# Supplementary Information
# Maximally informative feature selection using Information Imbalance: Application to COVID-19 severity prediction

**Romina Wild[1], Emanuela Sozio[2,3], Riccardo G. Margiotta[1], Fabiana Dellai[2], Angela Acquasanta[2], Fabio Del Ben[4], Carlo Tascini[2,4], Francesco Curcio[2,3], and Alessandro Laio*[1,5]**

[1]International School for Advanced Studies (SISSA), Via Bonomea 265, Trieste, Italy
[2]Infectious Disease Unit, Azienda Sanitaria Universitaria Integrata di Udine (ASU FC), Via Pozzuolo 330, Udine, Italy
[3]University of Udine, 33100, Via delle Scienze 206, Udine, Italy
[4]Department of Medicine (DAME), University of Udine, Via Palladio 8, 33100, Udine, Italy
[5]The Abdus Salam International Centre for Theoretical Physics (ICTP), Strada Costiera 11, Trieste, Italy
*laio@sissa.it

July 2023

## 1 Predictive power for patients without the optimal input tuples

The optimal 13-plet from our study is only available for about 100 patients. Our method, however, can find for each patient their patient-specific best n-plet, with a slight loss of predictive accuracy when averaged over all patients. This makes sense since 102 patients have the features of the best 13-plet, while the rest the patients do not have complete data for these features. Hence, their respective $\Delta_w$-optimized n-tuple has higher (worse) Information Imbalance than the optimal 13-plet, which influences the prediction accuracy. The patient-specific optimal n-plets were found in a leave-one-out (LOO) approach by calculating the Weighted Information Imbalance using all other patients who had the relevant features for each tested feature-tuple. Then we performed a 10-NN prediction of severity for each patient, using their optimal n-plet features in a LOO cross validation.

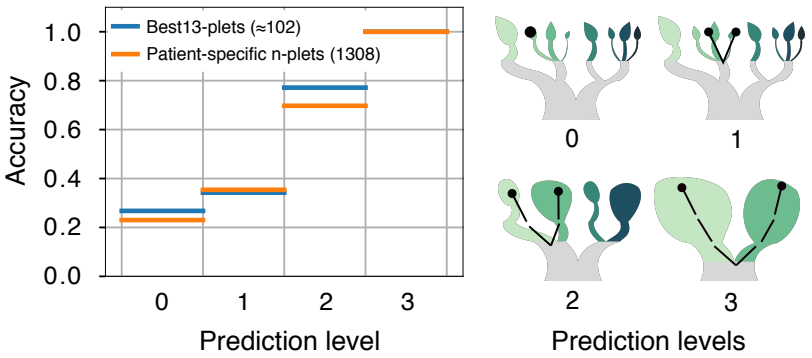


**Figure S1.** Accuracies of 10-NN predictions at given prediction levels by the optimal 13-plets (averaged over the ten best), which are available in roughly 102 patients, *vs.* accuracies using for each patient in the database (1308 patients) their optimal input feature tuple. The accuracy corresponds to the fraction of patients predicted correctly at a given prediction level. The prediction level corresponds to the maximum tolerated distance of the true *vs.* the predicted class on the severity tree, as depicted on the right. Together, the lines of one color can be considered the CDF of the fraction of patients predicted correctly.

# 2 Correlation and Information Imbalance between input features

## 2.1 Information Imbalance and correlation within $\Delta_w$-optimized n-plets

The Weighted Information Imbalance approach (main paper) automatically leads to the selection of tuples of features which are practically uncorrelated. This is demonstrated by calculation of the Pearson correlation coefficient and the pairwise classic Information Imbalance (introduced in ref. [3]) for all the numerical features contained in the $\Delta_w$-optimized n-plets (Fig. S2.). The mean of pairwise correlations of numerical features in *e.g.* the best 13-plet (Fig. S2 blue outline), is $r = 0.02$, and the pairwise Information Imbalances mean is $\Delta = 0.96$, both of which point towards a high degree of orthogonality.

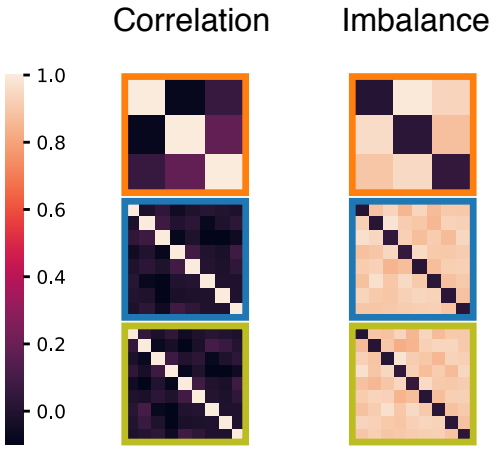


**Figure S2.** The pairwise Pearson correlation heat mat of the numerical features of the $\Delta_w$-optimized n-plets(3, 13, and 20) from Figure 2.a of the main paper, and the pairwise, classical Information Imbalances of the same n-plets.

## Information Imbalance between numerical patient features

We also use the classic Information Imbalance to investigate the relationships between the input features. We consider the 90 numerical input features for which it is possible to estimate the standard Information Imbalance introduced in ref. [3]. We computed the Information Imbalance $\Delta$ between each pair of features using the implementation in the Python package DADApy [2]. $\Delta(A \to B)$ is close to zero if feature A predicts feature B well. It is close to one if feature A does not provide information on feature B. For each pair of features we also computed the standard Pearson and Spearman correlation coefficients $r$ and $\rho$, which are $\pm 1$ in the case of a perfect positive or negative correlation, and 0 if there is no correlation. If two features correlate strongly, $\Delta(A \to B)$ and $\Delta(B \to A)$ should both be small and similar numbers, if both predict each other to an equal amount. However, if one feature predicts the other, but not vice versa, there exists an asymmetric correlation, and this is reflected in an asymmetric Information Imbalance. This phenomenon, as we will see, is not captured by Pearson and Spearman correlations.

In the table in Fig. S3a, we report the Information Imbalance and the correlation coefficients between the 20 pairs of features with the lowest $\Delta(A \to B)$. To highlight some possible relationships, we plot some of these feature as a function of each other in the bottom panels (Fig. S3b).

The Information Imbalance faithfully captures features which have strong correlations with each other. The top-eight positive correlation couples (all $r > 0.8$) are contained in the top-20 Information Imbalances. A clear sanity check is displayed in the first rows: the two different laboratory methods for the glomerular filtration rate (GFR and GFR.1), which hold the same values, display perfect correlations and extremely low Information Imbalances. This is also true for the prothrombin time (- and international normalized) ratios (PT/INR and PTR), where one is just a normalized version of the other. Correlation and Information Imbalance pick up on the linear relationship between hematocrit (EHCT) and hemoglobin (EHB). EHCT is the percentage of volume occupied by red blood cells relative to whole blood, and therefore is often related to hemoglobin (EHB). Also, the strongest negative correlation pairing is in the top-20 imbalance table (row 6). Thus the strongest correlations account for nine rows in the top-20 Information Imbalances. The other eleven rows are made up by pairings which have less strong correlations, but six of them have

a

| No. | Feature A | Feature B | Pearson r(A,B) | Spearman ρ(A,B) | Δ(A→ B) | Δ(B→ A) |
|---|---|---|---|---|---|---|
| 1 | GFR.1 | GFR | 1.000 | 1.000 | 0.037 | 0.037 |
| 2 | PT/INR | PTR | 1.000 | 0.998 | 0.049 | 0.050 |
| 3 | EHCT | EHB | 0.981 | 0.976 | 0.236 | 0.261 |
| 4 | EWBC | FLNEU | 0.874 | 0.959 | 0.276 | 0.283 |
| 5 | PaO2/FiO2 | PaO2 | 0.177 | 0.279 | 0.351 | 0.805 |
| 6 | FiO2 | PaO2/FiO2 | -0.906 | -0.770 | 0.425 | 0.441 |
| 7 | EMCH | EMCV | 0.947 | 0.916 | 0.448 | 0.450 |
| 8 | EHCT | ERBC | 0.861 | 0.853 | 0.527 | 0.541 |
| 9 | CRE | GFR | -0.533 | -0.842 | 0.560 | 0.592 |
| 10 | CRE | GFR.1 | -0.533 | -0.842 | 0.560 | 0.579 |
| 11 | EHB | ERBC | 0.820 | 0.819 | 0.561 | 0.592 |
| 12 | Oxygen saturation | PaO2 | 0.599 | 0.807 | 0.571 | 0.593 |
| 13 | BILD | BILT | 0.962 | 0.874 | 0.573 | 0.598 |
| 14 | TROP | FIBCL | -0.496 | -0.507 | 0.574 | 0.713 |
| 15 | IGM | HDL | -0.003 | 0.109 | 0.612 | 0.810 |
| 16 | IGG | HDL | -0.295 | -0.055 | 0.628 | 0.876 |
| 17 | A-a gradient | PaCO2 | -0.455 | -0.695 | 0.651 | 0.691 |
| 18 | BNP | Birth | -0.504 | -0.708 | 0.662 | 0.792 |
| 19 | TROP | Anion gap | 0.063 | -0.382 | 0.699 | 0.858 |
| 20 | HBA1CM | IGA | 0.100 | 0.042 | 0.711 | 0.804 |
| ... | ... | ... | ... | ... | ... | ... |
| 21 | BMI | GFR | -0.077 | -0.044 | 1.029 | 1.040 |

b

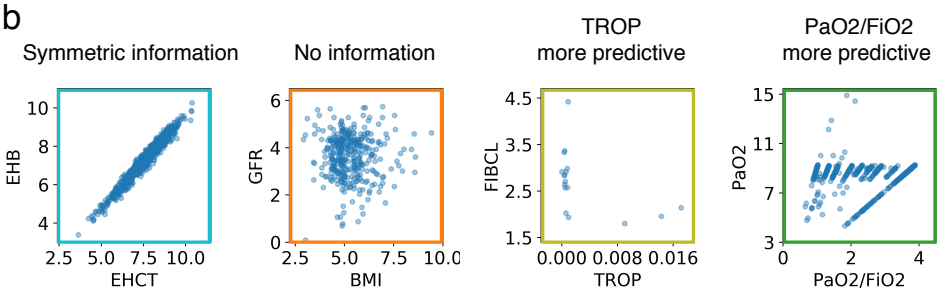


**Figure S3.** **a**: Features ordered according to the lowest Information Imbalances towards another feature, and their Pearson and Spearman correlation coefficients. Yellow colored rows have notably asymmetric Information Imbalances, where $|\Delta(A \to B) - \Delta(B \to A)| > 0.1$. **b**: Scatter plots of several of the feature *vs.* each other from A. The values are the normalized features.

high asymmetries in their Information Imbalances towards each other (yellow rows in S3a), describing a relationship where one variable is more informative about the other than *vice versa*. These six pairings have predominantly very low correlations, showcasing that correlation fails to identify these asymmetric relationships. The effect is especially pronounced in row five, where PaO2/FiO2 (oxygen partial pressure over fractional inspired oxygen) has a low Information Imbalance towards PaO2 (oxygen partial pressure) and such explains this feature space well, while the same is not true *v.v.* This can be used as a proof of concept because indeed PaO2/FiO2 is the value of PaO2 divided by FiO2 (fractional inspired oxygen) - a simple relationship via one confounding variable which is not detected by correlation (r=0.177). It should be noted that, from a clinical point of view, measuring the PaO2/FiO2 ratio can become very challenging: if patients are not on invasive mechanical ventilation, it is almost impossible to know the exact FiO2, because the devices deliver a variable inspired oxygen concentration. Information Imbalance also detected similar cases where the exact relationship is not known: Here we report asymmetric relationships between troponin (TROP), a well-known marker of cardiac injury, and tissue damage marker fibrinogen (FIBCL), as well as between the immunoglobulins IGM / IGG and the high-density lipoprotein (HDL). TROP values are more predictive of FIBCL values than the other way around. Fibrinogen is a plasma acute-phase reactant protein produced by the liver and is a major coagulation factor. Its concentration increases with inflammation, and it is traditionally considered a risk factor for cardiovascular disease [7, 5], which might explain the connection to troponin. Troponin, on the other hand, is a very specific marker: recent studies showed that troponin dosage should be considered as a prognostic indicator in all patients with

moderate/severe COVID-19 at hospital admission and in the case of clinical deterioration [**Lippi2020**]. Retrospective data have placed a strong emphasis on the possibility that acute myocardial injury represents a critical component in the development of serious complications in patients hospitalized with COVID-19 [1, 6, 4]. To the best of our knowledge, there is no literature concerning the exact relationships between IGM / IGG and HDL.

While causality-based feature selection methods only aim at finding causalities between the predicting features and the target classes [8], we point out here that inter-feature asymmetric relationships, such as inter-feature causalities, could also be important. They are not captured by standard correlation analyses, and they could lead to identify redundancy effects when tuples of features are used for predictive purposes. This pairwise Information Imbalance between the features was used as a sanity check of the prediction tuples used in the previous section. All $\Delta_w$-optimized tuples were indeed non-redundant, with Information Imbalances of $> 0.85$ between each other.

## References


[1] Tao Chen et al. "Clinical characteristics of 113 deceased patients with coronavirus disease 2019: retrospective study". In: *BMJ* 368 (2020). DOI: https://doi.org/10.1136/bmj.m1091. URL: https://www.bmj.com/content/368/bmj.m1091.

[2] Aldo Glielmo et al. "DADApy: Distance-based analysis of data-manifolds in Python". In: *Patterns* 3.10 (2022), p. 100589. ISSN: 2666-3899. DOI: https://doi.org/10.1016/j.patter.2022.100589. URL: https://www.sciencedirect.com/science/article/pii/S2666389922002070.

[3] Aldo Glielmo et al. "Ranking the information content of distance measures". In: *PNAS Nexus* 1.2 (Apr. 2022). ISSN: 2752-6542. DOI: https://doi.org/10.1093/pnasnexus/pgac039. URL: https://doi.org/10.1093/pnasnexus/pgac039.

[4] Tao Guo et al. "Cardiovascular Implications of Fatal Outcomes of Patients With Coronavirus Disease 2019 (COVID-19)". In: *JAMA Cardiology* 5.7 (July 2020), pp. 811–818. ISSN: 2380-6583. DOI: https://doi.org/10.1001/jamacardio.2020.1017. URL: https://doi.org/10.1001/jamacardio.2020.1017.

[5] Gordon DO Lowe, Ann Rumley, and Ian J Mackie. "Plasma fibrinogen". In: *Annals of Clinical Biochemistry* 41.6 (2004), pp. 430–440. DOI: https://doi.org/10.1258/0004563042466884. URL: https://doi.org/10.1258/0004563042466884.

[6] Shaobo Shi et al. "Association of Cardiac Injury With Mortality in Hospitalized Patients With COVID-19 in Wuhan, China". In: *JAMA Cardiology* 5.7 (July 2020), pp. 802–810. ISSN: 2380-6583. DOI: https://doi.org/10.1001/jamacardio.2020.0950. URL: https://doi.org/10.1001/jamacardio.2020.0950.

[7] Thomas M Stulnig. "C-reactive protein, fibrinogen, and cardiovascular risk". In: *New England Journal of Medicine* 368.1 (2013), pp. 84–86. DOI: https://doi.org/10.1056/NEJMc1213688.

[8] Kui Yu et al. "Causality-Based Feature Selection: Methods and Evaluations". In: *ACM Comput. Surv.* 53.5 (Sept. 2020). ISSN: 0360-0300. DOI: https://doi.org/10.1145/3409382.