Appendix A Supporting Information

A.1 Mapping metastasis bio-markers to gene names

The most important probe identified by Shapley values: **'1563882_a_at'** maps to Rap guanine nucleotide exchange factor 5 (RAPGEF5). According to the information extracted from the DAVID online tool, the gene Guanine nucleotide exchange factor (GEF) for RAP1A, RAP2A and MRAS/M-Ras-GTP is associated with MRAS, and inhibits Rap1 activation. It is widely expressed with highest levels in the brain.

The Ras-related protein M-Ras, is a protein that is encoded by the MRAS gene on the third chromosome in humans. It is expressed in many tissues and different cell types. The protein has the function of a signal transducer for a wide variety of signaling pathways, including those promoting neural and bone formation as well as tumor growth [30].

Rap1 is a Ras-like small GTPase that is activated by many extracellular stimuli and strongly implicated in the control of integrin-mediated cell adhesion. Rap1 also plays a key role in the formation of cadherin-based cell-cell junctions. Furthermore, the inhibition of Rap1 generates immature adherens junctions, whereas activation of Rap1 tightens cell-cell junctions. Interestingly, Rap1 guanine nucleotide exchange factors, such as C3G and PDZ-GEF, are directly linked to E-cadherin or to other junction proteins. Moreover, several junction proteins, such as afadin/AF6 and proteins controlling the actin cytoskeleton, function as effectors of Rap1 [31].

For the bone metastasis data set, one of the most important bio-markers has the identifier **'1558688_at'**. This bio-marker corresponds to the STX17 divergent transcript (STX17-DT), which is identified in other research as controlling the cell growth in tissues of colorectal cancer [32]. However, here we show that this gene can also have predictive power in the process of identifying bone metastasis from breast cancer. The association with colorectal cancer can be a reason for some of the missclassified patient samples with this model, since colorectal cancer might be one of the prevalent metastases in the other metastasis data set.

The probe **'201094_at'** has been identified as one of the most significant for identifying breast metastasis. However, we must keep in mind that the model identifying breast metastasis was not very accurate, so this bio-marker should not be considered as very important with high certainty in this particular case. The bio-marker corresponds to ribosomal protein S29(RPS29). Previous studies have associated this ribosomal protein with colorectal cancer [33].

Additionally, surfactant protein C (SFTPC), with probe identifier **'214387_x_at'** has been recognized as the most important bio-marker for predicting lung cancer. Indeed, this gene is associated with the process of gaseous exchange between an organism and its environment. Apart from some samples which have an extremely high expression of this gene and result with a larger probability of predicting lung metastasis from breast cancer, other values which are not as extreme, either high, low or moderate, result in a small lowering of the probability of lung metastasis from breast cancer.

What is also interesting to emphasize is the one bio-marker identified for detecting of whether an arbitrary metastasis is present or not. This bio-marker has the identifier

'1552768_at', which maps to calcium/calmodulin dependent protein kinase kinase 1 (CAMKK1). CAMKK1 has been associated with risk of lung cancer [34][35][36]. In the general case, more active expression of the gene corresponds to a higher probability of predicting an occurrence of an arbitrary metastasis, which is consistent with other research in the field. Furthermore, in the case pf our study, this bio-marker was solely enough discriminative to correctly separate all patients not having a metastasis from breast cancer, from all of those having an arbitrary metastasis (lung, bone, brain, breast, other). This might mean that, not only is this gene associated with severity of lung cancer, but it might also be associated with other types of metastases in the human body. Further research is necessary to provide novel insights in this area.

Furthermore, we focus our efforts on investigating the selected bio-markers, using the Boruta search algorithm, for the other metastasis data set. The analysis of the other metastasis bio-markers has two implications: firstly, we have provided results which identify a concrete set of bio-markers, able to detect the occurrence of an arbitrary type of breast cancer metastasis, and secondly, by mapping the probe identifiers to actual genes and their functions, we provide an unsupervised characterization of the types of metastasis which might have been grouped under the label 'other'. The explainable figures of the importance of each of the probes for other metastasis is given in Fig.A1. In the following text we give a brief overview of the 5 most important identified bio-markers and their implications and connections to certain types of tumors.

1. **RAB14** - Rab GTPases are localized to various intracellular compartments and are known to play important regulatory roles in membrane trafficking. Rab proteins have key functions in intracellular trafficking in eukaryotic cells, controlling accurate fission and fusion of transport vesicles with their target membranes during biosynthetic/secretory and endocytotic pathways [37]. It acts as Oncogene and Induce Proliferation of Gastric Cancer Cells via AKT Signaling Pathway [38]. In this study it is explored that high expression of RAB14 refers to high probability of other metastasis occuring.
2. **MORN1** - Membrane Occupation and Recognition Nexus (MORN). In other proteins, MORN repeats are responsible for membrane association and stabilization of protein complexes. MORN1 specifically is a general cytoplasmic expression with high expression in glandular cells and renal tubules. It is shown that it is mostly present in prostate cancer [39]. In our case, MORN1 is present in the genes refered to the other metastasis prediction.
3. **PON2** - Paraoxonase-2 (PON2) is an intracellular enzyme exerting a protective role against production of reactive oxygen species within mitochondrial respiratory chain. It is shown that PON2 is a potential bio-marker for skin cancer aggressiveness [40]. In our study this bio-marker is showing high probability of other metastasis occuring.
4. **BCAP29** - BCAP29 is often associated with gastric cancer. A novel DUS4L–BCAP29 fusion transcript was identified in recent research, and it was confirmed that this novel fusion transcript has tumorigenic potential [41]. In our study, we show that high expressions of BCAP29 leads to a higher probability of predicting other metastasis.

5. **CHD5** - Newer evidence has shown that CHD5 functiones as a tumor suppressor gene in gliomas and a variety of other tumor types, including breast, colon, lung, ovary, and prostate cancers. One copy of CHD5 is deleted frequently, however inactivating mutations of the remaining allele are rare. Often, low CHD5 expression is strongly associated with unfavorable clinical and biologic features as well as outcomes in neuroblastomas and many other tumor types [42]. This is consistent with the results of our explainable ML model. From Fig.A1 it can be seen that low expressions of CHD5 (probe id: 213965_s_at lead to a higher probability of predicting the occurence of other metastasis.

From the information presented about the mapping between the Affymetrix identifiers and the human genes for the corresponding bio-markers, we can conclude that the other metastasis data set typically refers to metastasis to the skin, prostate, lung, stomach, and blood cells. Furthermore, bio-markers which previous research relates to lung cancer are also present in the other metastasis data set, although we have included lung metastasis as a separate case in this study. This might suggest that those bio-markers related to lung cancer can also be related to another type of cancer in other organs.