StackFeat: a convergent algorithm for optimal predictor selection in genomic data

Akbar Yermekov[1,*] and *D. A.* Herrera Martí[2]

[1]PAfoS.AI (Predictive Analytics for Science), Almaty, Kazakhstan
[2]CEA List, Université Grenoble Alpes (formerly at Atos/Eviden Quantum), France

Abstract. In high-dimensional genomic data, the curse of dimensionality ($d \gg n$) and limited sampling make feature selection inherently unstable—a critical barrier to biomarker discovery. We introduce StackFeat, an iterative algorithm that accumulates two statistics across repeated cross-validation: signed coefficients (measuring effect strength and direction) and selection frequencies (estimating selection probability). Only features ranking highly by both criteria are retained. On a COVID-19 miRNA dataset (GSE240888), StackFeat identified a stable 5-miRNA signature from 332 features (98.5% reduction), achieving AUC 0.922, significantly outperforming the benchmark 9-gene set (AUC 0.907, p = 0.0016). The signature includes hsa-miR-150-5p, a marker implicated in both COVID-19 survival and Dengue infection. This dual-criterion approach provides convergence guarantees absent in single-criterion methods, enabling discovery of known biomarkers, novel candidates, and previously unknown relationships.

Keywords: marker selection, feature selection, bioinformatics, dimensionality reduction, robust algorithm, stacking, miRNA, COVID-19

1 Introduction

The identification of reliable biomarkers from high-dimensional genomic data is a primary challenge in bioinformatics, defined by the ($d > n$) problem. Datasets in this domain often contain 30000+ features (for typical transcriptomics datasets) or 300+ features (for miRNA data) for fewer than 100-200 samples, as in this study's case. A key failing of many feature selection methods in this context is instability: the selected feature sets are not reproducible and vary wildly with small data perturbations. This severely limits their clinical and biological reliability.

This instability arises from the curse of dimensionality: when $d \gg n$, many feature subsets achieve similar predictive performance, making the solution highly sensitive to small perturbations in the data. Correlated features exacerbate this, as the regularization path can arbitrarily select one feature over an equally predictive alternative.

The two dominant approaches to this problem present a trade-off. The Lasso algorithm [1] provides minimal (sparse) solutions but can be unstable when predictors are highly correlated - its L1 penalty tends to select one feature from a correlated group and zero out the others, and the choice is sensitive to minor changes in the data. The Elastic Net algorithm [2] was

*e-mail: ak.yermek@pafos.ai

designed to solve this by grouping correlated variables, resulting in high stability but at the cost of minimalism.

Identifying robust biomarkers is particularly important when they have demonstrated clinical relevance — for example, markers predictive of survival in critically ill patients. Such markers may otherwise be missed by unstable feature selection methods.

The goal is to achieve *both* minimalism and robustness. We present StackFeat, a novel iterative algorithm that cumulatively aggregates feature scores to find a stable, minimal, and highly predictive set of markers. This method adapts the concept of "Stacked Generalization" [3]; instead of stacking *predictions*, it stacks *feature importance signals* (score and frequency) from multiple methods to build a robust consensus. We validate this method against a recent 2023 study by Gao et al., demonstrating our method's superiority on the same dataset (GSE240888) [4].

2 Materials and Methods

2.1 Dataset and Benchmark

We utilized the public miRNA expression dataset GSE240888 from the NCBI Gene Expression Omnibus (GEO). This dataset, first analyzed by Gao et al., contains profiles from 122 patients (COVID-19 vs. healthy controls). After preprocessing, the feature space consists of **332 microRNA genes (miRNAs)**, presenting a significant dimensionality challenge (332 features vs 122 samples). We benchmarked our algorithm against the 9-marker set from Gao et al., which achieved a 0.907 AUC.

2.2 The StackFeat Algorithm

We designed a fully automated algorithm to identify a minimal, stable set of predictive features. The method is an iterative process where gene scores are cumulatively aggregated until the classifier's aggregate performance score converges.

The algorithm's logic is depicted in Figure 1 and proceeds in the following steps:

1. **Initialization:** Set iteration counter $t = 1$. Initialize empty dictionaries w and c to accumulate gene-level statistics across all iterations, where w_j stores cumulative signed coefficients and c_j stores selection counts for each gene j.

2. **Reshuffling:** Using seed $s_0 + t$, randomly reshuffle the dataset to create new fold assignments for iteration t.

3. **Nested Cross-Validation:**

   - *Outer CV Loop:* Perform k-fold stratified cross-validation ($k = 10$) on the dataset
   - *Inner CV Loop:* Within each outer fold f, run ElasticNetCV to obtain coefficients $\hat{\beta}^{(t,f)}$
   - *Model Training:* Train an ensemble estimator on the selected features and evaluate performance

4. **Score Aggregation:** After completing all k folds, update cumulative statistics:

   - Signed coefficients: $w_j \leftarrow w_j + \sum_{f=1}^{k} \hat{\beta}_j^{(t,f)}$
   - Selection counts: $c_j \leftarrow c_j + \sum_{f=1}^{k} \mathbf{1}[\hat{\beta}_j^{(t,f)} \neq 0]$

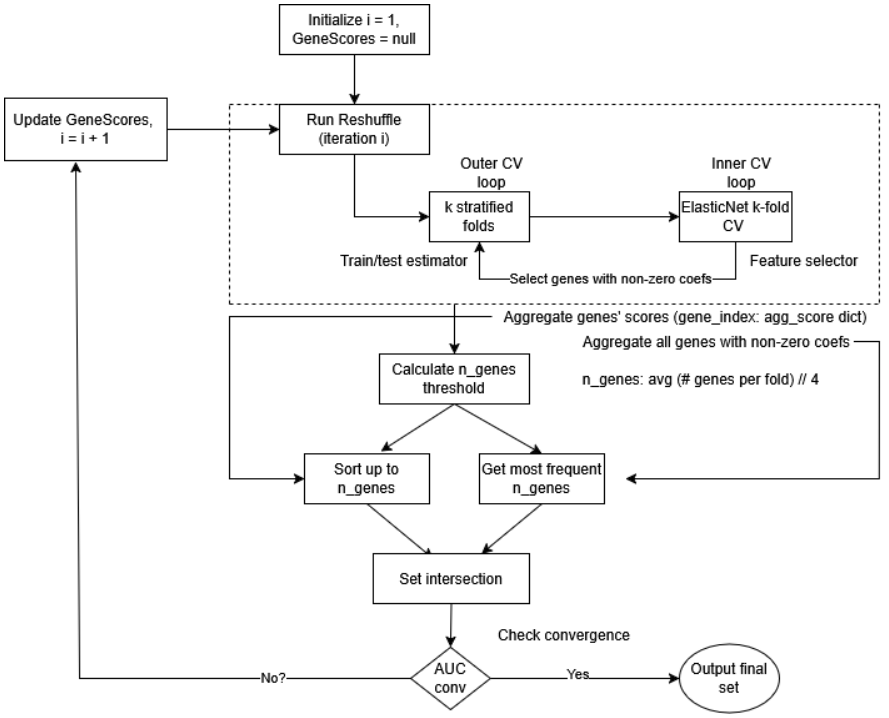

Figure 1. StackFeat algorithm workflow.

5. **Dynamic Thresholding:** Set candidate set size:

$$m = \left\lfloor \frac{\text{mean(genes\_per\_fold)}}{4} \right\rfloor$$

6. **Dual-Criterion Set Intersection:** Generate the current feature set:

$$S^{(t)} = \underbrace{\{j : |w_j| \text{ in top } m\}}_{S_w^{(t)}} \cap \underbrace{\{j : c_j \text{ in top } m\}}_{S_c^{(t)}}$$

7. **Convergence Check:** Convergence requires two consecutive iteration differences below tolerance ε:

$$|AUC^{(t)} - AUC^{(t-1)}| < \varepsilon \quad \text{and} \quad |AUC^{(t-1)} - AUC^{(t-2)}| < \varepsilon$$

8. **Iteration:**

- If converged: Output final feature set $S^{(t)}$
- If not converged: Set $t \leftarrow t + 1$, return to Step 2

We set $\varepsilon = 0.02$ empirically for this dataset. A more stringent threshold (e.g., 0.01) was unnecessary because the two-consecutive-iterations requirement already guards against false convergence.

2.3 Dual-Criterion Selection

Accumulated Statistics. Over T iterations with k folds each, we accumulate two statistics for each feature j:

Cumulative signed coefficient:

$$w_j^{(T)} = \sum_{t=1}^{T} \sum_{f=1}^{k} \hat{\beta}_j^{(t,f)} \tag{1}$$

Selection count:

$$c_j^{(T)} = \sum_{t=1}^{T} \sum_{f=1}^{k} \mathbf{1}[\hat{\beta}_j^{(t,f)} \neq 0] \tag{2}$$

Here $w_j^{(t)}$ denotes the cumulative value of w_j after iteration t, and similarly for $c_j^{(t)}$.

Population Quantities. Define the expected coefficient and selection probability:

$$\bar{\mu}_j = \mathbb{E}\left[\hat{\beta}_j^{(t,f)}\right], \quad p_j = P\left(|\hat{\beta}_j^{(t,f)}| > 0\right) \tag{3}$$

where t indexes iterations and $f \in \{1, \ldots, k\}$ indexes folds.

Convergence. Let $\tilde{\beta}_j^{(t)} = \sum_{f=1}^{k} \hat{\beta}_j^{(t,f)}$ denote the iteration-level sum. Since iterations are independent, by the law of large numbers:

$$\frac{w_j^{(T)}}{T} \to \mathbb{E}[\tilde{\beta}_j^{(t)}] = k \cdot \bar{\mu}_j \tag{4}$$

where the last equality follows from linearity of expectation. Dividing by k:

$$\frac{w_j^{(T)}}{T \cdot k} \to \bar{\mu}_j, \qquad \frac{c_j^{(T)}}{T \cdot k} \to p_j \tag{5}$$

Repeated cross-validation thus provides a resampling-based estimate of population-level feature importance.

Selection Rule. At iteration t, using the cumulative statistics $w_j^{(t)}$ and $c_j^{(t)}$ accumulated so far, we rank features by $|w_j^{(t)}|$ and by $c_j^{(t)}$, selecting the top m under each criterion:

$$S_w^{(t)} = \{j : |w_j^{(t)}| \text{ ranks in top } m\} \tag{6}$$

$$S_c^{(t)} = \{j : c_j^{(t)} \text{ ranks in top } m\} \tag{7}$$

The final selected set is their intersection:

$$S^{(t)} = S_w^{(t)} \cap S_c^{(t)} \tag{8}$$

This dual requirement guards against two failure modes that single-criterion methods miss:

Sign-inconsistent: A feature selected frequently but with inconsistent coefficient sign across folds—contributions cancel, yielding low $|w_j|$.

Infrequent but consistent: A feature overshadowed by correlated alternatives, selected rarely but with consistent direction when selected.

Comparison to Stability Selection. Unlike stability selection [5], which thresholds on selection frequency alone, StackFeat requires both coefficient consistency and selection frequency. A feature selected in every fold but with alternating sign would pass stability selection but be rejected by StackFeat. Additionally, StackFeat uses convergence-based stopping rather than a fixed iteration count B, allowing the algorithm to adapt to dataset complexity.

Feature type	$\lvert w_j \rvert$	c_j	Outcome
True signal	high	high	Selected
Noise	low	low	Rejected
Sign-inconsistent	low	high	Rejected by S_w
Infrequent but consistent	high	low	Rejected by S_c

Table 1. Failure modes addressed by dual-criterion selection.

3 Results

3.1 Benchmarking and Dimensionality Reduction

The most significant result of our methodology is the extreme dimensionality reduction. The algorithm successfully identified a minimal 5-feature set from the initial 332 miRNAs, a **>98% reduction in features**.

This 5-marker signature was:

- hsa-miR-181b-5p
- hsa-miR-4433b-5p
- hsa-miR-1185-1-3p
- hsa-miR-484
- hsa-miR-150-5p

This minimal 5-gene set also outperformed the benchmark. Our analysis first re-validated the findings from Gao et al. Their "Set 1" (9 markers) achieved a 10x10-fold CV average AUC of **0.907**. As shown in **Figure 2 (Bottom-Left)**, our iterative algorithm identified a 5-feature set that achieved a **peak AUC of 0.925** (at iteration 2).

Figure 2. Feature selection and performance evolution across iterations.

Of the 332 initial features, ElasticNet selected 36 ± 10 features per fold (range: 19–62). After applying the dual-criterion intersection, this was reduced to approximately 5 features, representing an 85% reduction before final convergence.

Of the final 5-gene signature, 4 features were selected in ≥77% of iterations (at least 10 of 13), demonstrating consistent selection across reshuffles.

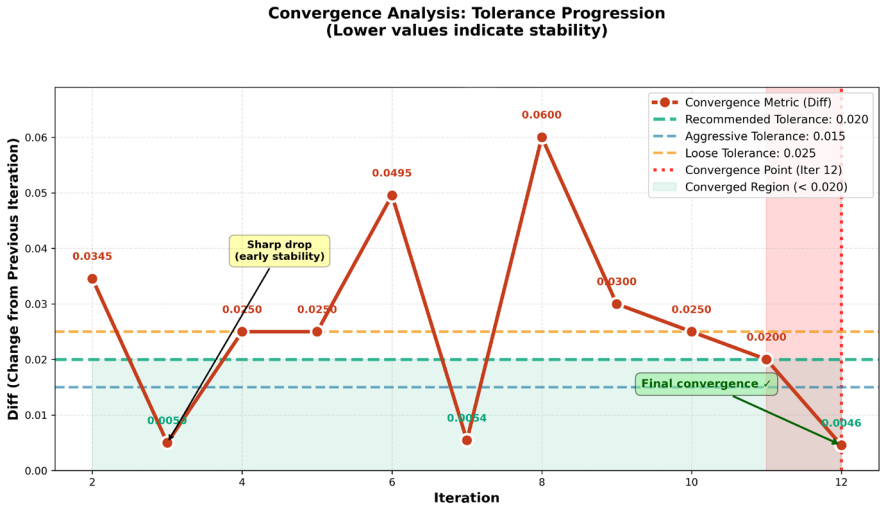

Figure 3. StackFeat convergence. Diff = $|AUC^{(t)} - AUC^{(t-1)}|$. Convergence requires two consecutive diffs below $\varepsilon = 0.02$.

To further validate the robustness of the 5-marker signature, we also compared its performance using 10x10-fold cross-validation, with 2 different classifiers (Ensemble classifier of: ExtraTrees, Logistic Regression, Gaussian Naive Bayes, and KNN, and separately via Random Forest) against the 9-marker set from Gao et al.

The ensemble combines classifiers with complementary assumptions: tree-based (ExtraTrees, 50 trees), linear (Logistic Regression, default parameters), probabilistic (Gaussian Naive Bayes, default parameters), and instance-based (KNN, k=3). This diversity reduces dependence on any single model. Random Forest was evaluated separately as a common baseline. Hyperparameters were chosen empirically; tuning was not the focus of this study.

As shown in **Figure 4**, our 5-marker set significantly outperformed the 9-marker set via ensemble classifier (AUC 0.922 ± 0.007 vs 0.907 ± 0.010; paired t-test, **p = 0.0016**), while achieving comparable performance via Random Forest (~0.88 AUC)—with a 44% smaller feature set.

3.2 Algorithm Convergence

The entire iterative process is visualized in the 4-panel summary of **Figure 2**. Early iterations show instability as cumulative statistics accumulate; later iterations stabilize as signal separates from noise.

AUC Performance Evolution (Fig. 2, Bottom-Left): This plot shows the performance of the selected feature set at each iteration. Each point represents the mean 10-fold CV AUC for the CurrentSet of that iteration, calculated on that iteration's unique data split. The AUC score fluctuated in early iterations while cumulative statistics accumulated, achieving a **peak**

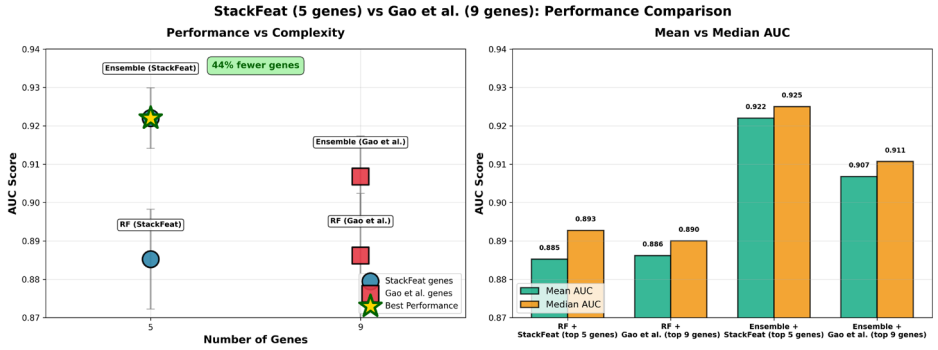

Figure 4. Performance comparison: StackFeat (5 genes) vs Gao et al. (9 genes). Right: Mean and median AUC across 10×10-fold CV. Close agreement indicates stable performance; median slightly higher suggests few low outlier folds.

score of 0.925 at iteration 2 before stabilizing. The convergence mean AUC across all the last 3 iterations (10-12) was 0.905 ± 0.011.

Feature Count Evolution (Fig. 2, Top-Right): This plot shows that the m threshold consistently selected a minimal set, with the feature count stabilizing at 5 features during the convergence phase (iterations 10-12).

Feature Survival Across Iterations (Fig. 2, Top-Left & Bottom-Right): This heatmap is the most critical result, visualizing the stability of the final 5-gene signature across iterations. It shows that two "anchor" genes (hsa-miR-181b-5p and hsa-miR-150-5p) were selected in 100% of iterations. The remaining 3 markers: hsa-miR-1185-1-3p and hsa-miR-484 stabilized quickly, while hsa-miR-4433b-5p flickered in and out (e.g., absent at iterations 0-6, present at 7, absent at 8) before being locked in as part of the final, converged 5-gene set.

The convergence process is detailed further in **Figure 3**. This plot tracks the diff (change from the previous iteration's AUC) against our convergence tolerance ($\varepsilon = 0.02$). Early iterations show high variance, with spikes at iterations 2, 6, and 8. The algorithm enters the converged region (diff < 0.02) after iteration 10, and meets the two-consecutive-diffs criterion at iteration 12.

3.3 Long-term Reproducibility Validation

To validate reproducibility, we re-executed the algorithm 1.5 years after the original analysis using updated software libraries. Each iteration uses a deterministic seed (seed $= s_0 + t$), ensuring reproducibility within a run while providing variation across iterations. The rerun required 12 iterations (vs. 6 originally) due to differences in library defaults, meaning 6 additional iterations with entirely different CV splits. Despite these differences, the algorithm converged to the identical 5-gene signature (hsa-miR-181b-5p, hsa-miR-4433b-5p, hsa-miR-1185-1-3p, hsa-miR-484, and hsa-miR-150-5p).

This demonstrates true solution convergence—the algorithm repeatedly identifies the same biological features regardless of the specific sequence of data perturbations—rather than merely achieving performance convergence. This strongly suggests that these 5 genes represent genuine discriminative markers for COVID-19 vs. healthy classification, not statistical artifacts.

4 Discussion

The results demonstrate a clear advantage of our iterative methodology. The algorithm achieves extreme dimensionality reduction (from 332 features to 5) while significantly improving predictive performance (AUC 0.922 vs 0.907, p = 0.0016) over the benchmark's 9-gene set.

The central innovation is the **dual-criterion iterative aggregation**. By accumulating both scores and frequencies over multiple reshuffles, the algorithm maintains two complementary views of feature importance. At each iteration, the feature set is refined through the intersection of top features by cumulative scores and top features by selection frequency.

Early iterations may show instability in the selected feature set as cumulative statistics build up. As iterations progress, signal accumulates while noise tends to cancel, and the stable feature set emerges. This is evident in Figure 2, where core markers (hsa-miR-181b-5p, hsa-miR-150-5p) are stable from the first iteration, while borderline features (hsa-miR-4433b-5p) fluctuate before stabilizing.

This approach differentiates StackFeat from other stability-focused methods, such as Stability Selection [5]. While Stability Selection passively filters for individual features that appear most frequently across subsamples, StackFeat actively integrates two stability metrics—coefficient magnitudes and selection frequencies—through iterative set intersection. The algorithm's convergence is determined by monitoring the stability of predictive performance (AUC), ensuring that feature set refinement stops when both feature composition and model performance have stabilized.

This process leverages the regularization of ElasticNet [2] while ultimately achieving Lasso-like [1] sparsity. While these results are strong, a full validation would require benchmarking against standard sparse methods (e.g., single-pass Lasso or full ElasticNet on all 332 features) to compare the final feature count and AUC.

Importantly, the 5 identified markers show strong biological coherence, suggesting they represent genuine mechanistic drivers rather than statistical artifacts.

hsa-miR-150-5p is a known inflammation marker, inversely correlated with disease severity in critically ill patients and identified as a key immune response regulator [4, 8]. Notably, it is also a significant marker for Dengue Hemorrhagic Fever and acute viral infection, implying a shared acute viral response pathway [9]. Additionally, this biomarker was found to be critical for survival of critically ill patients of COVID-19 [10], and immune response to successfully clear SARS-CoV-2 virus [11]. In our methodology, in all of the iterations / reshuffles, hsa-miR-150-5p was present (among the top-2 in terms of stability) and was at the top.

hsa-miR-1185-1-3p was identified in a separate study as having one of the most significant FDRs for COVID-19 and is linked to the immune response [12], as well as one of miRNA with a significant predicted binding site in the SARS-CoV-2 reference genome [13].

hsa-miR-4433b-5p is listed among the top 4 DE miRNAs for COVID-19 patients with oxygen requirements [14].

The remaining markers, **hsa-miR-181b-5p** and **hsa-miR-484**, are also cited in connection with the host immune response to viral infection and COVID-19 [14, 15]. This strong external validation confirms the biological relevance of the 5-gene set.

At the same time, **hsa-miR-181b-5p** specifically has not yet been widely cited in the literature in connection with COVID-19, which is a surprising finding, as in our study it is one of the 2 top stable (and therefore predictive) gene sets, one that was present across all iterations / reshuffles and thus had a higher score. Perhaps future studies will elucidate the role of this biomarker in response to SARS-CoV-2 / COVID-19 specifically, as nevertheless there

are pre-covid studies linking hsa-mir-181b-1 (along with its target CYLD) with regulation of inflammatory pathways [16].

Finally, the automation and efficiency of the algorithm are a significant practical advantage. The entire process completed in 197 seconds on a consumer-grade CPU.

5 Conclusion

We have presented StackFeat, a novel, fully automated convergent algorithm for robust feature selection in high-dimensional ($d > n$) data. By applying it to a recent COVID-19 miRNA dataset, we demonstrated a >98% dimensionality reduction, producing a 5-gene set that is more minimal and more predictive than the benchmark. The algorithm's design provides a fast, reliable, and reproducible tool for identifying robust biomarkers.

The miRNA expression data used in this study are publicly available from the NCBI Gene Expression Omnibus under accession number GSE240888.

References

[1] R. Tibshirani, Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B **58**, 267–288 (1996)

[2] H. Zou, T. Hastie, Regularization and variable selection via the elastic net. J. R. Stat. Soc. B **67**, 301–320 (2005)

[3] D.H. Wolpert, Stacked generalization. Neural Netw. **5**, 241–259 (1992)

[4] J. Gao, E. Kyubwa et al., Circulating miRNA profiles in COVID-19 patients and meta-analysis: implications for disease progression and prognosis. Sci. Rep. **13**, 21656 (2023). https://doi.org/10.1038/s41598-023-48227-w

[5] N. Meinshausen, P. Bühlmann, Stability selection. J. R. Stat. Soc. B **72**, 417–473 (2010)

[6] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees. Mach. Learn. **63**, 3–42 (2006). https://doi.org/10.1007/s10994-006-6226-1

[7] L. Breiman, Random forests. Mach. Learn. **45**, 5–32 (2001)

[8] Y. Ding, S. Tang et al., Plasma miR-150-5p as a biomarker for chronic obstructive pulmonary disease. Int. J. Chron. Obstruct. Pulmon. Dis. **18**, 399–406 (2023). https://doi.org/10.2147/COPD.S400985

[9] H. Hapugaswatta, P. Amarasena et al., Differential expression of selected microRNA and putative target genes in peripheral blood cells as early markers of severe forms of dengue. medRxiv (2019). https://doi.org/10.1101/19002725

[10] A. Fernandez-Pato et al., Plasma miRNA profile at COVID-19 onset predicts severity status and mortality. Emerg. Microbes Infect. **11**, 676–688 (2022). https://doi.org/10.1080/22221751.2022.2038021

[11] Y. Yang, D. Fang et al., Circulating microRNAs as emerging regulators of COVID-19. Theranostics **13**, 125–147 (2023). https://doi.org/10.7150/thno.78164

[12] A. Fernandez-Pato, Host miRNA differences by COVID-19 severity: identification of age and sex bias, Master thesis, Universidad Autónoma de Madrid (2021)

[13] J.T. Chow, L. Salmena, Prediction and analysis of SARS-CoV-2-targeting microRNA in human lung epithelium. Genes **11**, 1002 (2020). https://doi.org/10.3390/genes11091002

[14] K. Pollet, N. Garnier et al., Host miRNAs as biomarkers of SARS-CoV-2 infection: a critical review. Sens. Diagn. **2**, 12–35 (2023). https://doi.org/10.1039/D2SD00140C

[15] S.R. Trampuz, D. Vogrinc et al., Shared miRNA landscapes of COVID-19 and neurodegeneration confirm neuroinflammation as an important overlapping feature. Front. Mol. Neurosci. **16**, 1123955 (2023). https://doi.org/10.3389/fnmol.2023.1123955

[16] A. Andalib, S. Rashed, The upregulation of hsa-mir-181b-1 and downregulation of its target CYLD in the late-stage of tumor progression of breast cancer. Indian J. Clin. Biochem. **35**, 312–321 (2019). https://doi.org/10.1007/s12291-019-00826-z