# A NEW SEMI-SUPERVISED CLASSIFICATION METHOD USING A SUPERVISED AUTOENCODER FOR BIOMEDICAL APPLICATIONS

*Cyprien Gille*[1], *Frederic Guyard*[2] *and Michel Barlaud*[1], *Fellow IEEE*

[1] Laboratoire I3S CNRS, Cote d'Azur University, Sophia Antipolis, France
[2]Orange Labs, Sophia Antipolis, France

## ABSTRACT

Annotation of biomedical databases by clinicians is a very difficult, sometimes imprecise, and time consuming task. An alternative is to ask the clinician expert for the annotations they are the most confident in, which results in a semi-supervised classification problem. In this paper, we present a new approach to solve semi-supervised classification tasks for biomedical applications, involving a supervised autoencoder network. We train the Semi-Supervised AutoEncoder (SSAE) on labelled data using a double descent algorithm. Then, we classify unlabelled samples using the learned network thanks to a softmax classifier applied to the latent space which provides a classification confidence score for each class. Experiments show that the SSAE outperforms Label Propagation and Spreading and the Fully Connected Neural Network both on a synthetic dataset and on four real-world biological datasets.

***Index Terms*—Semi-supervised learning, Autoencoder neural networks.**

## I. RELATED WORKS

Annotation of biomedical databases from clinical data by clinicians is a very difficult and time consuming task. The annotated labels are sometimes imprecise and noisy due an heterogeneous treatment depending on detail filled in by several clinicians. An alternative is to only ask the clinician expert for the annotations they are the most confident in : we then obtain a semi-supervised classification problem. Semi-supervised learning is a machine learning paradigm [1] using both labelled and unlabelled data to perform two tasks, classically classification and clustering. Semi-supervised learning algorithms attempt to improve performance in one of these two tasks by utilizing information generally associated with the other [2], [3], [4], [5], [6], [7]. Subramanya and Talukdar (2014) provided an overview of several graph-based techniques [8]; Triguero et al. analyzed pseudo-labelling methods [9]; Oliver et al. compared several semi-supervised neural networks, on two image classification problems [10]. Recent research on semi-supervised learning is focused on neural network-based methods (see [11] for a survey). Autoencoders were used in semi-supervised learning to improve classification performance [12], [1], [13]. However, classical Variational Autoencoder networks encourage their latent space to fit a prior distribution [12], typically a Gaussian but they are non-adaptive and unfortunately may not match the specific data distribution.

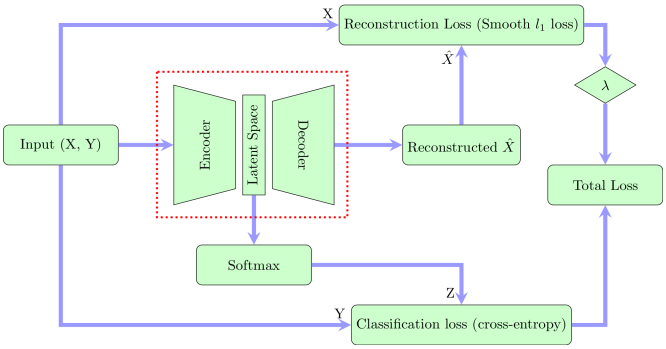


**Fig. 1**. Supervised autoencoder framework

In this paper we propose a new approach using a supervised autoencoder (SSAE) where our network encourages the latent space to fit a distribution learned from the labels rather than a parametric prior distribution. On top of that, we propose a constrained regularization approach that takes advantage of available efficient projection algorithms for the $\ell_1$ constraint [14], [15], and the structured constraint $\ell_{1,1}$ [16].

## II. SEMI-SUPERVISED AUTOENCODER FRAMEWORK

Figure 1 depicts the main constituent blocks of our proposed approach. Note that we added a "soft max" block to our autoencoder to compute the classification loss.
Let $X$ be the dataset in $\mathbb{R}^d$, and $Y$ the labels in $\{0, \ldots, k\}$, with $k$ the number of classes. Let $\mathrm{Z} \in \mathbb{R}^k$ be the encoded latent vectors, $\widehat{X} \in \mathbb{R}^d$ the reconstructed data and W the weights of the neural network. Note that the dimension of the latent space $k$ corresponds to the number of classes..

The goal is to compute the network weights $W$ minimizing the total loss which includes both the classification loss and the reconstruction loss. To perform feature selection, as biomedical datasets often present a relatively small number

of informative features, we also want to sparsify the network, following the work proposed in [16], [17] and [18]. To do so, instead of the classical computationally expensive lagrangian regularization approach [19], we propose to minimize the following constrained approach :

$$Loss(W) = \mathcal{H}(Z, Y) + \lambda\psi(\widehat{X} - X) \text{ s.t. } \|W\|_1^1 \leq \eta. \quad (1)$$

We use the Cross Entropy Loss for the classification loss $\mathcal{H}$. We use the robust Smooth $\ell_1$ (Huber) Loss [20] as the reconstruction loss $\psi$. Let us recall that $\ell_{1,1}$ norm is computed as the maximum $\ell_1$ norm of a column. We propose the following algorithm: we first compute the radius $t_i$ and then project the rows using the $\ell_1$ adaptive constraint $t_i$ (see [16] for more details).

Following the work by Frankle and Carbin [17] further developed by [18], we follow a double descent algorithm, originally proposed as follows: after training a network, set all weights smaller than a given threshold to zero, rewind the rest of the weights to their initial configuration, and then retrain the network from this starting configuration while keeping the zero weights frozen (untrained). We train the network using the classical Adam optimizer [21].
To achieve structured sparsity, we replace the thresholding by our $\ell_{1,1}$ projection and devise algorithm 1.

**Algorithm 1** Double descent algorithm. $\phi$ is the total loss as defined in (1), $\nabla\phi(W, M_0)$ is the gradient masked by the binary mask $M_0$, $A$ is the Adam optimizer, $N$ is the total number of epochs and $\gamma$ is the learning rate.

*# First descent*
**Input:** $W_{init}, \gamma, \eta$
**for** $n = 1, \ldots, N$ **do**
  $W \leftarrow A(W, \gamma, \nabla\phi(W))$
**end for**
*# Projection*
**for** $i = 1, \ldots, d$ **do**
  $t_i := proj_{\ell_1}((\|v_i\|_1)_{i=1}^{l}, \eta)$
  $w_i := proj_{\ell_1}(v_i, t_i)$
**end for**
$(M_0)_{ij} := \mathbb{1}_{x \neq 0}(w_{ij})$
**Output:** $M_0$
*# Second descent*
**Input:** $W_{init}, M_0, \gamma$
**for** $n = 1, \ldots, N$ **do**
  $W \leftarrow A(W, \gamma, \nabla\phi(W, M_0))$
**end for**
**Output:** $W$

$proj_{\ell_1}(V, \eta)$ is the projection onto the $\ell_1$-ball of radius $\eta$, which can be computed using fast algorithms [14], [15]. Low values of $\eta$ imply high sparsity of the network. Using the $\ell_{1,1}$ constraint specifically gives us structured sparsity [22].

## III. EXPERIMENTAL RESULTS

We implemented our SSAE method using the PyTorch framework for the model, optimizer, schedulers and loss functions. We used a symmetric linear fully connected network, with the encoder comprised of an input layer of $d$ neurons, one hidden layer followed by a ReLU activation function and a latent layer of dimension $k$.
We compare the SSAE with two classical semi-supervised classification techniques based on similarity graphs, Label Spreading (LabSpread) and Label Propagation (LabProp) [23], using their respective implementations in scikit-learn. We also provide comparison with a Fully Connected Neural Network (FCNN) implemented in PyTorch, corresponding to the encoder section of the SSAE. We evaluated our method on synthetic data and two different biological datasets. We apply classical log-transform, zero-mean and scaling to the biological datasets. The code to reproduce our results is made freely available on GitHub[1]. Note that our SSAE provides a two-dimensional latent space where the samples can be visualized, and their respective classifications interpreted. Moreover our supervised autoencoder specifically provides informative features [24] which are especially insightful for biologists.

### III-A. Synthetic data

To generate artificial data to benchmark our SSAE, we use the $make\_classification$ utility from scikit-learn. This generator creates clusters of points that are normally distributed along vertices of a $k$-dimensional hypercube. We are able to control the length of those vertices and thus the separability of the generated dataset. We generate $n = 1000$ samples (a number related to the number of samples in large biological datasets) with a number $d$ of features. We chose $d = 1000$ as the dimension to test because this is the typical range for biological data. We chose a low number of informative features ($< 64$ ) realistically with single cell or metabolomic biological databases [25], [26]. Within this dataset, we randomly pick 40% of the samples to be considered as unlabelled. We fit or train the algorithms on the remaining samples, and then compute the classification accuracy by comparing the labels predicted by the learned network to the original labels. We report the results in figure 2.

| Synthetic | SSAE | LProp LSpread | FCNN |
|---|---|---|---|
| Accuracy % | 85.55 | 63.6 | 71.0 |
| AUC | 0.921 | 0.701 | 0.803 |
| F1 score | 0.87 | 0.647 | 0.725 |

**Table I**. Synthetic dataset : comparison of LabelPropagation, LabelSpreading, FCNN and SSAE. 40% of unlabeled data, Mean over 3 seeds, separability=0.8, n=1000, d=1000, 8 informative features.

[1]https://github.com/CyprienGille/Semi-Supervised-AutoEncoder

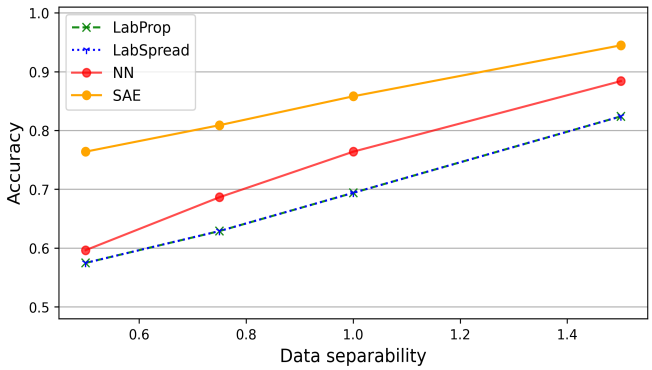


**Fig. 2**. Accuracy as a function of separability: Comparison on synthetic data of our SSAE, the FCNN, Label Propagation and Label Spreading: Mean over 3 seeds, 40% unlabeled samples, 8 informative features $n = 1000, d = 1000$.

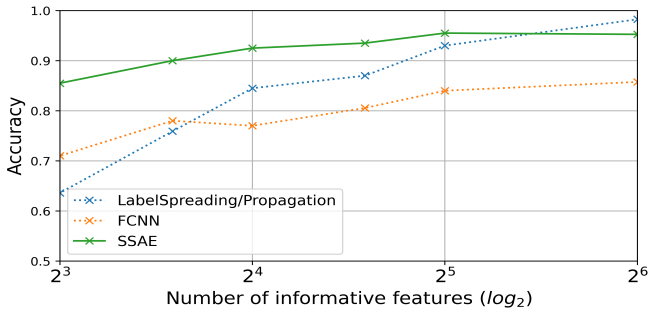


**Fig. 3**. Accuracy as a function of the number of informative features: Comparison on synthetic data of our SSAE, the FCNN, Label Propagation and Label Spreading : 40% unlabeled samples, Mean over 3 seeds, separability= 0.8, $n = 1000, d = 1000$.

Figure 2 shows that our SSAE largely outperforms the classical methods for any low number of informative features. Classical methods such as label propagation and label spreading are based on a similarity matrix and thus suffer from the curse of dimensionality (the similarity function used in this paper is a $kNN$ algorithm, which has to compute the distance between two samples). As the dimension increases, vectors become indiscernible [27] and the predictive power of the aforementioned methods is substantially reduced. Notably, figure 2 demonstrates that in high dimension, the performance of the FCNN also falls off in comparison to that of the SSAE.

Table I confirm the previous results: the SSAE outperforms the two classical methods across all metrics. More notably, our SSAE also outperforms the F1 score of the FCNN by 12% for $n = 1000, d = 1000$.

Figure 5 shows the distribution of the scores of the predicted labels: the two classical methods show poor decision capabilities, as the score of the predicted class will often be close to 0.5. On the other hand, the neural networks are more confident about their predictions, with the SSAE being a better discriminator than the FCNN, which is reflected in their respective metrics in table I.

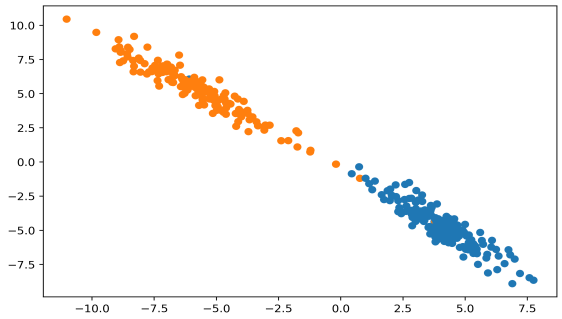

**Fig. 4**. Synthethic dataset, $n = 1000, d = 1000$, 40% of unlabeled samples, Separability=0.8, 8 informative features. Unlabeled samples represented in the latent space of the SSAE.

Figure 4 shows the latent space of the SSAE. We can see that, after learning, the SSAE is able to accurately separate the labeled samples, and that the unlabeled samples have been relatively clustered according to their labels. This feature, offered only by the SSAE, provides interpretability to the results and an insightful tool for practical use.

### III-B. Biological datasets

We now present the results of our SSAE on four biological datasets : two single-cell databases and two metabolomics databases.

The **IPF** dataset is a single cell RNA seq published database which is made of human fibroblasts transcriptomic profiles, obtained from lung explants of patients with Idiopathic Pulmonary Fibrosis and from healthy donors. This dataset comes from a study [28] aimed at characterizing the transcriptional changes induced by the pathology in pulmonary cell types. The 1443 samples are described by 14369 numerical features with high sparsity. From this labelled dataset, we randomly pick a subset of samples to be considered unlabelled, following the same procedure as described in section III-A.

| IPF | SSAE | LProp | LSpread | FCNN |
|---|---|---|---|---|
| Accuracy % | 96.66 | 72.14 | 72.14 | 95.55 |
| AUC | 0.9947 | 0.7792 | 0.7730 | 0.9903 |
| F1 score | 0.9633 | 0.6977 | 0.6976 | 0.9510 |

**Table II**. **IPF** dataset: Mean Metrics over 3 seeds : comparison of LabelPropagation, LabelSpreading, FCNN and SSAE. 40% of unlabeled data.

Table II shows that for the **IPF** dataset (which is a high-dimensional dataset) only the neural networks manage

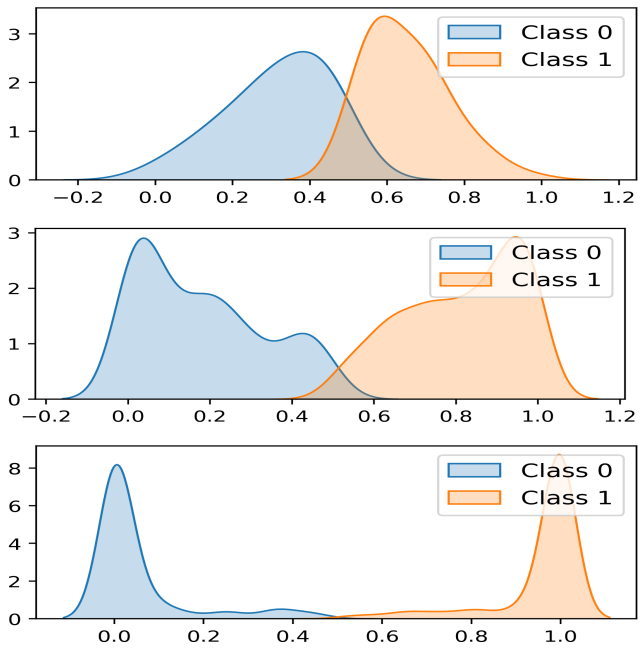


**Fig. 5**. Synthetic dataset, $n = 1000, d = 1000$, separability= 0.8, unlabeled proportion of 40%, 8 informative features. Comparison of the prediction score distributions. From top to bottom : LabelPropagation/LabelSpreading, FCNN, SSAE with $\ell_{1,1}$ constraint.

to accurately classify the unlabelled samples; they both do so almost optimally, reaching very high metrics.

The **IFNGR2** dataset is a single cell RNA seq published database from [25]. Its goal is to decipher the CRISPRI-induced signature as well as the heterogeneity of the perturbation response.

| IPF | SSAE | LProp | LSpread | FCNN |
|---|---|---|---|---|
| Accuracy % | 90.06 | 68.8 | 68.8 | 88.33 |
| AUC | 0.874 | 0.605 | 0.605 | 0.89 |
| F1 score | 0.879 | 0.45 | 0.45 | 0.863 |

**Table III**. **IFNGR2** dataset: Mean Metrics over 3 seeds : comparison of LabelPropagation, LabelSpreading, FCNN and SSAE. 40% of unlabeled data.

Table III shows that on this second single-cell dataset, the SSAE outperforms the other three methods : Label Propagation and Label Spreading by over 22% of accuracy; and the FCNN by 2% of accuracy and 3% of F1 Score.

The **LUNG** dataset was provided by Mathe et al. [26]. This dataset includes metabolomic data concerning urine samples from 469 Non-Small Cell Lung Cancer (NSCLC) patients prior to treatment and 536 control patients. Each sample is described by 2944 features.

| Lung | SSAE | LProp | LSpread | FCNN |
|---|---|---|---|---|
| Accuracy % | 82.59 | 59.27 | 58.66 | 78.15 |
| AUC | 0.9009 | 0.6569 | 0.6593 | 0.8713 |
| F1 score | 0.8258 | 0.5489 | 0.5399 | 0.7806 |

**Table IV**. **LUNG** dataset: Mean Metrics over 3 seeds : comparison of LabelPropagation, LabelSpreading, FCNN and SSAE. 40% of unlabeled data.

Table IV shows that our SSAE also outperforms the classical methods on this smaller, less balanced dataset. Note that SSAE also outperforms the FCNN by 3% of AUC and 4% of both accuracy and F1 Score.

The **BREAST** dataset was provided by Dr. Jan Budczies and can be found in the supplementary material of Budczies et al. [29]. It includes metabolomics data concerning 271 breast tumor samples: 204 tumors with over-expression of estrogen receptors (ER) and 67 tumors without over-expression of ER. Each sample is described by 161 features.

| Lung | SSAE | LProp | LSpread | FCNN |
|---|---|---|---|---|
| Accuracy % | 87.9 | 73.14 | 73.14 | 85.18 |
| AUC | 0.87 | 0.73 | 0.73 | 0.87 |
| F1 score | 0.837 | 0.52 | 0.52 | 0.780 |

**Table V**. **BREAST** dataset: Mean Metrics over 3 seeds : comparison of LabelPropagation, LabelSpreading, FCNN and SSAE. 40% of unlabeled data.

Table V confirms that our SSAE outperforms both the classical methods and the simple FCNN, by a wide margin for the former and 5% of F1 Score for the latter.

## IV. CONCLUSION AND PERSPECTIVES

In this paper we have presented a new framework to solve semi-supervised classification tasks, involving a supervised autoencoder network. Results on synthetic data and on four real-world biomedical datasets show that this approach outperforms classical semi-supervised classification techniques, and provides insightful features for biological applications, such as a confidence score and a latent space, compared to classical methods and fully connected neural networks.
In a future work, we propose to apply our method to other architectures, such as large Convolutional Neural Networks for image processing.

### ACKNOWLEDGMENTS

The authors thank Marin Truchi and Bernard Mari (IPMC Laboratory) for providing the IPF and IFNGR2 dataset and Thierry Pourcher (TIRO Laboratory) for providing the Lung and BREAST dataset.

## V. REFERENCES


[1] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," pp. 3581–3589, 2014.

[2] O. Chapelle, V. Sindhwani, and S. S. Keerthi, "Optimization techniques for semi-supervised support vector machines." *Journal of Machine Learning Research*, vol. 9, no. 2, 2008.

[3] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in Neural Information Processing Systems 16*. MIT Press, 2004, pp. 321–328.

[4] X. Zhu, "Semi-supervised learning literature survey," *University of Wisconsin-Madison Department of Computer Sciences*, 2005.

[5] Q. Zhu and R. Zhang, "A classification supervised auto-encoder based on predefined evenly-distributed class centroids," *arXiv, cs.CV, 1902.00220v3*, january 2020.

[6] L. Le, A. Patterson, and M. White, "Supervised autoencoders: Improving generalization performance with unsupervised regularizers," in *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018, pp. 107–117.

[7] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *International Conference on International Conference on Machine Learning*, ser. ICML'03, 2003, p. 912–919.

[8] A. Subramanya and P. Talukdar, "Graph-based semi-supervised learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning, 29*, November 2014.

[9] I. Triguero, S. García, and F. Herrera, "Self-labeled techniques for semi-supervised learning: Taxonomy, software and empirical study," *Knowledge and Information Systems*, vol. 42, 02 2015.

[10] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. J. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," 2018.

[11] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning." *Machine Learning*, vol. 109, no. 2, 2020.

[12] D. Kingma and M. Welling, "Auto-encoding variational bayes," *International Conference on Learning Representation*, 2014.

[13] J. Snoek, R. Adams, and H. Larochelle, "On non parametric guidance for learning autoencoder representations," ser. Proceedings of Machine Learning Research, vol. 22. PMLR, 2012, pp. 1073–1080.

[14] L. Condat, "Fast projection onto the simplex and the l1 ball," *Mathematical Programming Series A*, vol. 158, no. 1, pp. 575–585, 2016.

[15] G. Perez, M. Barlaud, L. Fillatre, and J.-C. Régin, "A filtered bucket-clustering method for projection onto the simplex and the $\ell_1$-ball," *Mathematical Programming*, May 2019.

[16] M. Barlaud and F. Guyard, "Learning sparse deep neural networks using efficient structured projections on convex constraints for green ai," *International Conference on Pattern Recognition, Milan*, pp. 1566–1573, 2020.

[17] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *International Conference on Learning Representations*, 2019.

[18] H. Zhou, J. Lan, R. Liu, and J. Yosinski, "Deconstructing lottery tickets: Zeros, signs, and the supermask," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 3597–3607.

[19] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu, "The entire regularization path for the support vector machine," *Journal of Machine Learning Research*, vol. 5, pp. 1391–1415, 2004.

[20] P. J. Huber, "Robust statistics. 1981."

[21] D. Kingma and J. Ba, "a method for stochastic optimization." *International Conference on Learning Representations*, pp. 1–13, 2015.

[22] M. Barlaud, A. Chambolle, and J.-B. Caillau, "Classification and feature selection using a primal-dual method and projection on structured constraints," *International Conference on Pattern Recognition, Milan*, pp. 6538–6545, 2020.

[23] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation." *Technical Report CMU-CALD-02-107, Carnegie Mellon University*, 2002.

[24] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Neural Information Processing Systems, Barcelone, Spain*, vol. 30, 2017.

[25] E. Papalexi *et al*, "Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens," *Nature Genetics*, vol. 53, no. 3, pp. 322–331, 2021.

[26] E. Mathé *et al*, "Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer," *Cancer research*, vol. 74, no. 12, p. 3259—3270, June 2014.

[27] C. Aggarwal, "On k-anonymity and the curse of dimensionality," *Proceedings of the 31st VLDB Conference, Trondheim, Norway*, 2005.

[28] T. Adams *et al*, "Single-cell rna-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis." *Science advances vol. 6,28 eaba1983*, July 2020.

[29] J. Budczies *et al*, "Comparative metabolomics of estrogen receptor positive and estrogen receptor negative breast cancer: Alterations in glutamine and beta-alanine metabolism," *Journal of Proteomics*, vol. 94, pp. 279–288, 2013.