. Hossen, T.H. Liew, M.S. Sayeed, K. Ong Michael Goh, Generative AI in clinical (2020–2025): a mini-review of applications, emerging trends, and clinical challenges, Front. Digit. Health 7 - 2025 (2025).
[8] A. Gosain, S. Sardana, Handling class imbalance problem using oversampling techniques: A review, Int. Conf. Adv. Comput. Commun. Inform. (ICACCI) 2017 (2017) 79–85.
[9] B.J. Sagarin, J.K. Ambler, E.M. Lee, An Ethical Approach to Peeking at Data, Perspect. Psychol. Sci. 9 (3) (2014) 293–304.
[10] G.M. Currie, K.E. Hawk, E.M. Rohren, Generative Artificial Intelligence Biases, Limitations and Risks in Nuclear Medicine: An Argument for Appropriate Use Framework and Recommendations, Semin. Nucl. Med. 55 (3) (2025) 423–436.
[11] J. Lötsch, I. Tegder, N. de Bruin, D. Thomas, G. Geisslinger, Targeted lipidomics dataset of central nervous system and plasma from mice with experimental autoimmune encephalomyelitis, Data Brief. 62 (2025) 111948.
[12] J. Lötsch, A. Himmelspach, D. Kringel, Dimensionality-modulated generative AI for safe biomedical dataset augmentation, iScience 29 (1) (2026) 114321.
[13] N.M. de Bruin, K. Schmitz, S. Schiffmann, N. Tafferner, M. Schmidt, H. Jordan, A. Haussler, I. Tegeder, G. Geisslinger, M.J. Parnham, Multiple rodent models and behavioral measures reveal unexpected responses to FTY720 and DMF in experimental autoimmune encephalomyelitis, Behav. Brain Res 300 (2016) 160–174.
[14] K. Schmitz, R. Brunkhorst, N. de Bruin, C.A. Mayer, A. Häussler, N. Ferreiros, S. Schiffmann, M.J. Parnham, S. Tunaru, J. Chun, S. Offermanns, C. Foerch, K. Scholich, J. Vogt, S. Wicker, J. Lötsch, G. Geisslinger, I. Tegeder, Dysregulation of lysophosphatidic acids in multiple sclerosis and autoimmune encephalomyelitis, Acta Neuropathol. Commun. 5 (1) (2017) 42.
[15] A. Sens, S. Rischke, L. Hahnefeld, E. Dorochow, S.M.G. Schäfer, D. Thomas, M. Köhm, G. Geisslinger, F. Behrens, R. Gurke, Pre-analytical sample handling standardization for reliable measurement of metabolites and lipids in LC-MS-based clinical research, J. Mass Spectrom. Adv. Clin. Lab 28 (2023) 35–46.
[16] R. Ihaka, R. Gentleman, R: A Language for Data Analysis and Graphics, J. Comput. Graph. Stat. 5 (3) (1996) 299–314.
[17] R Core Team, R: A Language and Environment for Statistical Computing, Vienna, Austria, 2021.
[18] G. Van Rossum, F.L. Drake Jr, Python tutorial, Centrum voor Wiskunde en Informatica Amsterdam1995.
[19] H.J. Motulsky, T. Head, P.B.S. Clarke, Analyzing lognormal data: A nonmathematical practical guide, Pharmacol. Rev. 77 (3) (2025) 100049.
[20] D.R. Mould, R.N. Upton, Basic concepts in population modeling, simulation, and model-based drug development-part 2: introduction to pharmacokinetic modeling methods, CPT Pharmacomet. Syst. Pharm. 2 (4) (2013) e38.
[21] L.F. Lacey, O.N. Keene, J.F. Pritchard, A. Bye, Common noncompartmental pharmacokinetic variables: are they normally or log-normally distributed? J. Biopharm. Stat. 7 (1) (1997) 171–178.

[22] S. Malkusch, L. Hahnefeld, R. Gurke, J. Lötsch, Visually guided preprocessing of bioanalytical laboratory data using an interactive R notebook (pguIMP), CPT Pharmacomet. Syst. Pharm. 10 (11) (2021) 1371–1381.
[23] M. Kuhn, Building Predictive Models in R Using the caret Package, J. Stat. Softw. 28 (5) (2018) 1–26, https://doi.org/10.18637/jss.v028.i05.
[24] T.K. Ho, Random decision forests, in: Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1), 1, IEEE Computer Society, 1995, p. 278.
[25] L. Breiman, Random Forests, Mach. Learn 45 (1) (2001) 5–32.
[26] D.J. Stekhoven, P. Bühlmann, MissForest–non-parametric missing value imputation for mixed-type data, Bioinformatics 28 (1) (2012) 112–118.
[27] Z. Šidák, Rectangular Confidence Regions for the Means of Multivariate Normal Distributions, J. Am. Stat. Assoc. 62 (318) (1967) 626–633.
[28] R.F. Wolff, K.G.M. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J. B. Reitsma, J. Kleijnen, S. Mallett, PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies, Ann. Intern. Med. 170 (1) (2019) 51–58.
[29] A. Lo, H. Chernoff, T. Zheng, S.-H. Lo, Why significant variables aren't automatically good predictors, in: Proceedings of the National Academy of Sciences of the United States of America, 112, 2015, pp. 13892–13897.
[30] C. Cortes, V. Vapnik, Support-Vector Networks, Mach. Learn. 20 (3) (1995) 273–297.
[31] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theor. 13 (1) (1967) 21–27.
[32] W.W. Cohen, Fast Effective Rule Induction, ICML (1995).
[33] M.B. Kursa, W.R. Rudnicki, Feature Selection with the Boruta Package, J. Stat. Softw. 36 (11) (2010) 13.
[34] C.J. Morgan, Use of proper statistical techniques for research studies with small samples, Am. J. Physiol. Lung Cell Mol. Physiol. 313 (5) (2017) L873–l877.
[35] J. Charan, N.D. Kantharia, How to calculate sample size in animal studies? J. Pharm. Pharm. 4 (4) (2013) 303–306.
[36] T.F.C. Kung, C.M. Wilkinson, L.J. Liddle, F. Colbourne, A systematic review and meta-analysis on the efficacy of glibenclamide in animal models of intracerebral hemorrhage, PLoS One 18 (9) (2023) e0292033.
[37] L.J. Liddle, C.A. Dirks, B.A. Fedor, M. Almekhlafi, F. Colbourne, A Systematic Review and Meta-Analysis of Animal Studies Testing Intra-Arterial Chilled Infusates After Ischemic Stroke, Front Neurol. 11 (2020) 588479.
[38] M.F. Festing, D.G. Altman, Guidelines for the design and statistical analysis of experiments using laboratory animals, Ilar J. 43 (4) (2002) 244–258.
[39] K. Neumann, U. Grittner, S.K. Piper, A. Rex, O. Florez-Vargas, G. Karystianis, A. Schneider, I. Wellwood, B. Siegerink, J.P. Ioannidis, J. Kimmelman, U. Dirnagl, Increasing efficiency of preclinical research by group sequential designs, PLoS Biol. 15 (3) (2017) e2001307.
[40] F. Rimet, J.-C. Druart, O. Anneville, Exploring the dynamics of plankton diatom communities in Lake Geneva using emergent self-organizing maps (1974–2007), Ecol. Inform. 4 (2) (2009) 99–110.
[41] A. Ultsch, D. Kämpf, Knowledge Discovery in DNA Microarray Data of Cancer Patients with Emergent Self Organizing Maps. Proceedings of the European Symposium on Artificial Neural Networks (ESANN 2004), 2004, pp. 501–506.
[42] T. Kohonen, Self-organized formation of topologically correct feature maps, Biol. Cyber 43 (1982) 59–69.
[43] A. Ultsch, Maps for Visualization of High-Dimensional Data Spaces, WSOM (2003) 225–230.
[44] J. Lötsch, A. Ultsch, Exploiting the structures of the U-matrix, in: T. Villmann, F.-M. Schleif, M. Kaden, M. Lange (Eds.), Advances in Intelligent Systems and Computing, Springer, Heidelberg, 2014, pp. 248–257.
[45] A. Ultsch, J. Lötsch, Machine-learned cluster identification in high-dimensional data, J. Biomed. Inf. 66 (2017) 95–104.
[46] T. Kohonen. Self-Organizing Maps, Springer Series in Information Sciences, Springer, Berlin, Heidelberge, 1995, https://doi.org/10.1007/978-3-642-97610-0.
[47] A. Ultsch, H.P. Siemon, Kohonen's self organizing feature maps for exploratory data analysis. INNC'90, Int. Neural Network Conference, Kluwer, Dordrecht, Netherlands, 1990, pp. 305–308.
[48] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, J. R. Stat. Soc. Ser. B 39 (1) (1977) 1–38.
[49] A. Ultsch, J. Lötsch, Generative learning with emergent self-organizing neuronal networks. Conference of the International Federation of Classification Societies, Tokyo, 2017, p. 266.
[50] J. Lötsch, A. Himmelspach, D. Kringel, Dimensionality modulated generative AI for safe biomedical dataset augmentation, iScience.
[51] J. Lötsch, S. Malkusch, A. Ultsch, Comparative assessment of automated algorithms for the separation of one-dimensional Gaussian mixtures, Inform. Med. Unlocked 34 (2022) 101113.
[52] M. Bayes, M. Price, An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S, Philos. Trans. 53 (1763) 370–418.
[53] H. Guo, H.L. Viktor, Boosting with data generation: improving the classification of hard to learn examples. International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Springer, 2004, pp. 1082–1091.
[54] P.M. Baggenstoss, Statistical modeling using gaussian mixtures and hmms with matlab, Naval Undersea Warfare Center, Newport RI, 2002.
[55] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).
[56] L. Xu, M. Skoularidou, A. Cuesta-Infante, K. Veeramachaneni, in: H. Wallach, H. Larochelle, A. Beygelzimer, Fdt Alch\'-Buc, E. Fox, R. Garnett (Eds.), Modeling Tabular data using Conditional GAN, 2019.
[57] C.E. Bonferroni, Teoria statistica delle classi e calcolo delle probabilita, in: Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8, 1936, pp. 3–62.
[58] K.H. Brodersen, C.S. Ong, K.E. Stephan, J.M. Buhmann, The Balanced Accuracy and Its Posterior Distribution, Pattern Recognition (ICPR), 2010 20th Int. Conf. (2010) 3121–3124.
[59] K. Pearson, Note on Regression and Inheritance in the Case of Two Parents, Proc. R. Soc. Lond. Ser. I 58 (1895) 240–242.
[60] D.C. Funder, D.J. Ozer, Evaluating Effect Size in Psychological Research: Sense and Nonsense, Adv. Methods Pract. Psychol. Sci. 2 (2) (2019) 156–168.
[61] R.N. Shepard, The analysis of proximities: Multidimensional scaling with an unknown distance function. II, Psychometrika 27 (3) (1962) 219–246.
[62] R.N. Shepard, The analysis of proximities: multidimensional scaling with an unknown distance function. I, Psychometrika 27 (2) (1962) 125–140.
[63] J. Lötsch, A. Ultsch, Pitfalls of Using Multinomial Regression Analysis to Identify Class-Structure-Relevant Variables in Biomedical Data Sets: Why a Mixture of Experts (MOE) Approach Is Better, BioMedInformatics 3 (4) (2023) 869–884.
[64] I. Kraljevski, Y.C. Ju, D. Ivanov, C. Tschöpe, M. Wolff, How Do Mach. Learn. Small Data? A Rev. Ind. Perspect. (2023).
[65] J. Lötsch, F. Lerch, R. Djaldetti, I. Tegeder, A. Ultsch, Identification of disease-distinct complex biomarker patterns by means of unsupervised machine-learning using an interactive R toolbox (Umatrix), BMC Big Data Anal. 3 (5) (2018), https://doi.org/10.1186/s41044-018-0032-1.
[66] Z. Pang, L. Xu, C. Viau, Y. Lu, R. Salavati, N. Basu, J. Xia, MetaboAnalystR 4.0: a unified LC-MS workflow for global metabolomics, Nat. Commun. 15 (1) (2024) 3675.
[67] D.C. Angus, R. Khera, T. Lieu, V. Liu, F.S. Ahmad, B. Anderson, S.V. Bhavani, A. Bindman, T. Brennan, L.A. Celi, F. Chen, I.G. Cohen, A. Denniston, S. Desai, P. Embí, A. Faisal, K. Ferryman, J. Gerhart, M. Gross, T. Hernandez-Boussard, M. Howell, K. Johnson, K. Lee, X. Liu, K. Lomis, A.J. London, C.A. Longhurst, K. D. Mandl, E. McGlynn, M.M. Mello, F. Munoz, L. Ohno-Machado, D. Ouyang, R. Perlis, A. Phillips, D. Rhew, J.S. Ross, S. Saria, L. Schwamm, C.W. Seymour, N. H. Shah, R. Shah, K. Singh, M. Solomon, K. Spates, K. Spector-Bagdady, T. Wang, J. W. Gichoya, J. Weinstein, J. Wiens, K. Bibbins-Domingo, AI, Health, and Health Care Today and Tomorrow: The JAMA Summit Report on Artificial Intelligence, Jama 334 (18) (2025) 1650–1664.
[68] P. Reinagel, Is N-Hacking Ever OK? The consequences of collecting more data in pursuit of statistical significance, PLoS Biol. 21 (11) (2023) e3002345.
[69] A. Zimek, E. Schubert, H.P. Kriegel, A survey on unsupervised outlier detection in high-dimensional numerical data, Stat. Anal. Data Min. 5 (5) (2012) 363–387.
[70] T. Yamazaki, J. Igarashi, H. Yamaura, Human-scale Brain Simulation via Supercomputer: A Case Study on the Cerebellum, Neuroscience 462 (2021) 235–246.
[71] J. Lötsch, A. Ultsch, Generative artificial intelligence based algorithm to increase the predictivity of preclinical studies while keeping sample sizes small, in: H. A. Kestler, M. Schmid, L. Lausser, A. Fürstberger (Eds.), Statistical Computing 2019, Ulmer Informatik-Bericht, Günzburg, Germany, 2019, pp. 29–30.
[72] A. Creswell, A.A. Bharath, Adversarial training for sketch retrieval. Computer Vision – ECCV 2016 Workshops, Springer International Publishing, Amsterdam, The Netherlands, 2016.
[73] G.E. Hinton, T.J. Sejnowski, Optimal Perceptual Inference, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
[74] N. Caporale, Y. Dan, Spike timing-dependent plasticity: a Hebbian learning rule, Annu Rev. Neurosci. 31 (2008) 25–46.
[75] G.E. Hinton, Training products of experts by minimizing contrastive divergence, Neural Comput. 14 (8) (2002) 1771–1800.
[76] H. Larochelle, M. Mandel, R. Pascanu, Y. Bengio, Learning algorithms for the classification restricted Boltzmann machine, J. Mach. Learn. Res. 13 (2012) 643–669.
[77] S. Ruslan, H. Geoffrey, Deep Boltzmann Machines, PMLR, 2009, pp. 448–455.
[78] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res. 12 (85) (2011) 2825–2830.
[79] S. Laxman, P.S. Sastry, K.P. Unnikrishnan, Discovering frequent episodes and learning hidden Markov models: a formal connection, IEEE Trans. Knowl. Data Eng. 17 (11) (2005) 1505–1517.
[80] H. Narimatsu, H. Kasai, State duration and interval modeling in hidden semi-Markov model for sequential data analysis, Ann. Math. Artif. Intell. 81 (3) (2017) 377–403.
[81] S. Malkusch, J.V. Rahm, M.S. Dietz, M. Heilemann, J.B. Sibarita, J. Lötsch, Receptor tyrosine kinase MET ligand-interaction classified via machine learning from single-particle tracking data, Mol. Biol. Cell 33 (6) (2022) ar60.
[82] R. Gulakala, B. Markert, M. Stoffel, Generative adversarial network based data augmentation for CNN based detection of Covid-19, Sci. Rep. 12 (1) (2022) 19186.
[83] D.-C. Li, S.-C. Chen, Y.-S. Lin, K.-C. Huang, A Generative Adversarial Network Structure for Learning with Small Numerical Data Sets, Applied Sciences, 2021.
[84] J. Röglin, K. Ziegeler, J. Kube, F. König, K.-G. Hermann, S. Ortmann, Improving classification results on a small medical dataset using a GAN; An outlook for dealing with rare disease datasets, Front. Comput. Sci. 4 (2022).
[85] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, T. Aila, Training Generative Adversarial Networks with Limited Data, in: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, H. Lin (Eds.) Advances in Neural Information Processing Systems, Curran Associates, Inc., pp. 12104–12114-12104–12114.

[86] R.H. Randhawa, N. Aslam, M. Alauthman, H.U.S.N.A.I.N. Rafiq, Evasion Generative Adversarial Network for Low Data Regimes, IEEE Trans. Artif. Intell. (2022), https://doi.org/10.48550/arXiv.2109.08026 arXiv:2109.08026.
[87] S. Gurumurthy, R.K. Sarvadevabhatla, V.B. Radhakrishnan, DeLiGAN: Generative Adversarial Networks for Diverse and Limited Data, 2017.
[88] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag, New York, 2009.
[89] Z. Gu, R. Eils, M. Schlesner, Complex heatmaps reveal patterns and correlations in multidimensional genomic data, Bioinformatics 32 (18) (2016) 2847–2849.