DISSERTATION FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (PHD)

Development and application of a new food composition database to reveal patterns of human milk composition

by

Mayara Lopes Martins

UNIVERSITY OF DEBRECEN DOCTORAL SCHOOL OF NUTRITION AND FOOD SCIENCES

DEBRECEN, 2023

DISSERTATION FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (PHD)

Development and application of a new food composition database to reveal patterns of human milk composition

by **Mayara Lopes Martins**

Supervisor: **Dr. József Baranyi**
Co-supervisor: **Dr. Katalin E. Müller**

UNIVERSITY OF DEBRECEN DOCTORAL SCHOOL OF NUTRITION AND FOOD SCIENCES

DEBRECEN, 2023

TABLE OF CONTENTS

LIST OF ABBREVIATION

HM = human milk

HMC = human milk composition

BMI = Body Mass Index

FAO = Food and Agriculture Organization

FCD= Food Composition Database

INFOODS = International Network of Food Data Systems, Food and Agriculture Organization

EuroFIR = European Food Information Resource Association Internationale Sans but Lucratif

VBA = Visual Basic for Application

1. INTRODUCTION

In early infancy, nutrition is key to the modulation the offspring's health and development driven by a unique plasticity period, in the first 1,000 days of life (KOLETZKO et al 2019; KOLETZKO et al 2017; INDRIO et al, 2017; ROBINSON; FALL, 2012; GLUCKMAN et al 2005). This period is a window of opportunity when infant nutrition plays a crucial role in forming a basis on which health in the adult life will depend (KOLETZKO et al 2019; KOLETZKO et al 2017; INDRIO et al, 2017; ROBINSON; FALL, 2012; GLUCKMAN et al 2005).

It is well-recognized worldwide that the best food for infants over the first period of life is their own mother's breastmilk, abbreviated as HM (human milk) in this study (BARDANZELLU et al, 2020; KOLETZKO et al 2019; ROBINSON; FALL, 2012, WHO, 2001). Exclusive breastfeeding is recommended until the 6th month of life and infant feeding should be continued, along with complementary food, for up to 2 years or beyond (BARDANZELLU et al, 2020; KOLETZKO et al 2019; ROBINSON; FALL, 2012, WHO, 2001).

The uniqueness of HM is rooted in its distinctive composition that is different from any other food (GEORGE et al, 2022; CHRISTIAN et al, 2021; CHRISTIAN et al, 2021; SAMUEL et al 2020; BARDANZELLU et al, 2020). Beyond delivering general nutritional requirements, in terms of both macro- and micronutrients, HM composition (HMC) comprises bioactive components such as cells, microbiota, hormones and antigens, and, together with the nutritional elements, they cover all metabolic, developmental and growth needs of the infant (GEORGE et al, 2022; CHRISTIAN et al, 2021; BARDANZELLU et al, 2020).

HM is a living system that is shaped individually, depending on the infant (gestational age, size, sex, etc.) and maternal (diet, environment, etc.) characteristics, as well as own factors like expression, storage (SHENHAV, AZAD, 2022; AHUJA et al 2022; CHRISTIAN et al 2021). HMC changes over

time, driven by various factors of the mother-milk-infant triad, that makes HM a dynamic "system within a system" (SHENHAV, AZAD, 2022; CHRISTIAN et al 2021; BODE et al 2020).

Nutrition and food scientists face key questions on HMC such as (i) how maternal and infant characteristics influence the dynamics of HM, (ii) how to tailor HM to positively influence the infant's health and development and, (iii) in circumstances, when breastfeeding is not an option, how to optimize and personalize infant formulae that can be also specific to mother-infant pairs (CHRISTIAN et al 2021).

Three main points hinder the advancement of our current knowledge: (i) most studies focus on the impacts of single factors, within the mother-milk-infant triad, on the individual components of HM; (ii) multi-omics methods, which are primary tools to analyse HMC, are used in cross-sectional rather than longitudinal sense and (iii) there is a notable lack of knowledge about advanced computational and statistical tools, and this hinders the adequate analysis of relevant data (SHENHAV, AZAD, 2022).

To explore HMC knowledge and advance the research field, a proper framework to capture and store quantitative data from published studies on HMC is a must (AHUJA et al 2022). Ideally, such database needs to comport the temporal variation of HMC, accompanied by information on maternal, infant and methodological factors affecting it (AHUJA et al 2022). Once this structure exists, available HMC data can be stored and analysed in an analogous manner to a biological system (AHUJA et al 2022; DE WEERTH et al 2022; SHENHAV, AZAD, 2022; CHRISTIAN et al 2021).

At the moment, there is no database that would help unravel the dynamics of HM (FERRAZ DE ARRUDA et al 2023; AHUJA et al, 2022; DELGADO et al 2021; TOURE et al, 2020; KAPSOKEFALOU et al, 2019). Therefore, the present study focuses on the principles of how to build a database on HMC (nutritional and bioactive components) that complies with scientific and IT standards, while taking into account the requirements above.

2. LITERATURE REVIEW

2.1 EARLY INFANCY AND NUTRITION

The nutrition in the early infancy is decisive to one's health in later life (DARLING et al, 2020; KOLETZKO et al 2019). Over the first years of life, our biological systems are rapidly evolving due to a unique developmental plasticity (GILLMAN, 2010). This means that the inherited genotype can trigger diverse (physiological or morphological) phenotypes as responses to different environmental stimuli (HOCHBERG, 2011; BARKER, 2004). The "first 1,000 days of life" covers the period from conception until the age of two years when nutrition highly influences gene expressions and can also affect outcomes in adult life (INDRIO et al, 2017; BHUTTA et al 2013, LUCAS et al, 2005).

Epidemiologist David J. P . Barker and his collaborators are considered pioneers on the topic (GLUCKMAN et al 2005; BARKER, 2004) as they described first, in 1986, the epidemiological correlations between geography-dependent occurrences of ischemic heart diseases in adults and their local dietary intake during early infancy (BARKER et al, 1986). Later, in 1989, the same group performed a study that traced back birth records of more than 5000 men. They found a relationship between low birthweight and higher rate of cardiovascular mortality caused by cardiovascular diseases (BARKER et al, 1989). WHERE IS GEOGRAPHY HERE -JB?

Afterwards, the "developmental origins of adult disease" hypothesis was created by Barker (2004), proposing a link between early infant nutrition and later disease onset in adulthood. The hypothesis was based on four points: (i) a relationship between low birthweight and risk to develop non-communicable disease in the adulthood shown by vast number of studies, (ii) the intra-uterine developmental plasticity reacting to mothers' diet, (iii) the impact of low birthweight and poor nutrition in the first years of life on the metabolism and health outcomes in adulthood, and (iv) the link between low birthweight, followed by rapid weight gain during the infancy, and higher risk of developing chronic diseases in adulthood (BARKER et al, 2004).

A long (and still running) longitudinal household survey confirmed that low birthweight was closely related to diseases before the age of 50 (JOHNSON & SCHOENI, 2011). The researchers found that low birth-weighted infants have greater chances to develop asthma, hypertension, diabetes mellitus, and cardiovascular diseases (stroke, heart attack, or heart disease; see JOHNSON & SCHOENI, 2011). Another milestone was the recognition of the impact of the Dutch famine during the World War II (1944-1945) on the risk of cardiovascular disease (PAINTER et al 2006). Infants of mothers who prenatally exposed their infants to a daily food intake with less than 1000 kcal during any 13-week period of gestation were studied and compared with those who were unexposed (PAINTER et al 2006). As it turned out, the probability that the famine-exposed newborns would develop coronary artery diseases in their later life, before their 61st birthday, increased by two-fold compared to 590 unexposed subjects (PAINTER et al 2006). This study is one of most important evidence that mothers' diet affects the long-term health of the offspring.

Such hypotheses and findings had generated the relatively new research of epigenetics that focuses on the phenomenon that the way genes are expressed, as responses to different external factors, can affect subsequent generations (KOLETZKO et al 2019; INDRIO et al, 2017; BHUTTA et al 2013). Epigenetics can be defined simply as the ability to produce different gene expressions, prompted by different external factors, without altering inherited genetic material (INDRIO et al, 2017). The term was first described in 1942 as the "whole complex of developmental processes" that connects genotype and phenotype (DEICHMANN, 2016).

Since then, numerous studies have established that environmental factors, such as diet, pollutants, lifestyle, are able to affect gene expression without changing the DNA sequence (RAMOS-LOPEZ et al, 2021). The significance of epigenetics has also been confirmed by several studies in the field over the years (KOLETZKO et al 2019; KOLETZKO et al 2017; ROBINSON; FALL, 2012; GLUCKMAN et al 2005).

On the other hand, many studies in developed countries confirmed the relationship between rapid weight gain in healthy term infants during the first year of life with higher probability of developing obesity and cardio-metabolic diseases (GILLMAN, 2010; ONG et al, 2006; BAIRD et al, 2005; MONTEIRO et al, 2005). It is well-explored and known that early life nutrition can be considered as a valuable tool to be

mastered to improve health-outcomes for the offspring (KOLETZKO et al 2019). For instance, the "Early Nutrition Project" is a scientific initiative funded by the European Union that explores the influence of early nutrition and lifestyle on programming the metabolism (KOLETZKO et al 2019). Based on the mentioned evidence of epigenetics, the Early Nutrition Project group educates families on how to promote healthier lives to their offspring by supporting weight maintenance, healthy diet and breastfeeding.

2.2. GUT MICROBIOME AND INFANCY

Considered as our "second genome", the human microbiome is another important player in the modification of gene expression and programming of metabolism and immune system in the early infancy (INDRIO et al 2017; GRICE; SEGRE, 2012). Just as the genome, the infant's gut microbiome is highly influenced by factors like delivery and feeding modes which can affect the "plasticity" towards a beneficial or pathogenic profile depending on the stimuli (INDRIO et al 2017). Besides, the gut microbiome can also interact directly with the gene expression (INDRIO et al 2017).

In that context, the triad: nutrition, microbiome and epigenome are in constant cooperation to fine-tune the development of the infant over the first 1,000 days of life (**Figure 1**).

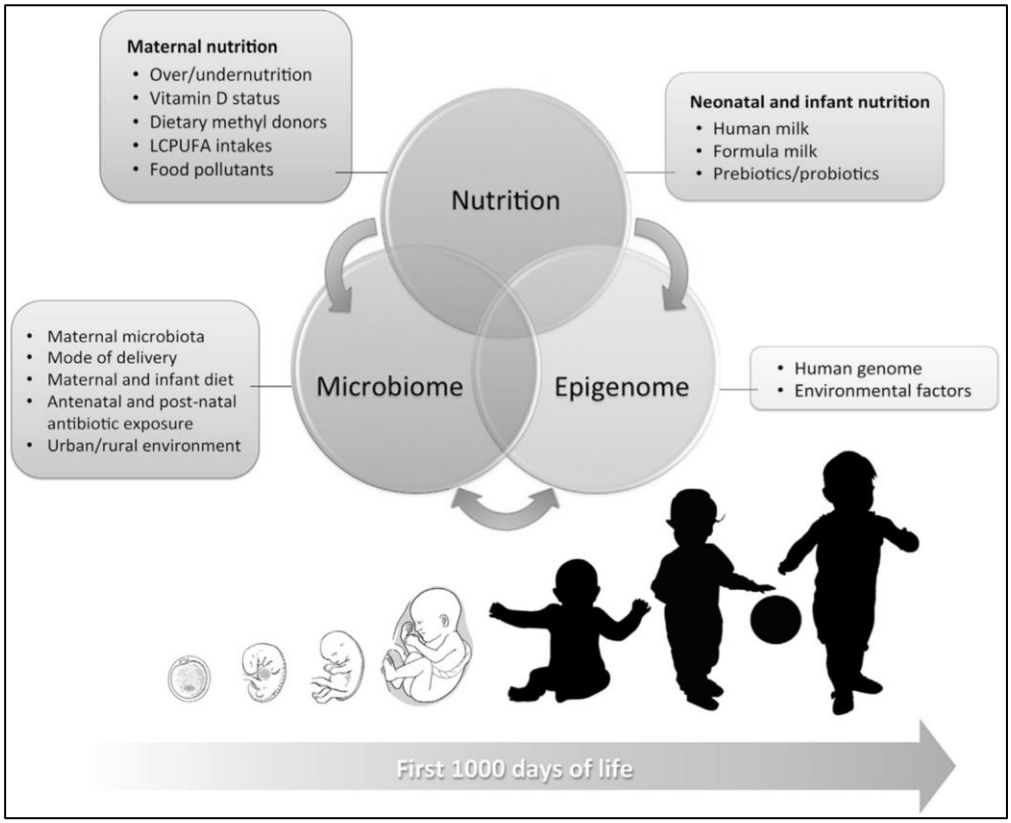


Figure 1. Relationship of nutrition, gut microbiota, and epigenetics over the first 1,000 days of life. Source: INDRIO et al 2017

Based on our current knowledge, the optimal nutrient source for infants is exclusively from breastfeeding for the first 6 months (WHO, 2001). After that period, appropriate food should be complemented by breastfeeding which should last until the infant is 2 years old or even older (BARDANZELLU et al, 2020; KOLETZKO et al 2019; ROBINSON; FALL, 2012, WHO, 2001). This strategy is recommended because HM is not an ordinary food and plays pivotal role in the early infant development (BARDANZELLU et al, 2020; KOLETZKO et al 2019 ROBINSON; FALL, 2012, WHO, 2001).

2.3 HUMAN MILK AND INFANT HEALTH

Mammalian milk is the only foodstuff in nature that is considered perfectly fit for purpose (RUMBOLD et al 2021). Besides, breastfeeding is more than just nutrition (Figure 2) (REY-MARINO; FRANCINO, 2022; HORTA et al 2022, GEORGE et al, 2022; CHEN et al 2022; CHRISTIAN et al, 2021;

CHRISTIAN et al, 2021; SAMUEL et al, 2020; AZAD et al 2018; VICTORA et al 2016). Apart from transferring water and essential macro- and micronutrients to the infant, HM also provides non-nutritional components (microbiota, cells, antigens, hormones, etc.) to create a complex biological system that is unique to a single mother-infant pair, and covers all nutritional, metabolic, developmental and growth needs together with immunological specificities for that particular infant (**Figure 2**) (GEORGE et al, 2022; CHRISTIAN et al, 2021; BARDANZELLU et al, 2020).

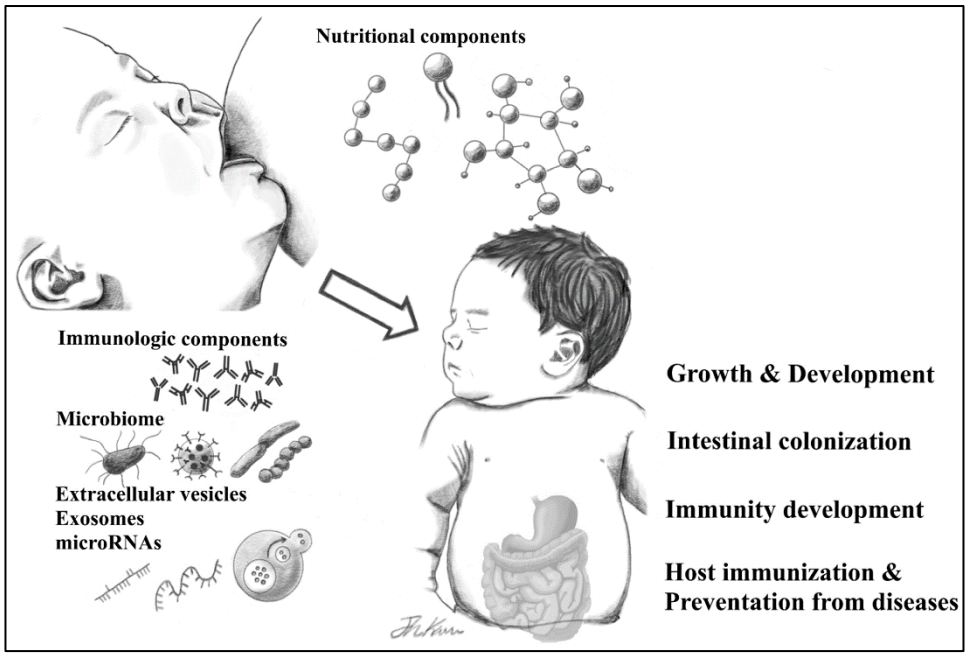


Figure 2. Human milk composition and infant's health. Source: **Yi; Kim, 2021**

By carrying bioactive components, HM has a critical role in the protection of infants from several diseases, which makes breastfeeding responsible for shielding infants from short- and long-term health-related consequences in their life (HORTA et al 2022, GEORGE et al, 2022; CHEN et al 2022; GILA-DIAZ et al 2019; AZAD et al 2018; VICTORA et al 2016).

The meta-analysis of Victora et al. (2016) established that breastfeeding mothers, were less likely to develop type 2 diabetes, ovarian and breast cancer, and their children were also less pruned to infections, malocclusion, overweight and diabetes (VICTORA et al 2016). The Canadian CHILD project showed, on a cohort consisting of more than 2,500 mother-infant pairs, that breastfeeding was inversely related to weight

gain velocity and body mass index (BMI) over the early infancy (AZAD et al 2018). The systematic review and meta-analysis (n=159) of Horta and collaborators (2022) confirmed this finding (HORTA et al 2022). Complementarily, the National Health and Nutrition Examination Survey (1999 - 2014) in the U.S. concluded that infants who were exclusively breastfed for 4 to 6 months had lower probability to develop asthma (CHEN et al 2022). Also, the meta-analysis of Horta et al (2018) showed that breastfeeding had a positive correlation with IQ test scores (HORTA et al, 2018).

HM transfers important factors to the child, such as insulin-like, epidermal, hepatocyte, vascular endothelial and neuronal growth factors that drive main evolving areas as dendrite development, angiogenesis, erythropoiesis and maturation of diverse organs and, generally, biological systems (**Figure 3**).

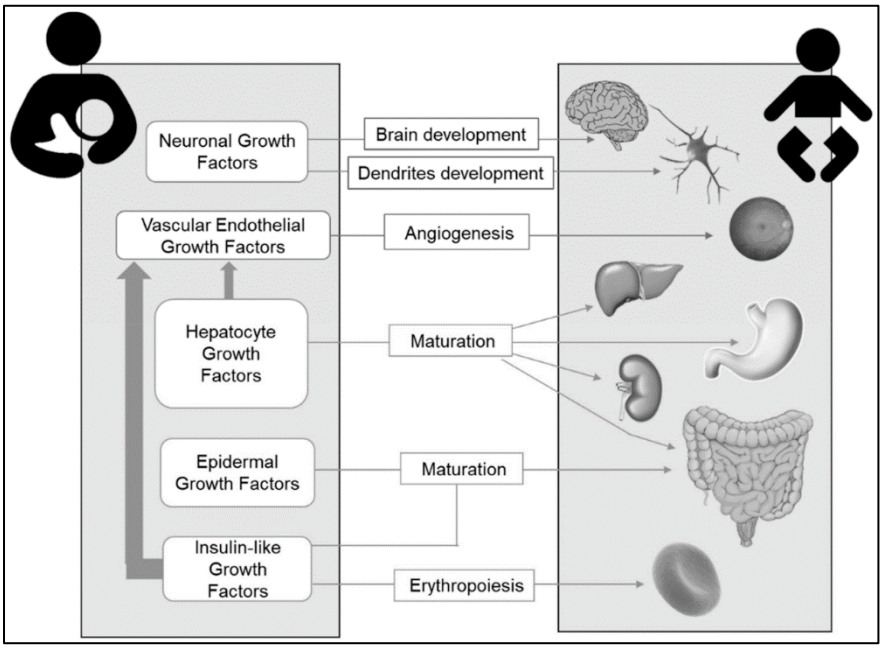


Figure 3. Growth factors transferred through human milk to support the maturation and development of infants' biological systems. Source: GILA-DIAZ et al 2019.

Interestingly, HM contains the highest number of bioactive compounds through the first postpartum days while its overall composition keeps changing during lactation (CHRISTIAN et al, 2021; SAMUEL et al, 2020; GILA-DIAZ et al 2019). Altogether, nutritional, and non-nutritional components in HM are interacting with one another, inducing temporal dynamics in their system. (CHRISTIAN et al, 2021).

2.4 HUMAN MILK COMPOSITION

HMC is considered one of the special research areas in which nutrition scientists should put more effort (BASSAGANYA-RIERA et al, 2021). HMC is changing with time and adapting to various maternal and infantile factors and this dynamic system has an important long-term influence on the health of individuals (CHRISTIAN et al, 2021; GARWOLIŃSKA et al 2018).

Even though HM is the gold standard for early infancy nutrition, sometimes it is not practical or viable for mothers to breastfeed their infants and, in those cases, infant formulae are recommended instead (KIM; YI, 2020; KOLETZKO et al 2019; MARTIN et al, 2016). Despite industrial efforts to mimic HMC, HM substitutes are focusing on macronutrients- or single bioactive component rather than providing a balanced biological system, as HM is. The complexity of HM raises significant challenges to create appropriate HM substitutes (KIM; YI, 2020; MARTIN et al, 2016).

In general, HM consists of mainly water and solid components including nutritional components such as 50 – 70 g of carbohydrates, 15 – 40 g of lipids and 8 – 16 g of proteins (KIM; YI, 2020). These nutritional components originate from maternal diet and body storages or synthesized in the breast cells (lactocytes) (BALLARD; MARROW, 2013). Besides, HM also carries bioactive components originating from various sources, such as mammary glands, HM cells and maternal serum (BALLARD; MARROW, 2013). (CHRISTIAN et al, 2021; SAMUEL et al 2020; BALLARD; MARROW, 2013) Together, nutritional

(macro- and micronutrients) and bioactive components (microbes, hormones, cells, antibodies, etc.) make HM the most valuable source of nutrition during infancy (**Figure 4**).

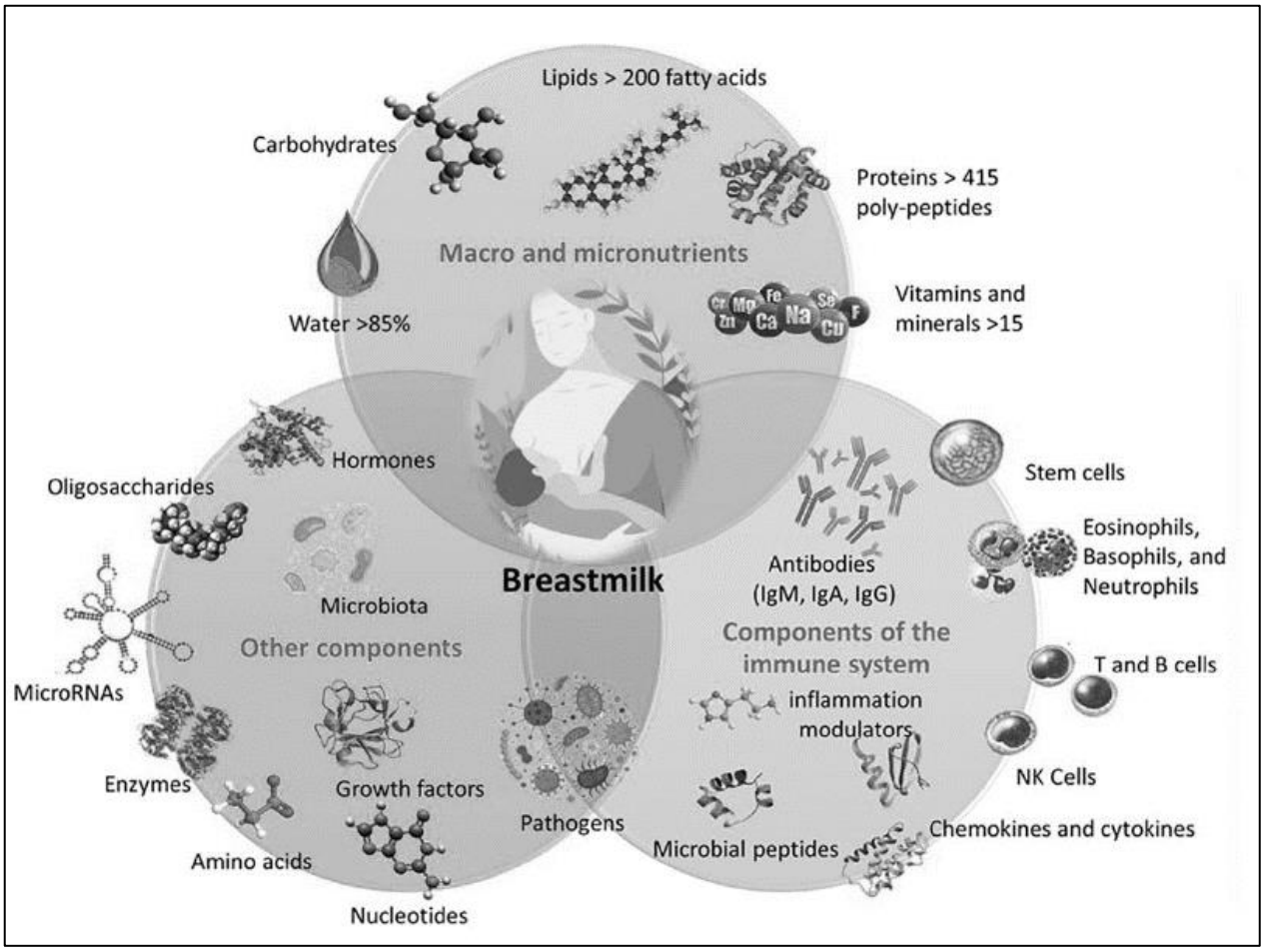


Figure 4. Components of human milk. Source: CABA-FLORES et al, 2022

Bioactive components can be defined as the ones impacting biological processes and having influence on body functions or state and, due course, on health (**Table 1**) (BALLARD; MARROW, 2013).

Table 1. Main bioactive components in human milk. Source: adapted from Ballard & Marrow (2013)

Bioactive component	**Biological impact**
Cells	
Macrophages	Defense against infection, T-cell activation
Stem cells	Regeneration and repair
Immunoglobulins	
Immunoglobulin A/secretory immunoglobulin A	Pathogen binding inhibition
Immunoglobulin G	Anti-microbial, activation of phagocytosis; anti-inflammatory, response to allergens
Immunoglobulin M	Agglutination, complement activation
Cytokines	
Interleukin-6	Stimulation of the acute phase response, B cell activation, pro-inflammatory
Interleukin-7	Increased thymic size and output
Interleukin-8	Recruitment of neutrophils, pro-inflammatory
Interleukin-10	Repressing Th1-type inflammation, induction of antibody production, facilitation of tolerance
Growth Factors	
Epidermal growth factor	Stimulation of cell proliferation and maturation
Heparin-binding epidermal growth factor-like growth factor	Protective against damage from hypoxia and ischemia
Vascular Endothelial growth factor	Promotion of angiogenesis and tissue repair
Nerve growth factor	Promotion of neuron growth and maturation
Insulin-like growth factor	Stimulation of growth and development, increased RBCs and hemoglobin
Erythropoietin	Erythropoiesis, intestinal development
Anti-microbial	
Lactoferrin	Acute phase protein, chelates iron, anti-bacterial, antioxidant
Lactadherin/Milk fat globule-epidermal growth factor 8 protein	Anti-viral, prevents inflammation by enhancing phagocytosis of apoptotic cells
Hormones	
Calcitonin	Development of enteric neurons
Somatostatin	Regulation of gastric epithelial growth
Adiponectin	Reduction of infant BMI and weight, anti- inflammatory
Leptin	Regulation of energy conversion and infant BMI, appetite regulation
Ghrelin	Regulation of energy conversion and infant BMI
Human Milk Oligosaccharides	Prebiotic, stimulate beneficial colonization, reduce inflammation

The influence of the HM on the brain functionality is widely cited as a model of synergy between different nutritional and bioactive components. Essentially, all of them (fats, vitamins, hormones, cells, immune factors, and oligosaccharides) affect the infant's cognitive functions (DE WEERTH et al, 2022). Other examples include memory, brain development, learning, etc. (**Figure 5)** (DE WEERTH et al, 2022).

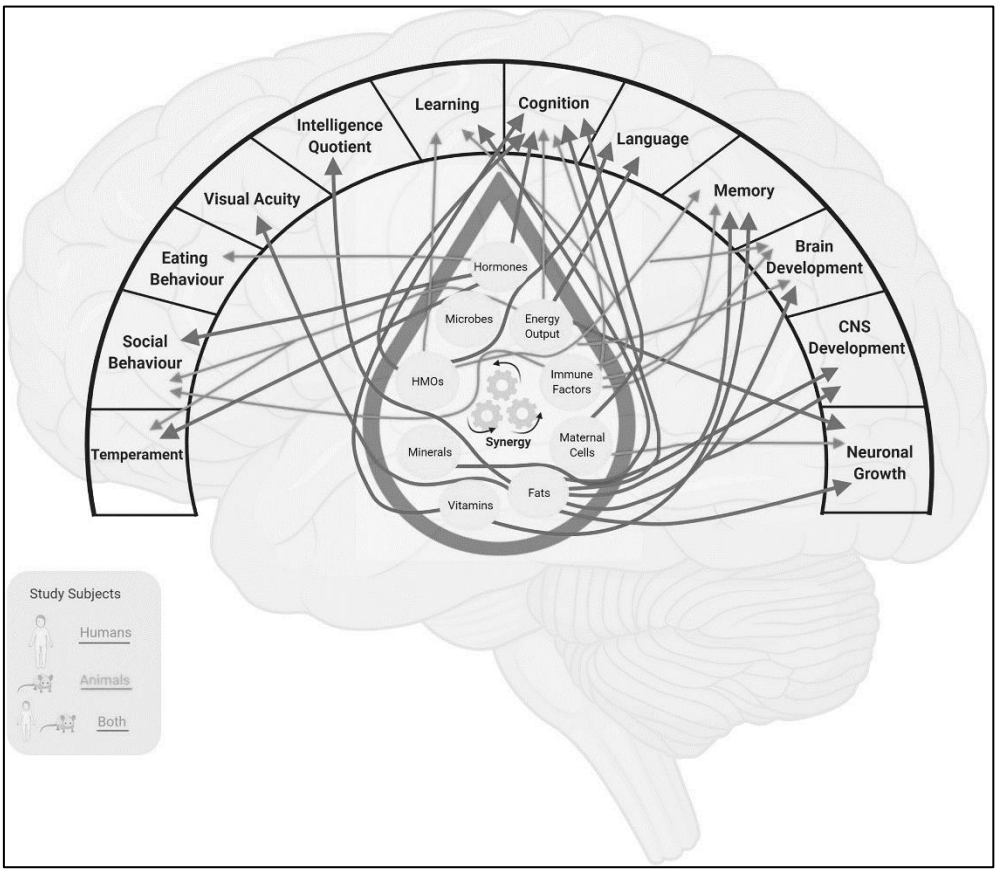


Figure 5. Human milk components and their impact on the brain development.

It is well studied that HM is a nutrition source full of interacting molecules that respond to other environmental factors, in order to optimally adapt and shape HMC over time (CABA-FLORES et al, 2022). There are longitudinal changes, occurring over the postpartum time, circadian and intrinsic variation, depending on various mother-infant characteristics (CABA-FLORES et al, 2022; SAMUEL et al 2020).

2.4.1 Longitudinal phases of human milk

HM is uniquely made for a single mother-infant pair, it responds to many different maternal and infantile factors such as maternal diet, mode of delivery, geography, gestational age, etc., which makes even more challenging to individualize infant formulae (CHRISTIAN et al, 2021; SAMUEL et al, 2020). HMC has three time-dependent stages: the colostrum (up to the 5^{th} postpartum day), the transitional (5^{th}- 10^{th} postpartum day) and the mature milk (from the 10^{th} postpartum day) (CHRISTIAN et al, 2021).

As a dynamic system, HM drastically changes during the first two weeks after birth, in order to adapt and provide optimal nutrition for the newborn (CHRISTIAN et al, 2021; SAMUEL et al 2020; BALLARD; MARROW, 2013). Longitudinal changes in one molecule can also affect the concentrations of other molecules, both nutritional and bioactive components (CHRISTIAN et al, 2021; SAMUEL et al 2020; BALLARD; MARROW, 2013). For instance, epidermal growth factors, insulin-like growth factors and most of the transforming growth factors reduce their concentration over lactation period, from colostrum to mature milk, while the transforming growth factor-alpha seems to increase during the same period of lactation (LU et al, 2018). Table 2 shows the compositional changes of some HMC concentrations that happen from the colostrum to the mature milk.

Table 2. Energy and nutritional composition of human milk.

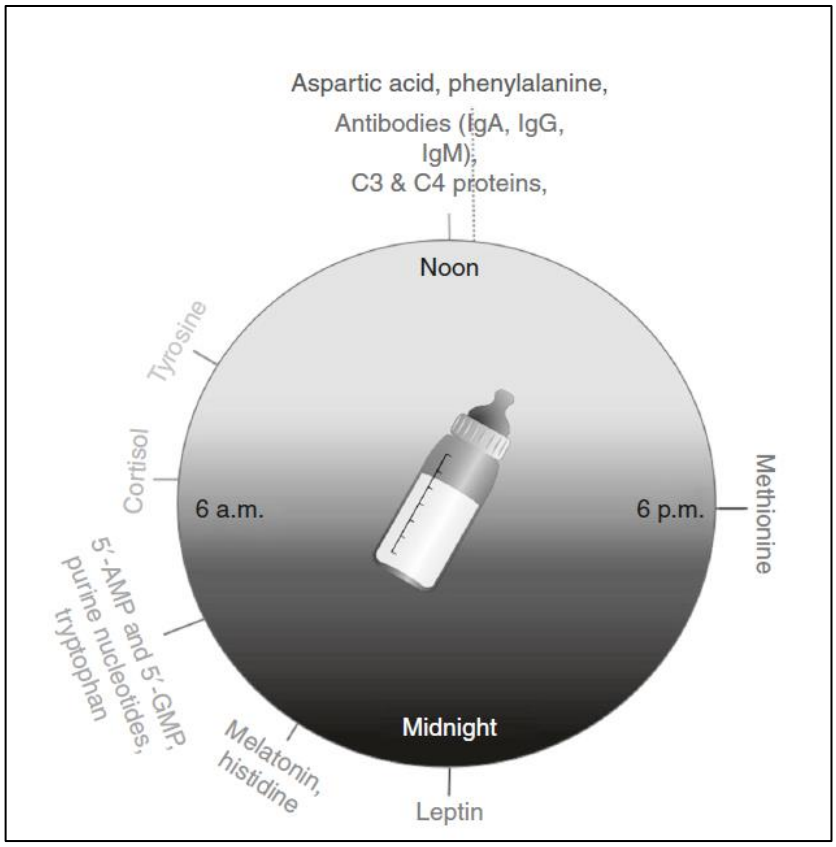
Component	Colostrum	Mature milk
Energy	50-60 kcal/100 mL	65–70 kcal/100 mL
Macronutrients		
Carbohydrate	50-62 g/L	60–70 g/L
Lactose	20–30 g/L	67–70 g/L
Oligosaccharides	20–24 g/L	12–14 g/L
Protein	14–16 g/L	8–10 g/L
Lipid	15–20 g/L	35–40 g/L
Micronutrients		
Iron	0.5–1.0 mg/L	0.3–0.7 mg/L
Calcium	250 mg/L	200–250 mg/L
Phosphorus	120–160 mg/L	120–140 mg/L
Magnesium	30–35 mg/L	30–35 mg/L
Sodium	300–400 mg/L	150–250 mg/L
Chloride	600–800 mg/L	400–450 mg/L
Potassium	600–700 mg/L	400–550 mg/L
Manganese	5–12 μg/L	3–4 μg/L
Iodine	40–50 μg/L	140–150 μg/L
Selenium	25–32 μg/L	10–25 μg/L
Copper	0.5–0.8 μg/L	0.1–0.3 μg/L
Zinc	5–12 μg/L	1–3 μg/L

Source: Adapted from KIM; YI, 2020

Colostrum differs from mature milk primary on having low fat content and high immunological components, such as human milk oligosaccharides and secretory immunoglobulins, whose role are more in immunology rather than a nutrition, this understandable as the infant has just left a sterile environment to be exposed to the external world (ANDREAS; LE-DOARE, 2015).

There is a rising field in nutrition research called “chrono-nutrition” which assesses the impact of mealtimes on health (FRANZAGO et al, 2023). When it comes to circadian variation, it is remarkable how HM components vary depending also on the time-of-day demand (CABA-FLORES et al, 2022; HAHN-HOLBROOK et al, 2019). These hour-based changes reassure the importance of the period of the day when samples are collected (**Figure 6**). (CABA-FLORES et al, 2022; HAHN-HOLBROOK et al, 2019).

Figure 6. Circadian variation of human milk composition. Source: HAHN-HOLBROOK et al, 2019



For instance, concentration of tryptophan varies in different lactation stages (colostrum, transitional and mature milk) and it gradually decreases towards mature milk. However, all the three milk stages present similar circadian pattern of this essential amino acid, according to exposure to sunlight (**figure 7**).

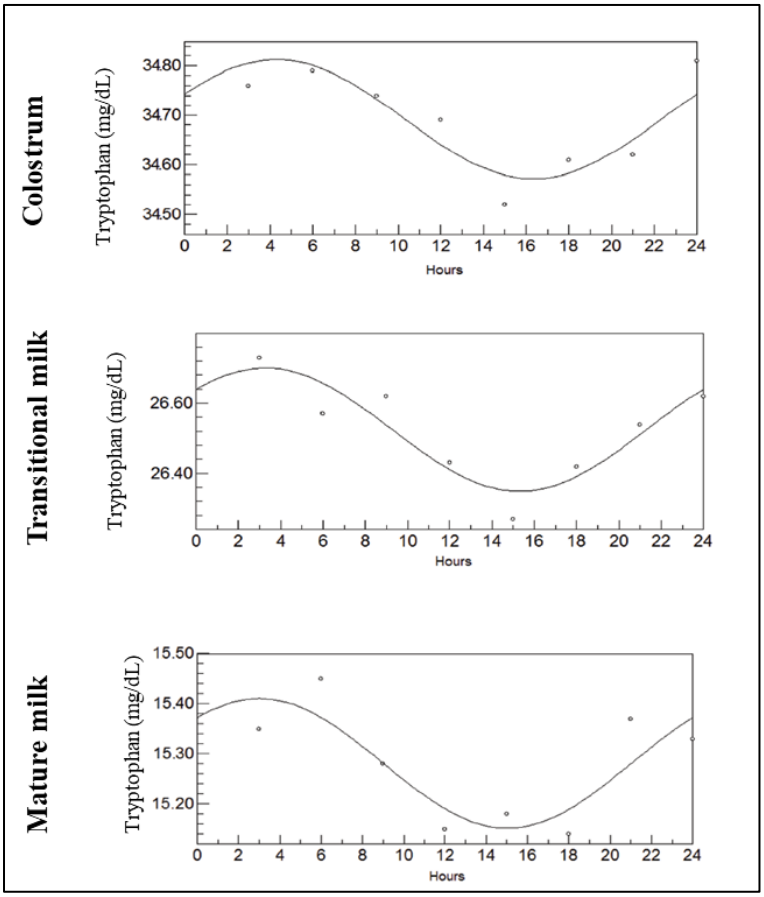


Figure 7. Circadian variation of tryptophan in human milk. Source: SANCHEZ et al, 2013

A plausible explanation for this is the fact that tryptophan is the precursor of serotonin (neurotransmitter) and melatonin (hormone), both responsible for sleep-stimulating activity in the human body (SANCHEZ et al, 2013). While varying drastically over the first post-partum month, HMC also changes periodically within a 24-h period, in order to provide optimal fulfilment for various circadian demands of the infant, such as tryptophan daily variation (CABA-FLORES et al, 2022; HAHN-HOLBROOK et al, 2019).

2.4.2 Variabilities in HM

By receiving sprints from different sources (e.g., environment), HM adapts accordingly, which makes HM a gold standard for personalized nutrition (CHRISTIAN et al, 2021; SAMUEL et al, 2020). Maternal, milk and infant factors are recognized as key elements responsible to shape HMC uniquely,

therefore they are described as the mother-milk-infant “triad” by Bode et al (2020) (CHRISTIAN et al, 2021; SAMUEL et al, 2020).

As a maternal condition, nutritional status of mothers does not have much impact on the macronutrients found in HM (BALLARD; MARROW, 2013), though HM of mothers with higher BMI shows higher level of saturated fatty acids, omega-6/omega-3 ratio (MAKELA et al 2013), and leptin (KUGANANTHAN et al 2017). Even so such studies on BMI were performed in high income countries, it is highly recommended that lactating mothers should keep a balanced dietary intake including adequate nutrients concentrations (KOLETZKO et al 2019).

As a function of maternal diet, HMC can considerably change at nutrient level (SAMUEL et al, 2020; KEIKHA et al, 2017). The intake of amino acids, omega-3 fatty acid, some minerals (iodine, selenium) and vitamins (thiamine, riboflavin, niacin, B6, B12, choline, C, A, D and K) can influence HMC, while changes in other minerals intake such as calcium, magnesium and copper may not have impact (SAMUEL et al, 2020; KEIKHA et al, 2017). Likewise, fatty acids from maternal intake are well-known to interfere significantly with HMC (DEMMELMAIR, KOLETZKO, 2018; VOORTMAN et al, 2016; INNIS, 2014), from which the polyunsaturated fatty acids play central role (BARRERA et al, 2018; AMARAL et al 2017). A study showed that the habit to eat salmon (2 portions/week), from the 20th gestational week until delivery time, provided higher levels of omega 3 fatty acids (eicosapentaenoic acid, docosahexaenoic acid and docosapentaenoic acid) in HM on the 5th postpartum day (URWIN et al 2012). Other factors related to mothers such as mode of delivery, age of the mother, health status and parity may alter HMC, but evidence are lacking or inconsistent (SAMUEL et al 2020).

As an infant condition, the gestational age is key for changing HMC. The lower gestational age is associated with decreased lactose concentration in colostrum and increased protein levels (BORRÀS-NOVEL et al 2023; GIDREWICZ FENTON, 2014). In addition, male infants may receive more energy (THAKKAR et al 2013; POWE et al, 2010) and lipid (THAKKAR et al 2013) in HM than female infants do..

It is important to highlight that the way HM samples are expressed, stored, and measured also affect the final reported values of HM, which represent limitations, in terms of accuracy, in all studies on HMC (SAMUEL et al, 2020). HM expressed manually or by big electric pumps shows higher levels of protein when compared to regular pumps while the manual expression alone provides higher amounts of sodium and lower potassium levels in HM when compared to pumps (BECKER et al, 2015).

The analytical method applied to the measurement of HMC, may also contribute to the heterogeneity of the data recorded from the literature. Lipids are the most sensitive macronutrients when it comes to storage (SAMUEL et al, 2020). There is evidence that even when stored at -20ºC for only 2 days, total lipid reduces ~9% in HM samples (CHANG et al, 2012). An explanation for this may be that bile salt-stimulated lipases can remain active even in cooler temperatures such -20ºC (LEV et al 2014).

Similarly, analytical methods on HMC should vary according to the nutrients of interest (SAMUEL et 2020). The bomb calorimeter accurately measures macronutrients and energy in HM (GIDREWICZ, (year); FENTON, 2014), while the Kjeldahl method measures the total protein content in HM and the Folch, Bligh and Dyer and Rose-Gottlieb methods are preferable to produce precise results for the lipid content in HM (BOYCE et al 2016). After a general literature review, Samuel et al (2020) proposed simple recommendations to minimize errors on the quantification of HM components (**Table 3**)

Table 3. Recommendations for quantifying components in human milk. Source: adapted from Samuel et 2020.

Expression	Storage	Analytical analysis
✓ Set a standardized time to express milk samples; ✓ Target only one breast to collect milk samples (right or left); ✓ Instruct mothers to empty the targeted breast ~2 hours before samples collection; ✓ Collect full sample (complete breastfeeding session: from fore- to hindmilk); ✓ Write down the volume expressed; ✓ Pick only one expression method for the whole sample expression (manual or electronic).	✓ If various analyses are needed, store small aliquots of milk samples to avoid freeze-thaw cycles; ✓ Preferably, use freezer -80ºC to store the aliquots; ✓ Do not waste time on the samples transportation (maintain low temperatures)	✓Choose the appropriate analytical method in accordance with the targeted nutrient.

As mentioned before, HMC is well-known for being affected by many conditions related to mother, infant and milk samples, and that is what makes research of HM so complex (**Figure 8**). HM is a biological system that cannot be studied simply as the sum of its components, rather as the interactions of nutritional and non-nutritional elements and their variation as a function of time and other conditions, i.e. the factors of the mother-milk-infant triad.

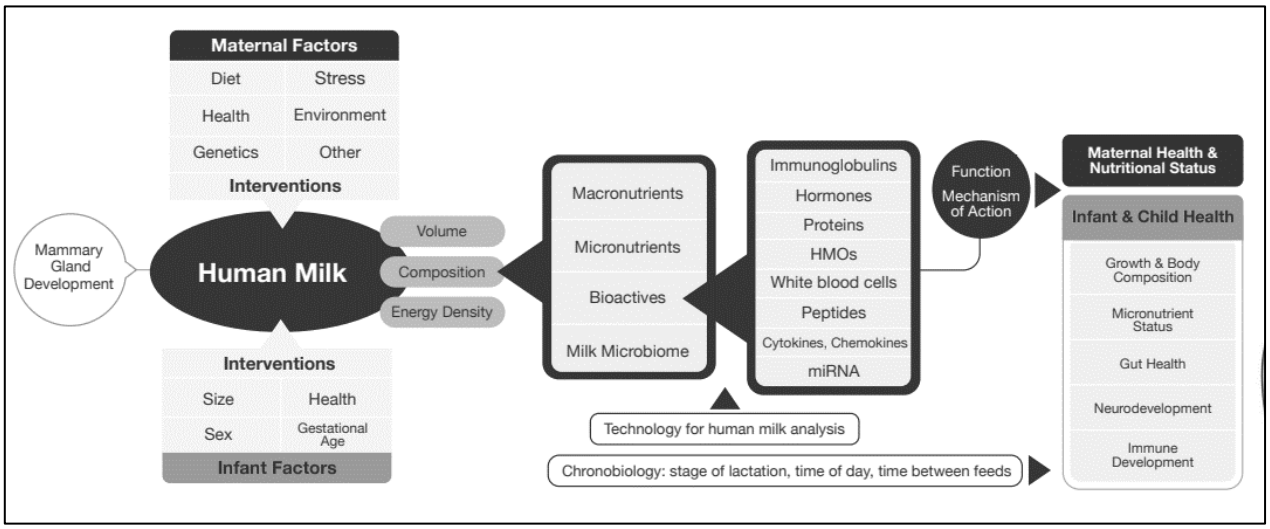


Figure 8. Complexity involving human milk composition. Source: NIH, 2023b

2.4.3 Gaps in HMC research

HM research has identified major gaps in our knowledge about HMC (SHENHAV, AZAD, 2022; AHUJA et al 2022; CHRISTIAN et al 2021). These gaps raised key questions, such as: (i) how maternal and infant factors influence HM dynamics, (ii) how to tailor HMC, in order to positively influence infant health and development and, in circumstances when breastfeeding is not an option, (iii) how to personalize infant formulae according to the specific aspects of mother-infant pairs (CHRISTIAN et al 2021).

Consequently, Shenhav and Azad (2022) recognised the three main facts on the HMC research that are considered as barriers to advance the current knowledge: (i) most studies focus on the impact of a single aspect/factor in the mother-milk-infant triad on individual components of HM, (ii) the so-needed multi-omics methods are used in cross-sectional studies rather than longitudinal ones and (iii) there is a

notable deficiency in the knowledge on advanced computational and statistical tools to properly analyse data on the field.

It is well-known that, by moving from a simplistic analysis towards a more complex one, new challenges have emerged on HMC research (SHENHAV, AZAD, 2022; AHUJA et al 2022; DE WEERTH et al 2022; CHRISTIAN et al 2021; SAMUEL et al, 2020). Insufficient effort is invested in combining, visualizing and analyzing distinct types of data on HMC in this present era of data science. That is why advanced computational skills are considered key in this shift of HMC research (SHENHAV, AZAD, 2022; AHUJA et al 2022; DE WEERTH et al 2022; CHRISTIAN et al 2021; SAMUEL et al, 2020). Multidisciplinary teams along with interdisciplinary collaborations are the main elements to evolve research on the HMC, especially when sophisticated data science tools include multi-omics methods (DE WEERTH et al 2022).

For instance, the "Breastmilk Ecology: Genesis of Infant Nutrition" project (BEGIN initiative) developed by the Eunice Kennedy Shriver National Institute of Child Health and Human Development, comprises a multidisciplinary team to address many questions related to the mentioned knowledge gap in HM (NIH, 2023a). The initiative has five working groups out of which the "human milk composition working group" is focused to analyse HMC via systems biology approaches, including ideation for large and diverse longitudinal studies combined with multi-omics and advanced computational techniques such as network analysis and predictive models (USAID , 2023).

While state-of-the-art initiatives such as the BEGIN project are ongoing, and gradually providing insights on the dynamics of HMC, there is a vast number of published data already available in the literature on this topic (DE WEERTH et al 2022; CHRISTIAN et al 2021). Most published data currently focus on understanding individual components, while the research priority shifts to a more holistic approach in order to see HM as a complex system (DE WEERTH et al 2022; SHENHAV, AZAD, 2022; CHRISTIAN et al 2021; GEDDES et al 2021).

Data originated from published studies have the potential to be pooled together in one common framework with advanced computational tools. Such need was recently raised by Ahuja et al. (2022) while advocating for a common framework on recording HM components in North America, mainly by taking advantage of data collected from ongoing longitudinal cohorts in United States and Canada (**Figure 9**).

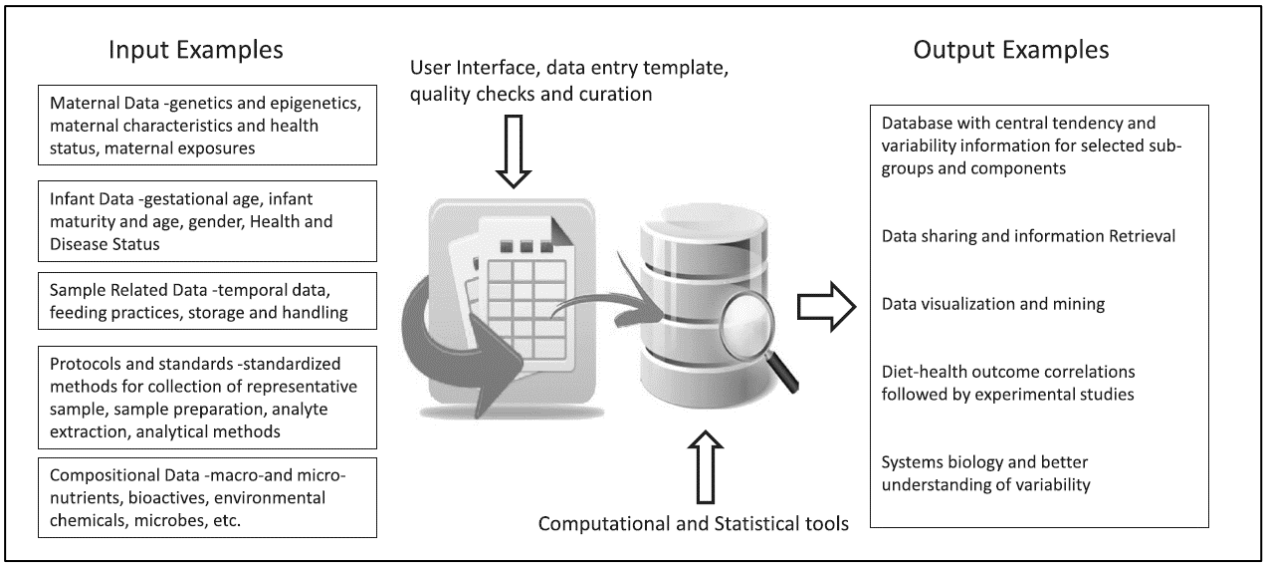


Figure 9. Scheme of a hypothetic database on HMC. Source: Ahuja et al 2022.

Beside storing data on HMC, the ideal database needs to accommodate temporal data (components changing over time) and information on maternal, infant and methodological factors that are known to affect HMC (AHUJA et al 2022).

Once such framework exists, HMC data can be stored and eventually analysed as a biological system (AHUJA et al 2022; DE WEERTH et al 2022; SHENHAV, AZAD, 2022; CHRISTIAN et al 2021). Also, the research field will be able to evolve towards a more accurate science- and data-based recommendations to pregnant/lactating women along with the infant formulae industry, in order to optimize recipes for HM substitutes (AHUJA et al 2022; DE WEERTH et al 2022; SHENHAV, AZAD, 2022; CHRISTIAN et al 2021).

Shenhav and Azad (2022) stated that the microbiome research should serve as guidance to advance HMC research towards a system approach. HMC has much in common with microbial community ecology, and computational microbiology tools would be useful to explore this (**Figure 10**).

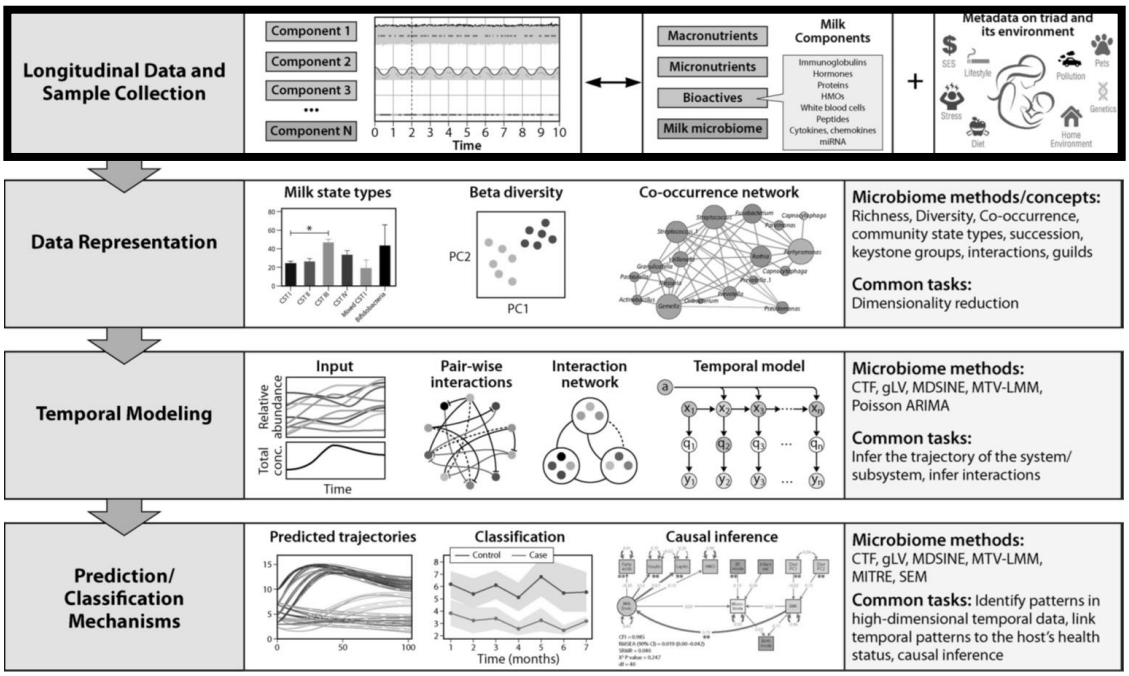


Figure 10. Suggested methodology to study human milk composition. Source: SHENHAV, AZAD, 2022

The first step towards a new era for HMC research consists of capturing longitudinal data including information on the mother-milk-infant triad (SHENHAV, AZAD, 2022; AHUJA et al 2022). Therefore, data on the time of collection (various timepoints) and concentration of components (nutritional and bioactive components) are fundamental to clarify such reported HMC.

The context of HM collection should be stored as well; that is, the characteristics of the infants and mothers (gestational age, mothers' diet, BMI etc) and information of the used HM analysis (expression, storage, analytics) (SHENHAV, AZAD, 2022; AHUJA et al 2022). Current Food Composition Databases (FCD-s) are not able to store such complexity of dynamic values of foodstuff in conjunction with influencing factors at once (FERRAZ DE ARRUDA et al 2023; MICHA, 2018).

2.5 FOOD COMPOSITION DATABASES

FCD-s are platforms that hold data on nutrients and other relevant components of various food items (DELGADO et al, 2021). Food industry stakeholders, scholars, healthcare professionals, public policymakers (food and health), consumers and educators utilize FCD-s for collective and individual usage (DELGADO et al, 2021; KAPSOKEFALOU et al, 2019) (**Figure 11**).

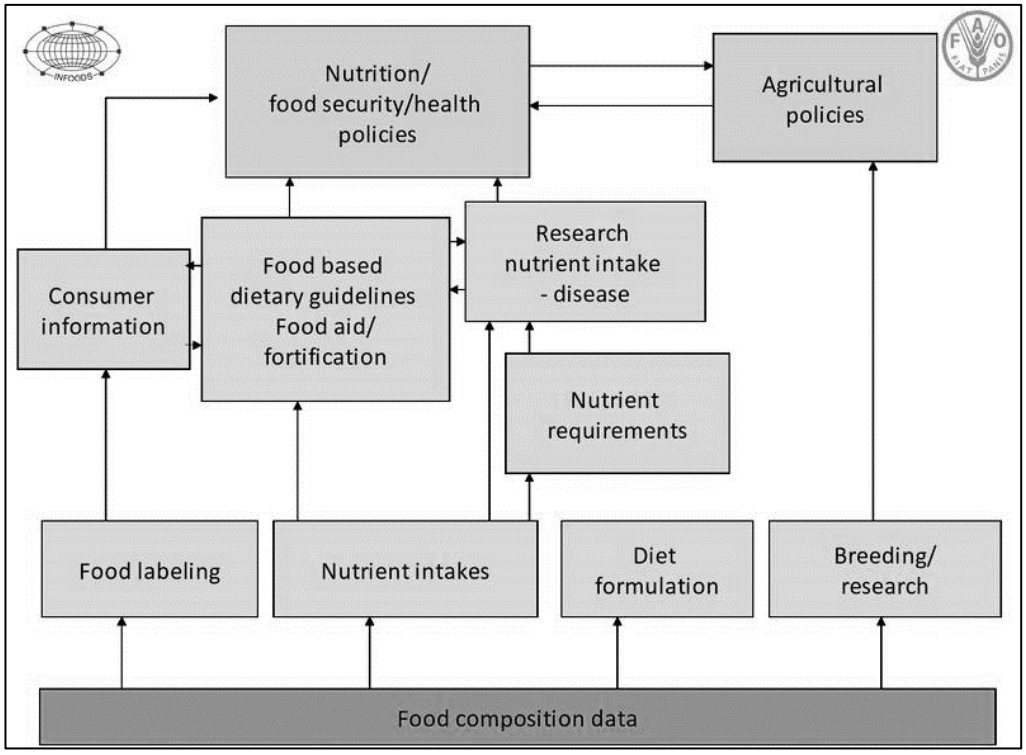


Figure 11. Various purposes for utilizing food composition databases. Source: INFOODS/FAO, 2023a

Conventionally, FCD-s are sources of nutritional values to convert food intake records into actual nutrients intake (DELGADO et al, 2021; KAPSOKEFALOU et al, 2019). Consequently, FCD-s are valuable tools for a list of functions such as: to assess health and nutritional status, to formulate appropriate individual or collective diets (schools, prisons, hospitals, etc), to support nutritional education, to promote health on nutrition, to train professionals on food composition and nutrition, to perform researches on dietary intake and diseases, to create nutrition labelling, to develop food products, to evaluate dietary patterns, and to monitor national policies on nutrition and health (CHURCH, 2006).

According to Church (2006), the first reported FCD was dated in 1818 as a “nutrition scale” to manage food supply in prisons, still the first structured FCD as we know today was issued at the end of the 19th century (CHURCH, 2006). Germany was the pioneer in Europe by publishing a contemporary-shaped European FCD in 1878 (KONIG,1878). In the United States, Atwater and Woods (1896) initiated a FCD by collating data from 2600 analyses of several popular American food items, varying from meats, cereals, fruits, vegetables to chocolates and other processed foods. However, it was only in 1949 when the Food and Agriculture Organization (FAO) published the first global FCD to be used internationally (CHATFIELD, 1949).

Though, generally, FCD-s concentrate on nutritional components (macro- and micronutrients). Though, bioactive components such as the polyphenols in plant-based foods, have also been gaining space recently (DELGADO et al 2021; KAPSOKEFALOU et al, 2019). The FoodData Central from the United States Department of Agriculture (USDA, 2023), the International Network of Food Data Systems, Food and Agriculture Organization (INFOODS) created by FAO (INFOODS/FAO, 2023b), and the European Food Information Resource Association Internationale *Sans but Lucratif* (EuroFIR, 2023) are examples of well-known FCD-s (FERRAZ DE ARRUDA et al 2023; DELGADO et al 2021; KAPSOKEFALOU et al, 2019) (**Table 4).**

Table 4. Main Food Composition Databases worldwide. Source: Adapted from Delgado et al. 2021

Name	Organization	Description	Source of food data	Updates
FoodData Central	United States Department of Agriculture	Targets relevant components for each food; highly differentiated	Laboratory analysis	Regularly
EuroFIR	EuroFIR AISBL, International non-profit association	Shows that on energy, macronutrients, vitamins, minerals, and other bioactive compounds	Estimations by expert panels and targets food and nutrition professionals	Regularly
INFOODS	Food and Agriculture Organization	Assembles food composition compilers, data generators and users (e.g., chemists, nutritionist, food scientists) and decision makers	Retrieved analytics from published data	Regularly

Even though FCD-s have evolved over the years, there are important limitations of data quality in relation to incompleteness and obsoleteness (outdated data) (AHUJA et al, 2022; DELGADO et al 2021; KAPSOKEFALOU et al, 2019). For instance, FoodData Central (USDA, 2023) provides data on HM from the 70s without details on the source, sampling, storage, or analysis of HMC (AHUJA et al, 2022).

Another key obstacle of FCD-s is their limitation to capture temporal variation of food items (FERRAZ DE ARRUDA et al 2023; KAPSOKEFALOU et al, 2019), and the challenge is even bigger when the issue expands to understanding to causal interactions with biological data as in HMC (TOURE et al, 2020). During the temporal changes of food components (trajectories) the external factors (soil, climate, harvest etc) may also change (FERRAZ DE ARRUDA et al 2023; KAPSOKEFALOU et al, 2019), therefore the condition-response effect should be considered more like an x(t)=>y(t) assignment with temporal variables rather than just and x=>y mapping between values is analogous to the impact that mother and infant factors have in the HMC.

2.6 OBJECTIVE

Our primary objective has been, to build a novel food composition database on HMC (nutritional and bioactive components), considering two elements of complexity: time-dependence and variability. The secondary objective is , to demonstrate the applicability of such database via its visual tools to find pattern in the total protein content of HM.

3. METHODOLOGY

FCD-s carry a large amount of data mostly focused on the composition of food as a static piece of data, neglecting the possible temporal variation of that composition or detailed information under what condition the data were acquired. Consequently, precious information is lost during populating the database (DE ARRUDA et al, 2023). MilkyBase was created to address this issue on HMC.

MilkyBase intended to capture the dynamics of HMC along with relevant information on the mother-milk-infant triad producing the data. Our vision is that MilkyBase could be a piece in the chain that could lead to a Big Data approach to advance HM research. Big Data can be defined by several V-s: Volume, Velocity, Veracity Variety, Variability and Value (CREMIN et al 2022; RISTEVSKI, CHEN, 2018). “Volume” indicates the big amount of generated data; “Velocity” refers to the fast access and processing of the data; “Veracity” implies that the generated data are verifiable, reliable and consistent. “Variety and Variability” denote the diversity of data due to its complexity and heterogeneity, and lastly, “Value” refers to the benefit provided by comprehensible data analysis using (RISTEVSKI, CHEN, 2018).

To make MilkyBase easily accessible for nutrition- and food-scientists, the database was created in Microsoft Excel, a popular software used by scientists of all fields, with user-friendly facets (NUNES et al, 2015 Additionally, Microsoft Excel is accompanied by the Visual Basic for Application (VBA) programming language that was utilized to support the functionality of MilkyBase. Macros written in VBA were used to validate the syntax and semantics of the database. It made users aware of errors in the data while they were adding them to MilkyBase.

3.1 VOLUME

Firstly, to cover the volume aspect, a targeted search of relevant literature was performed to find large amount of quantitative data on HMC. The search focused on PubMed with the following MeSH terms and Boolean operators: (“human milk” OR “breast milk” OR “mothers’ milk”) AND (“nutrients” OR

"components" OR "composition" OR "biochemical" OR "quantification" OR "bioactive"). The search prioritized (but was not limited to) English language literature.

In parallel, FoodMine (HOOTON et al 2020) was used to systematically evaluate title and abstract of published studies related to HMC in PubMed. As the volume, the goal was to get enough quantitative and longitudinal data on HMC to start building a prototype framework to comport them all and, progressively, while more data were coming in, more adjustments were made until the final MilkyBase template was created.

3.2 VELOCITY

The principle of velocity was not stressed in MilkyBase mainly because the current volume of data is modest for a Big Data philosophy and does not affect the functionality of the database. As the number of records will reach tens of thousands, this will be an issue. Then, Microsoft Excel will be the transit area to transfer MilkyBase into a system that would comply with Big Data criteria.

3.3 VARIETY AND VARIABILITY

After literature search and before inputting data in MilkyBase, two key questions regarding the source of knowledge (scientific papers) needed to be asked, (i) Does it provide sufficiently large set of quantitative and longitudinal data on HMC? and (ii) What data are important and should be added? If the scientific paper provides numerous temporal datapoints in addition to information on the conditions under which the data were measured, then such paper was prioritized to be inputted in MilkyBase.

Later, following the variety principle, a record in MilkyBase can be divided into two types of fields: quantitative data on the HMC (response fields) and the information on the conditions (named as explanatory fields) to which the responses were measured. The latter ones could be related to the mother (BMI, diet, age, etc.), infant (gestational age, sex, weight, etc.) and milk sample (expression, pasteurization, storage temperature, etc.).

In terms of veracity, the original data were extracted from refereed scientific papers and their verification was not an issue.

Variety was addressed in many ways. The default unit used for HMC was set as "gram *per* Liter" of milk, therefore, conversion had to be performed on the original data whenever other unit was used. Similarly, if values were given in mass of component per mass of HM, they were converted to g/L where 1 kilogram of milk was assumed to be 1 liter (USDA, 1992). However, an entry can be a derived quantity, too, like the ratio between two HMC. Such is the proportion of a (group of) molecule(s) compared to a bigger group, like omega 3/Polyunsaturated Fatty Acid.

Variety meant more effort to validate the syntax of the data. As described before, checking the syntax check is vital to maintain the structural integrity of the database, while the semantic check makes sure that the data are interpreted correctly. While the first one can be automated, the latter one must be mainly a human-based work, necessitating nutritional expertise, which affects directly what data are important to be collected.

Besides, MilkyBase allows estimations on HMC even when quantitative data of a certain component is missing. That was facilitated by organizing all the data (response and explanatory fields) in a hierarchical tree-structured order. For instance, in the response fields, the root represents the total HMC that is followed by many nodes - the groups of components - until it reaches the leaves, representing the component itself at molecular level. So, if the information of a certain component is needed (leaf), but only data on its former group (node) was available, estimations (at certain confidence level) could be performed to find an expected value for that component.

As a practical example, **Figure 12** shows the data hierarchy of a component in MilkyBase. The root represents the overall HMC while the fatty acid itself (C22:6 n-3) is the leaf and the four groups of components between them are the nodes (Lipids, Fatty Acids, Omega 3), where the fatty acids belong to.

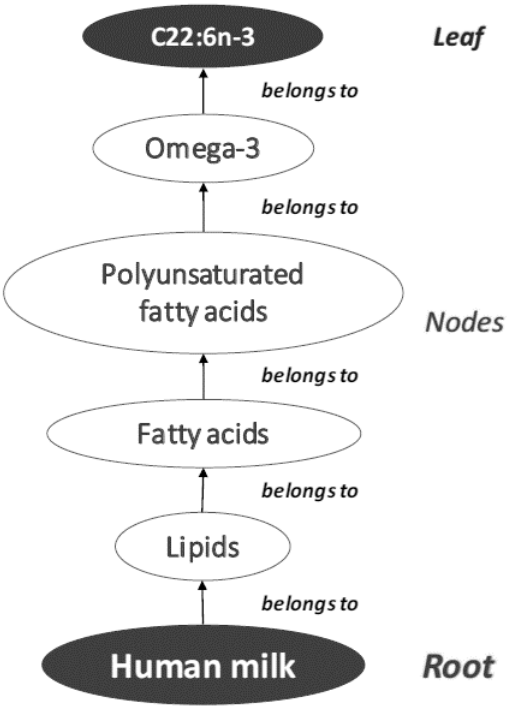


Figure 12. Example of tree-structured data of a human milk component in MilkyBase.

In an alternative presentation, but displaying the same tree-structured concept, **Figure 13** shows an example coming from data related to maternal supplementation where vitamin D supplement is located in the inner part of the onion chart (leaf) which belongs to "Dietary supplement of Mother" group (node) that is part of the "Diet of Mother" group (node), until reaches the outer group (root) "Condition of Mother".

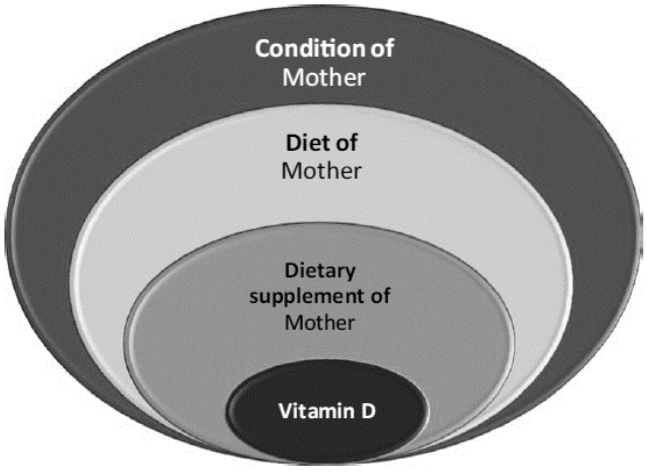


Figure 13. Example of tree-structured data of maternal supplementation in MilkyBase.

MilkyBase contains information on the uncertainties of HMC measurements. Typically, a numerical entry has two parts, where the first part represents the arithmetical average of measured data (including a single measurement), while the second part characterizes its uncertainty or spread. Alternatively the second part can be an interval, such as the 95% confidence interval. This format requires a minimum of one single value, then the remaining info is optional. Therefore, if the entry represents an interval, then a (possibly stochastic) interval-analysis can be applied in succeeding calculations.

In summary, as default format, a centroid value on the component concentration is recorded (target data), along with the quantification of its spread /or uncertainty (standard deviation or minimax boundaries). This information may be accompanied by estimations/predictions, supplied by their uncertainties, commonly as the standard error of the mean, or its 95% confidence interval.

To capture HMC temporal variability, MilkyBase holds an independent location for temporal measurements of HMC (dynamic data), that is, in a separate sheet, time-dependent variables are represented by a table of time-HMC pairs. This feature in MilkyBase allows the plotting and use of dynamic data and analyzing their parameters such as rates and their dependence on the conditions of the data generation. This template focusing on dynamic HMC data was inspired by the primary and secondary model method in the field of predictive microbiology following mathematical modelling principles when storing data (BARANYI; TAMPLIN, 2004). Namely, in predictive microbiology, primary model describes the temporal profile of a variable under constant conditions, characterized by a few (most importantly rate-) parameters. The variation of the primary parameters is described by secondary models, as a function of the explanatory variables characterizing the conditions under which the primary temporal profiles were produced. The way HMC data were structured in MilkyBase follows this scheme.

3.4 VALUE

MilkyBase brings three novel and valuable features in the FCD field, (i) the target is temporal data (both in response and explanatory fields); (ii) the quantification of the uncertainties and (iii) the tree structured organization of both the explanatory and response variables, allowing the users to execute probabilistic estimations equivalent to interval arithmetic.

3.5 MILKYBASE STRUCTURE

Ultimately, MilkyBase is a flat database, which is a system of linked sheets in a single Microsoft Excel workbook. The core of MilkyBase is named Master sheet that contains records (Excel rows) identified by unique keys. The fields (columns) can be divided into three groups: (i) administrative fields with details on how the data were collected, (ii) the explanatory fields related to the conditions under which the data on mother-milk-infant factors were produced and (iii) the response field that holds the measured/reported data on one or more HMC (**Figure 14)**

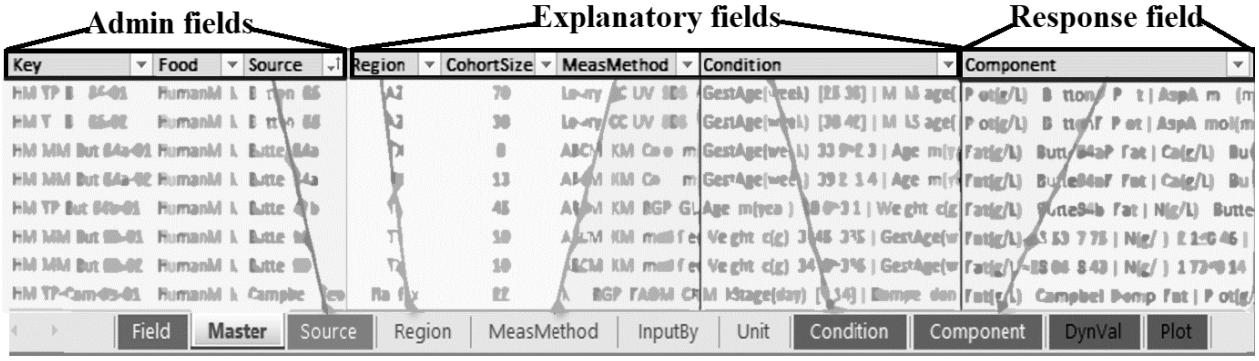


Figure 14. Screenshot of MilkyBase.

The "Region", "MeasMethod" (measurement methods), "Condition" and "Component" sheets are definition sheets with tree-structured data. Complementary to **Figure 14, Table 5** explains in detail the description of each field in MilkyBase and their respective data structure.

Table 5. Description of MilkyBase ontology

Field	Description	Typical structure and example
Key	A unique identifier for each record in the **Master** sheet.	**AA-BB-Ccc-xx-yy**, where: **- AA** – Code for the Food ("HM" for human milk) **- BB** - Initials of the person who inputted the record (e.g., AS for Anne Smith) - **Ccc** - First three letters of the surname of the first author of the scientific paper (e.g., Sil for Silva) - **xx** - Year of the publication (e.g., 07 represents 2007) - **yy** – An ID number within the source (e.g., 01, 02...) For instance, the Key could be inputted as: HM-AS-Sil-07-01, HM-AS-Smi-07-02, etc
Food	Name or category of the food item.	MilkyBase only accepts "HumanMilk"
Source	The source of the information stored by the record. It is an abbreviation defined in the **Source** definition sheet.	Example: Silva_07 (First Author surname + last 2 digits of year of publication)
Region	Geographic region of the cohort. The (tree-structured) value set is defined in the **Region** definition sheet.	Example: USA
CohortSize	Size of the cohort of mothers measured. When the data is longitudinal, this is the average of the sample size, from the first sample collection to the last measurement.	Example: 100
MeasMethod	Information on analytical methods used to measure human milk components. The (tree-structured) value set is defined in the **MeasMethod** definition sheet.	The entries are separated by semi-colon (;). Example: Calorimetry ; microKjeldahl
Condition	Various factors in relation to mother-milk-infant factors. The (tree-structured) value set is defined in the **Condition** definition sheet.	The entries are separated by vertical bars (|) and the unit are defined into parathesis (unit) Example: **GestAge(week)=**39±0.9 | **StorageTemp(C)**=-70 | **MilkStage(day)=**[3,56]

Table 5 Cont. Description of MilkyBase ontology

Field	Description	Typical structure, with example
Component	Concentration of human milk components. The (tree-structured) value set is defined in the **Component** definition sheet.	Entries are quantified (g / Liter milk by default) Biochemical components separated by vertical bars (e.g.,vitC); or groups of compounds (e.g, vitamins). Example: VitC(g/L)=0.0217±0.01929 | C18:1n-9/Fat(-)=0.3497±0.0521 If the component has dynamic data, a "!" should be entered for the value, followed by a unique ID (usually first author surname + component) to link the entry in the Master sheet to the DynVal sheet that holds all longitudinal data. Example: Fat(g/L)=17.4±0.5777 | Prot(g/L)=15.8±0.15 | BetaCarotene(g/L)=2.58E-5±1.58E-6 | VitB1(g/L)=2.20E-4±4.00E-5

3.5.1 Numeric value

Numeric values in MilkyBase can be integer (e.g., 6), fixed point format (e.g., 6.12), exponential format (6.12e5) and interval (e.g., [5.12e5, 9.9e6]. The Hashtag (#) is a special character, indicating a placeholder for an unknown number. Thus, [6, #] indicates a number greater than 6. In the Condition and Component sheets, the values can be combined with error margins like 6±2 or 6@[4,8]. Respectively, 6±2 or 6@[4,8] mean that either (i) the real value is between 4 and 8; or (ii) 6 is the centre with [4,8] as quantiles; or (iii) if 6 is an estimation then its 95% confidence interval is defined as [4,8]. Scientific notation is used when a number is less than 0.001.

In the Component and Condition sheets, both singular and dual numerical formats are permitted. The dual format comprises of two parts: the first refers to raw (measured) data, possibly with its scatter (standard deviation or inter-quantile range or minmax interval); the second part describes an estimation and the confidence in this estimation: either 95% confidence interval or standard error of the mean). So, the first part is descriptive, and the second part is predictive (or an estimate).

For instance, as typical numeric entry X±Y; X@[x,y] can be described as: X, a centroid of the raw data (average, median, etc.). X in the second part (after the semicolon) is about its estimation, generally the observed mean. Y is the Standard Deviation of the measured data, but if it was presented in the second part,

then it would represent the standard error of the mean. The interval represented by [x, y] represents the estimated value of interval, such as Confidence Interval (normally 95% CI), but if it was placed in the first part (before semicolon), then it could also mean the raw minmax values or interquartile etc.

4. RESULTS

The workflow of MilkyBase was set in five steps: (i) manual and automated literature search, (ii) selection of papers that gave quantitative and longitudinal data on HMC. Once such papers were selected, (iii) relevant information on the mother-milk-infant triad and on temporal data of HMC were spotted and, finally, (iv) the chosen data were inputted in their corrected location as explanatory or response fields (**Figure 15**).

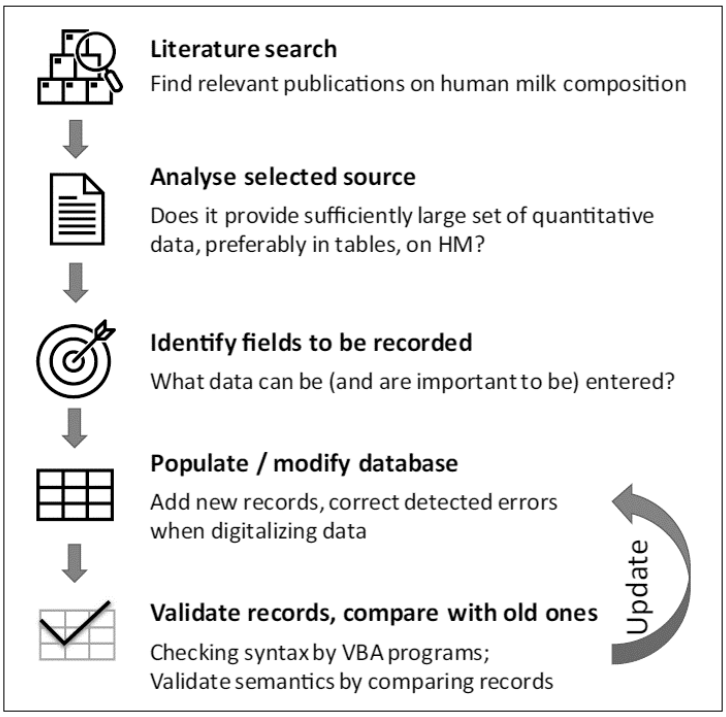


Figure 15. MilkyBase workflow.

By the time MilkyBase was in the process of being populated, the database structure underwent major modifications to be able to capture new information. It took around 150 records in the Master sheet to achieve a consistent structure. Currently, MilkyBase holds roughly 10,000 datapoints and around 600 components of HM (nutritional and bioactive).

MilkyBase can be downloaded and opened in any computer with Excel installed. A Fig Share repository holds the links to download MilkyBase, its user-manual and macros (see PACZA et al, 2022).

4.1 FINAL FRAMEWORK

The heart of MilkyBase is the Master sheet. It has eight fields to be filled with data related to HMC (**Figure 16**). The fields carry either numerical values (possibly with data on uncertainty as defined above) or list of allowed category values. The numerical domains are intervals, the allowed category values, as an interpretation domain, are given in a separate sheet with a name that is the same as that of the field in question in the Master sheet. Such sheets are called "Definition sheets", for obvious reasons. The interpretation intervals and the Definition sheets are the main resources for the syntax check programs accompanying MilkyBase.

Key	Food	Source	Region	Cohor	MeasMethod	Condition	Component
HM-MM-Bau-11-01	HumanMilk	Bauer_11	Germany	4	CREAM; mLowry; Orcinol_CH; AS; Colorimetry	GestAge(week)=23 | StorageTemp(C)=-70 | MilkStage(day)=[3,56]	Prot(g/L)=!Bauer_prot_GA23
HM-MM-Bau-11-02	HumanMilk	Bauer_11	Germany	10	CREAM; mLowry; Orcinol_CH; AS; Colorimetry	GestAge(week)=32 | StorageTemp(C)=-70 | MilkStage(day)=[3,56]	Prot(g/L)=!Bauer_prot_GA32
HM-MM-Bau-11-03	HumanMilk	Bauer_11	Germany	10	CREAM; mLowry; Orcinol_CH; AS; Colorimetry	GestAge(week)=33 | StorageTemp(C)=-70 | MilkStage(day)=[3,56]	Prot(g/L)=!Bauer_prot_GA33

Field	Description	Value Set	Type of data	Comment
Key	Unique record ID in the Master sheet		Alphanumeric	Unique one word (can be an integer, too); possibly consisting underscore or hyphen as separators
Food	Food name	HumanMilk	Alphanumeric	*Likely to be extended*
Source	Source of information	_Source	Alphanumeric	Syntax checked like the Master sheet
Region	Geographic region of cohort	_Region	Alphanumeric	Tree-structured definition sheet
CohortSize	Cohort size	[1, 99999]	Numeric	
MeasMethod	Ananlytical method of measurement	_MeasMethod	Alphanumeric	Tree-structured definition sheet
Condition	Various conditions and history around birth	_Condition	Alphanumeric	*Tree-structured definition sheet.*
Component	Food component per-volume concentration or relationships bet	_Component	Alphanumeric	*Tree-structured definition sheet.*

Figure 16. Description of fields in The Master sheet of MilkyBase.

Definition sheets are organized in a tree structure. The tree-structure is utilized by the syntax check Excel VBA Add-ins accompanying the database. It is also useful to carry out estimations on specific molecule content via Bayesian and/or interval arithmetic. Namely, if for example immunoglobulin-A and immunoglobulin-G were measured together, then a new variable handily named "immunoglobulinA+immunoglobulinG" can be introduced and the measured data will be the input for this new, composite variable. This operation corresponds to the merge of two numerical intervals. When the concentration of only one of the two immunoglobulin molecules is to be estimated, then this can be done with a Bayesian conditional probability if, from other, independent data, the proportion of that specific immunoglobulin concentration within the "immunoglobulinA+immunoglobulinG" composite is available.

Another useful feature of the tree-structured data recording emerges when the proportion between two variables is available. For example the "X/Y" is the name of a variable that indicates the proportion of

X compared to Y. This is important because (especially with fatty acids), sometimes it is not clear what "a percentage of a molecule". For example, the proportions "omega6/PUFA" and "omega6/TotalFat" are obviously not the same, still the article scanned assumes that it is clear for the reader. In this case, other MilkyBase records or simply the expertise of the data inputter should be used to estimate the denominator, so the definition (and its interpretation interval) will become clear (hence the name "interval arithmetics").

Figure 17 shows an example of a specific protein which is included in the Definition sheet assigned to the field "Component". Immunoglobulin A is the final leaf of the tree, representing the target molecule. The node "immunoglobulins" represents a wider group of molecules (all immunoglobulins) through which a specific immunoglobulin molecule (e.g., immunoglobulin-A) can be reached. And so on; "protein" is a wider group (all the protein molecules) that includes the immunoglobulin subset; a node in the tree though which that subset can be reached.

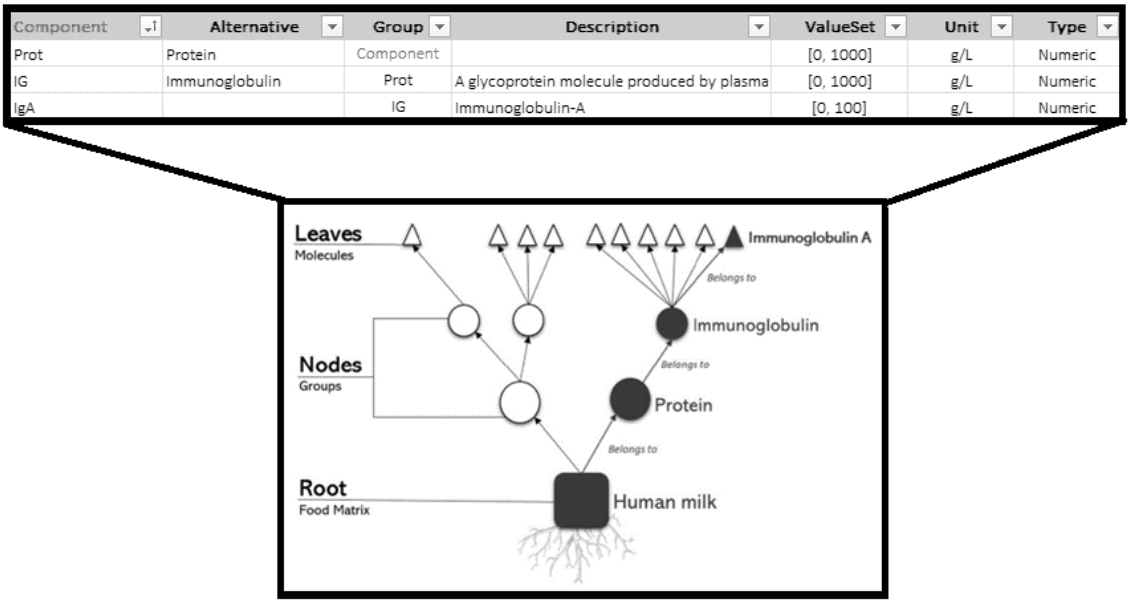


Figure 17. Example of tree structure in the Component sheet.

Beside the Master and definition sheets, MilkyBase includes a dedicated sheet for temporal data, named "DynVal" sheet. This is for "time-value" pairs, representing the temporal variation of the variable in question. **Figure 18** shows an example of such dynamic component, where the concentration of sodium was

measured at six timepoints, from the 11th until the 81st, postpartum days, therefore its temporal profile was easily plotted and visualized.

DynVal	sar	time	tim	time	timeH	timeUni	Component	value	vauleSD	valu	valueH	valueU
Butte84aP_Na		11		8	14	day	Na	0.266	0.098			g/L
Butte84aP_Na		25		22	28	day	Na	0.165	0.034			g/L
Butte84aP_Na		39		36	42	day	Na	0.149	0.03			g/L
Butte84aP_Na		53		50	56	day	Na	0.131	0.03			g/L
Butte84aP_Na		67		64	70	day	Na	0.135	0.037			g/L
Butte84aP_Na		81		78	84	day	Na	0.15	0.044			g/L

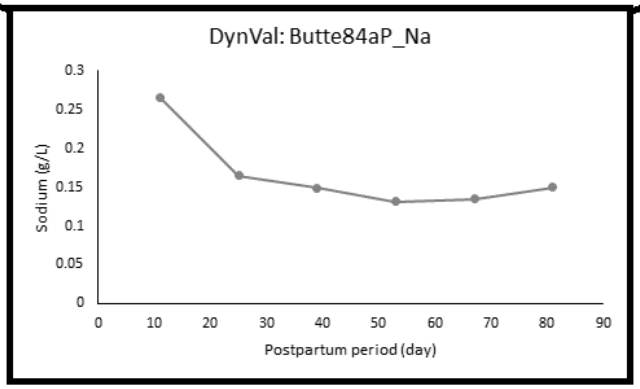


Figure 18. Example of dynamic data and analysis of temporal profile in MilkyBase.

Temporal profiles are the main target entries of Milkybase, the ideal singular entities, because they are the variables of the dynamic mathematical models on which the predictive power of the database is based. In summary, MilkyBase holds 841 records in the Master sheet, from 141 different scientific publications, adding up ca 10,000 of data points on HMC, including more than 7,000 dynamic datapoints (longitudinal data).

4.2 FINDING INCONSISTENCIES IN PUBLICATIONS VIA MILKYBASE

Through the construction of MilkyBase, various errors were found in publications. One of them is the confusing use of the terms “standard deviation”, and “standard error”. The first quantifies the scatter of the sample around its measured mean, the second quantifies the expected error of the estimation of the real mean (BARDE & BARDE, 2012; NAGELE, 2003). As MilkyBase ontology separates the measured data from the estimates, it is straightforward to find such inconsistencies.

MilkyBase is structured in a way that allows simple checks, by visualization, via comparing suspicious data with other published data. For instance, **Figure 19** shows the plotted concentration of alpha-tocopherol content of HM from six publications. Axis y shows the concentration in grams per liter of human milk, and the horizontal axis shows the spread of the data for better visualization. In two of these publications the standard deviation values were given and demonstrated by the first data point with error bars (paper #1) which represents 10 mothers in Canada who had their samples collected at the 14th postpartum day (key HM-TP-Eli-11-01), following by four datapoints with error bars (paper #2) from a sample of 13 mothers in the USA who had their mature milk collected (>28 postpartum days) and processed via different heating methods, each datapoint there represent a different heating procedure (HM-LQ-Lim-20-01 to HM-LQ-Lim-20-04).

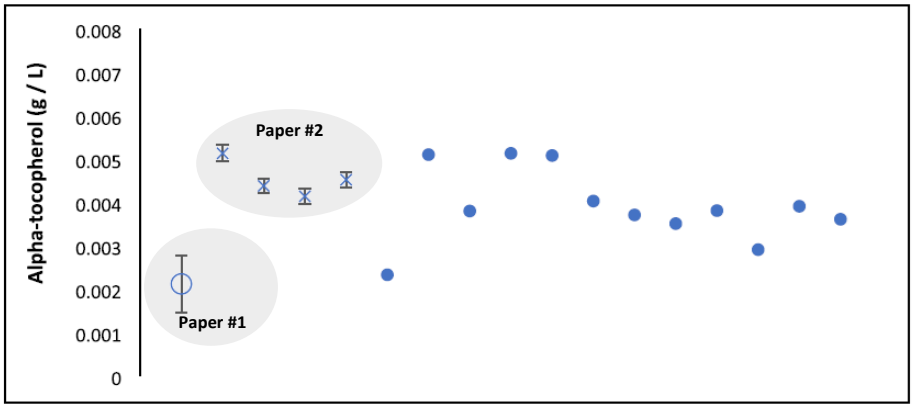


Figure 19. Example of published errors regarding confusion on standard deviation and standard error of the mean found in MilkyBase.

It is notorious that the reported standard deviation of paper #1 was much larger than the following 4 datapoints extracted from paper #2. When paper #2 passed through a re-evaluation, it turned out that the authors should not have reported it as standard deviation (which is the scatter of the observed raw data) but the standard error of the mean.

4.3 DEMONSTRATING THE TEMPORAL VARIATION OF HM PROTEIN CONTENT

It is to be noted that the remit of this first version of MilkyBase was not a systematic review of the literature, much rather to collect and clean many of them to reach a critical mass for data analysis. Thus, it

may not include all the relevant data that are available, and any scientific conclusion should be interpreted with caution. Protein content in HM was chosen for a demo of the potential of MilkyBase due to the vast number of available longitudinal data. In total, 21 publications were inputted in MilkyBase on the temporal changes of protein levels in HM (**Table 6**).

Table 6. List of publications in MilkyBase used on protein analysis of human milk.

Key of MilkyBase	Title of paper	DOI
HM-MM-And-83	Length of Gestation and Nutritional Composition of Human Milk	doi.org/10.1093/ajcn/37.5.810
HM-MM-Arn-87	Protein, Lactose and Fat Concentration of Breast Milk of Mothers of Term and Premature Neonates	doi.org/10.1111/j.1440-1754.1987.tb00276.x
HM-MM-Bau-1	Longitudinal Analysis of Macronutrients and Minerals in Human Milk Produced by Mothers of Preterm Infants	doi.org/10.1016/j.clnu.2010.08.003
HM-TP-Bri-86	Milk Protein Quality in Mothers Delivering Prematurely: Implications for Infants in the Intensive Care Unit Nursery Setting	doi.org/10.1097/00005176-198601000-00021
HM-TP-Cam-09	Effect of Domperidone on the Composition of Preterm Human Breast Milk	doi.org/10.1542/peds.2008-3441
HM-TP-Cha-81	Protein and Amino Acids of Breast Milk From Thai Mothers	doi.org/10.1093/ajcn/34.6.1126
HM-MM-Czo-18	Breast milk macronutrient components in prolonged lactation	doi.org/10.3390/nu10121893
HM-LQ-Czo-19	Lactoferrin in Human Milk of Prolonged Lactation	doi.org/10.3390/nu11102350
HM-TP-Joh-19	Macronutrient variability in human milk from donors to a milk bank. Implications for feeding preterm infants	doi.org/10.1371/journal.pone.0210610
HM-CR-Kul-81	Changes in Human Milk Composition During the Initiation of Lactation	doi.org/10.1038/icb.1981.6
HM-TP-Lem-82	Differences in the Composition of Preterm and Term Human Milk During Early Lactation	doi.org/10.1203/00006450-198202000-00007
HM-MM-Mal-18	Preterm human milk macronutrient concentration is independent of gestational age at birth	doi.org/10.1136/archdischild-2016-312572
HM-CR-Mic-90	Variation in Macronutrients in Human Bank Milk: Influencing Factors and Implications for Human Milk Banking	doi.org/10.1097/00005176-199008000-00013
HM-MM-Mon-99	Immunological and nutritional composition of human milk in relation to prematurity and mothers' parity during the first 2 weeks of lactation	doi.org/10.1097/00005176-199907000-00018
HM-MM-Nom-91	Determinants of energy protein lipid and lactose concentrations in human milk during the first 12 mo of lactation the DARLING Study	doi.org/10.1093/ajcn/53.2.457
HM-TP-Pau-20	Circadian Changes in the Composition of Human Milk Macronutrients Depending on Pregnancy Duration: A Cross-Sectional Study	doi.org/10.1186/s13006-020-00291-y

Table 6 Cont. List of publications in MilkyBase used on protein analysis of human milk.

Key of MilkyBase	Title of paper	DOI
HM-CR-Rei-85	Vitamin B6- And Protein Concentrations in Breast Milk From Mothers of Preterm and Term Infants	doi.org/10.1055/s-2008-1033924
HM-MM-Saa-05	Macronutrient and energy contents of human milk fractions during the first six months of lactation	doi.org/10.1111/j.1651-2227.2005.tb02070.x
HM-LQ-San-86	Changes in the Protein Fractions of Human Milk During Lactation	doi.org/10.1159/000177172
HM-LQ-San-81	Comparison of the Composition of Breast Milk From Mothers of Term and Preterm Infants	doi.org/10.1111/j.1651-2227.1981.tb07182.x
HM-MM-Van-93a	Milk of patients with tightly controlled insulin-dependent diabetes mellitus has normal macronutrient and fatty acids composition	doi.org/10.1093/ajcn/57.6.938

Based on these 21 publications, the time of collection (postpartum day) and the concentration of protein (g/L) were entered in MilkyBase, along with the correspondent explanatory fields (region, measurement analytical methods and mother-milk-infant conditions.

Fig. 20A presents the original data of **Table 6** and **Fig. 20B** shows the logarithm of the protein concentrations data on the logarithmic timescale. The log-log plot provides a better visualization of the temporal variation of the protein content of HM. It is noticeable that, between the 60^{th} and 250^{th} postpartum days (months 2 to 8), the drop in the protein concentration versus the logarithm of time is close to linear. So, visually, the days 60 and 250 appear to be key markers in the temporal variation of protein in HMC.

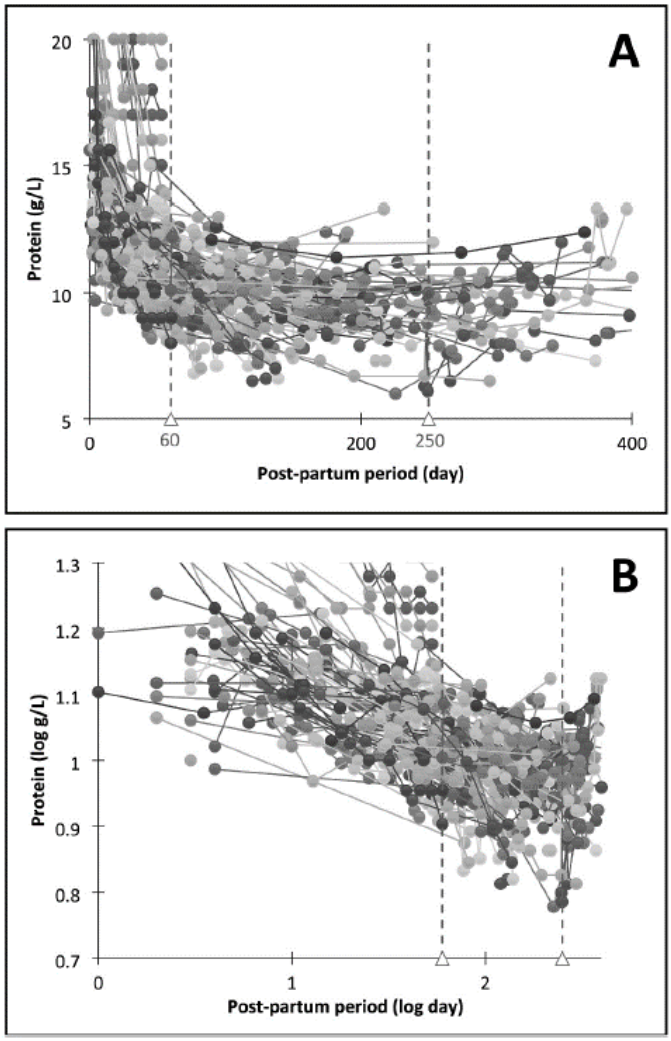


Figure 20. Protein concentration of human milk on the arithmetical (A) and on the logarithmic (B) timescale

Complementarily, **Figure 21** shows another example to validate such finding with four different papers that reported near-to-parallel trajectories (changes over time of protein content) on the log-time scale, which is aligned with the overall pattern found in **Figure 20B.**

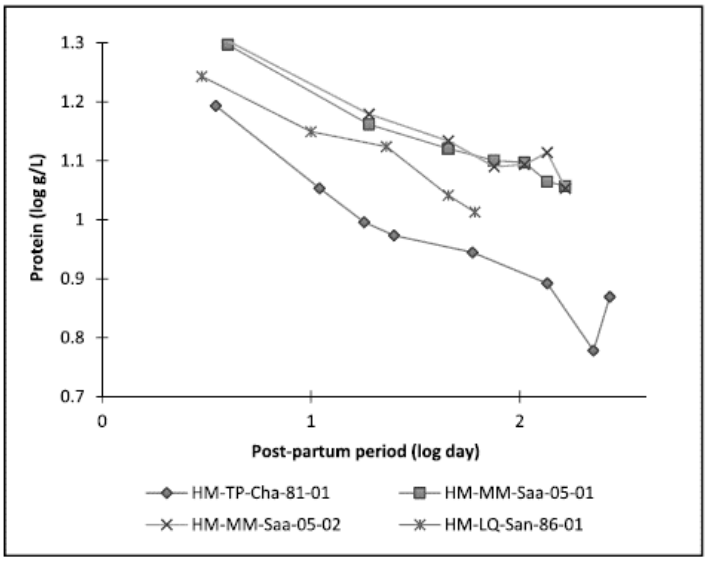


Figure 21. Published longitudinal data on the protein concentration of human milk from different publications, plotted on log-log scale.

5. DISCUSSION

MilkyBase was constructed following the primary – secondary modelling structure taken over from predictive microbiology ideas. Its focus is the temporal variation (primary model) of a HM component as a response to various factors (secondary model) related to the mother-milk-infant triad. The database has been supplied with VBA macros to support syntax and semantic checks along with its user manual. Therefore, the present study created a brand-new FCD to cover gaps in HMC research as mentioned previously.

FCD-s carry fundamental data that allow stakeholders from private and public sectors to gather and utilize food composition information. National food databases are the most popular among users, including renown researchers of the field (KAPSOKEFALOU et al., 2019).The main purpose of national FCD-s are data extraction and validation. However, frequently, vital details of the recorded information (origin, way of sample collection, etc) is missing probably due to difficulties in tracking the process how the FCD was populated (CLANCY et al, 2015). An alternative source of information for food composition may be extracting individual data from retailers and producers via their websites or private databases. However, this alternative has not been explored by researchers yet (KAPSOKEFALOU et al., 2019).

The most popular national and international FCD-s are the FoodData Central from the United States Department of Agriculture (USDA, 2023), the International Network of Food Data Systems, Food and Agriculture Organization (INFOODS) created by FAO (INFOODS/FAO, 2023b), and the European Food Information Resource Association Internationale *Sans but Lucratif* (EuroFIR, 2023) (FERRAZ DE ARRUDA et al 2023; DELGADO et al 2021; KAPSOKEFALOU et al, 2019).

Regarding HM research, there are important limitations related to data quality in a conventional FCD. FoodData Central (USDA, 2023) is based on HMC-information from the 70s, without any details on the maternal or infant characteristics, sampling, storage, or the methodology applied to measure components

(AHUJA et al, 2022).In contrast, MilkyBase does includes such background along with details such as sample size, region and measurement methods.

A traditional FCD commonly focuses on the general energy, macro- (protein, fat, carbohydrate) and micronutrients (vitamins and minerals), with less emphasis on bioactive components, such as antioxidants (DELGADO et al 2021; KAPSOKEFALOU et al, 2019). MilkyBase, on the other hand, comports data from macro-, micronutrients and bioactive components, at a similar priority level.

Likewise, to date, the structure of current FCD-s do not easily accommodate dynamic data of food composition (ALETA ET AL., 2022; KAPSOKEFALOU ET AL., 2019), possibly due to the complexity behind gathering and harmonizing different databases into a single template. Unlike other FCD-s, the very focus of MilkyBase is temporal data.

Current FCD-s lack information on some factors influencing food composition, such as, environmental aspects (temperature, harvest, soil, etc) (FERRAZ DE ARRUDA et al 2023; KAPSOKEFALOU et al, 2019). It is known that from the beginning of food production, raw foods do change their composition in relation to agricultural practices and animals' diet (CARNOVALE et al, 2000; SISSENER et al, 2018), and that is similar to the role that factors related to the mother-milk-infant triad play in HMC.

Data related to milk samples include information on the mother (age, gestational age, BMI, etc.), infant (weight at birth, allergy, etc.) and the sampling of the milk (storage, handling, etc.). These are captured by MilkyBase in its explanatory fields followed, in the same record, by HMC components in the response fields. A very cornerstone of MilkyBase was to follow the structure explanatory-response variables in order to allow predictive modelling based on the data it stores.

To guarantee veracity and quality of collected data, MilkyBase can record uncertainty and spread (interval) values of HMC measurements. The standard deviation or minimax values along with any published estimation derived from them (confidence interval or standard error of the mean) are examples of

these. The (i) availability of uncertainty data in addition to facts that (ii) HMC is organized in a hierarchical tree structure and (iii) the focus is on dynamic data, are the main novelties of MilkyBase.

The tree structure recording is a major novelty of the database; it is vital not only for both syntax and semantics checks. Indeed, it played a crucial role in making use of uncertain information as well as finding outliers and errors in the published data entered in the database. Although MilkyBase is currently far from being a big data project, with appropriate support. It does have the potential to increase its volume to a big data level, when its value can be truly appreciated. To date, the capability of MilkyBase has been shown by brief demos, such as the HM protein dynamics.

In summary, there was a need for a new HMC database with computational tools in HM research (AHUJA et al 2022) and MilkyBase is an answer to that need. It was developed with novelties that can be used to make more precise decisions in infant nutrition, with the potential of being used generally in the FCD field.

6. SUMMARY

The main barriers to advance the current knowledge in HMC research were (i) most of the studies are focused on studying the impact of single aspects of mother-milk-infant triad on individual components of HM, (ii) multi-omics methods are used in cross-sectional studies rather than longitudinal ones and (iii) there is a notable deficiency of knowledge on advanced computational and statistical tools to properly analysed data on the field.

At the moment, there is no FCD prepared to comport such complexity behind a HMC, which behaves as a biological system. By following the low-hanging fruit principle to unlock HMC knowledge and evolve the research field, the present dissertation initiated a brand-new framework to hold and analyze HMC data in an innovative manner.

MilkyBase stores dynamic data of HMC and carries information on explanatory factors related to the mother-milk-infant triad. It brings together three novelties and valuable features for HM and FCD research field, (i) the target on temporal data (response and explanatory fields); (ii) the quantification of the uncertainties and (iii) the tree structured organization of explanatory and response variables, allowing users to execute probabilistic estimations equivalent to interval arithmetic.

Our database came to cover needs in HM science and proposes a new era for data management in FCD research where dynamic data can be stored and analysed in order to make more precise decisions in nutrition and food science.

7. REFERENCES

7.1 DISSERTATION

- Ahuja JKC, Casavale KO, Li Y, Hopperton KE, Chakrabarti S, Hines EP, Brooks SPJ, Bondy GS, MacFarlane AJ, Weiler HA, Wu X, Borghese MM, Ahluwalia N, Cheung W, Vargas AJ, Arteaga S, Lombo T, Fisher MM, Hayward D, Pehrsson PR. Perspective: Human Milk Composition and Related Data for National Health and Nutrition Monitoring and Related Research. Adv Nutr. 2022 Dec 22;13(6):2098-2114. doi: 10.1093/advances/nmac099. PMID: 36084013; PMCID: PMC9776678.

- Ahuja JKC, Casavale KO, Li Y, Hopperton KE, Chakrabarti S, Hines EP, Brooks SPJ, Bondy GS, MacFarlane AJ, Weiler HA, Wu X, Borghese MM, Ahluwalia N, Cheung W, Vargas AJ, Arteaga S, Lombo T, Fisher MM, Hayward D, Pehrsson PR. Perspective: Human Milk Composition and Related Data for National Health and Nutrition Monitoring and Related Research. Adv Nutr. 2022 Dec 22;13(6):2098-2114. doi: 10.1093/advances/nmac099. PMID: 36084013; PMCID: PMC9776678.

- Amaral YN, Marano D, Silva LM, Guimaraes AC, Moreira ME. Are there changes in the fatty acid profile of breast milk with supplementation of omega-3 sources? a systematic review. Rev Bras Ginecol Obstet. (2017) 39:128–141. doi: 10.1055/s-0037-1599094
- Andreas NJ, Kampmann B, Mehring Le-Doare K. Human breast milk: A review on its composition and bioactivity. Early Hum Dev. 2015 Nov;91(11):629-35. doi: 10.1016/j.earlhumdev.2015.08.013. Epub 2015 Sep 12. PMID: 26375355.
- Azad MB, Vehling L, Chan D, Klopp A, Nickel NC, McGavock JM, Becker AB, Mandhane PJ, Turvey SE, Moraes TJ, Taylor MS, Lefebvre DL, Sears MR, Subbarao P; CHILD Study Investigators. Infant Feeding and Weight Gain: Separating Breast Milk From Breastfeeding and Formula From Food. Pediatrics. 2018 Oct;142(4):e20181092. doi: 10.1542/peds.2018-1092. PMID: 30249624.

- Bachour P, Yafawi R, Jaber F, Choueiri E, Abdel-Razzak Z. Effects of smoking, mother's age, body mass index, and parity number on lipid, protein, and secretory immunoglobulin a concentrations of human milk. Breastfeed Med. (2012) 7:179–88. doi: 10.1089/bfm.2011.0038
- Baird J, Fisher D, Lucas P, et al. Being big or growing fast: systematic review of size and growth in infancy and later obesity. BMJ. 2005; 331:929–934. [PubMed: 16227306]
- Baranyi J, Tamplin ML. ComBase: a common database on microbial responses to food environments. J Food Prot. 2004 Sep;67(9):1967-71. doi: 10.4315/0362-028x-67.9.1967. PMID: 15453591.
- BARDANZELLU, F., PERONI, D.G. & FANOS, V. Human Breast Milk: Bioactive Components, from Stem Cells to Health Outcomes. Curr Nutr Rep 9, 1–13 (2020). https://doi.org/10.1007/s13668-020-00303-7
- Barde, M., & Barde, P. (2012). What to use to express the variability of data: Standard deviation or standard error of mean? Perspectives in Clinical Research, 113–116. https://doi.org/10.4103/2229-3485.100662
- Barker D.J. & Osmond C. (1986) Infant mortality, childhood nutrition, and ischaemic heart disease in England and Wales. Lancet, 1, 1077–1081.
- Barker DJ. The developmental origins of adult disease. J Am Coll Nutr. 2004 Dec;23(6 Suppl):588S-595S. doi: 10.1080/07315724.2004.10719428. PMID: 15640511.
- Barrera C, Valenzuela R, Chamorro R, Bascunan K, Sandoval J, Sabag N, et al. The impact of maternal diet during pregnancy and lactation on the fatty acid composition of erythrocytes and breast milk of chilean women. Nutrients. (2018) 10:839. doi: 10.3390/nu10070839
- Batko K, Ślęzak A. The use of Big Data Analytics in healthcare. J Big Data. 2022;9(1):3. doi: 10.1186/s40537-021-00553-4. Epub 2022 Jan 6. PMID: 35013701; PMCID: PMC8733917.
- Becker GE, Smith HA, Cooney F. Methods of milk expression for lactating women. Cochrane Database Syst Rev.2015:CD006170. doi: 10.1002/14651858.CD006170.pub4

- Bhutta ZA, Das JK, Rizvi A, Gaffey MF, Walker N, Horton S, Webb P, Lartey A, Black RE; Lancet Nutrition Interventions Review Group, the Maternal and Child Nutrition Study Group. Evidence-based interventions for improvement of maternal and child nutrition: what can be done and at what cost? Lancet. 2013 Aug 3;382(9890):452-477. doi: 10.1016/S0140-6736(13)60996-4. Epub 2013 Jun 6. Erratum in: Lancet. 2013 Aug 3;382(9890):396. PMID: 23746776.
- Bode L, Raman AS, Murch SH, Rollins NC, Gordon JI. 2020. Understanding the mother-breastmilk-infant 'triad'. Science 367:1070–1072. https://doi.org/10.1126/science.aaw6147.
- Borràs-Novell, C., Herranz Barbero, A., Balcells Esponera, C. et al. Influence of maternal and perinatal factors on macronutrient content of very preterm human milk during the first weeks after birth. J Perinatol 43, 52–59 (2023). https://doi.org/10.1038/s41372-022-01475-6
- Boyce C, Watson M, Lazidis G, Reeve S, Dods K, Simmer K, et al. Preterm human milk composition: a systematic literature review. Br J Nutr. (2016) 116:1033–45. doi: 10.1017/S0007114516003007
- Caba-Flores MD, Ramos-Ligonio A, Camacho-Morales A, Martínez-Valenzuela C, Viveros-Contreras R, Caba M. Breast Milk and the Importance of Chrononutrition. Front Nutr. 2022 May 12;9:867507. doi: 10.3389/fnut.2022.867507. PMID: 35634367; PMCID: PMC9133889.
- Carnovale E, Nicoli S. Changes in fatty acid composition in beef in Italy. J Food Compos Anal. (2000) 13:505–10. doi: 10.1006/jfca.2000.0908
- Chang YC, Chen CH, Lin MC. The macronutrients in human milk change after storage in various containers. Pediatr Neonatol. (2012) 53:2059. doi: 10.1016/j.pedneo.2012.04.009
- Chatfield C (1949) Food Composition Tables for International Use FAO Nutritional Study no. 3. FAO UN: Washington, DC. Available at: http://www.fao.org/documents/show_cdr.asp?url_file=/docrep/x5557e/x5557e00.htm.
- Chen CN, Lin YC, Ho SR, Fu CM, Chou AK, Yang YH. Association of Exclusive Breastfeeding with Asthma Risk among Preschool Children: An Analysis of National Health and Nutrition

Examination Survey Data, 1999 to 2014. Nutrients. 2022 Oct 12;14(20):4250. doi: 10.3390/nu14204250. PMID: 36296941; PMCID: PMC9607098.

- CHRISTIAN P et al. The need to study human milk as a biological system. Am J Clin Nutr. 113(5):1063-1072. DOI:10.1093/ajcn/nqab075 (2021)
- Church, S.M. The history of food composition databases. Nutr. Bull. 2006, 31, 15–20.
- Clancy, A.K.;Woods, K.; McMahon, A.; Probst, Y. Food Composition Database Format and Structure: A UserFocused Approach. PLoS ONE 2015, 10, e0142137.
- De Weerth, A. A. et al (2022) Human milk: From complex tailored nutrition to bioactive impact on child cognition and behavior, Critical Reviews in Food Science and Nutrition, DOI: 10.1080/10408398.2022.2053058
- Deichmann U. Epigenetics: The origins and evolution of a fashionable topic. Dev Biol. 2016 Aug 1;416(1):249-254. doi: 10.1016/j.ydbio.2016.06.005. Epub 2016 Jun 9. PMID: 27291929.
- Delgado A, Issaoui M, Vieira MC, Saraiva de Carvalho I, Fardet A. Food Composition Databases: Does It Matter to Human Health? Nutrients. 2021 Aug 17;13(8):2816. doi: 10.3390/nu13082816. PMID: 34444976; PMCID: PMC8399939.
- Demmelmair H, Koletzko B. Lipids in human milk. Best Pract Res Clin Endocrinol Metab. (2018) 32:57–68. doi: 10.1016/j.beem.2017.11.002 21.
- Eunice Kennedy Shriver National Institute of Child Health and Human Development (NIH). (2023b) Breastmilk Ecology: Genesis of Infant Nutrition (BEGIN) concept. Available online: https://www.nichd.nih.gov/sites/default/files/inline-files/BEGIN_Human_Milk_concept.pdf , (accessed on 22 April 2023).
- Eunice Kennedy Shriver National Institute of Child Health and Human Development (NIH). (2023a). Breastmilk Ecology: Genesis of Infant Nutrition (BEGIN) project. [Internet]; [cited 2023 Mar 18]. Available from: https://www.nichd.nih.gov/research/supported/begin
- European Food Information Resource (EuroFIR). (2023). Food data: List of EuroFIR databases. Central. Available online https://www.eurofir.org/food-information/food-composition-databases/

- Ferraz de Arruda H, Aleta A, Moreno Y. Food composition databases in the era of Big Data: Vegetable oils as a case study. Front Nutr. 2023 Jan 5;9:1052934. doi: 10.3389/fnut.2022.1052934. PMID: 36687693; PMCID: PMC9851468.
- Franzago M, Alessandrelli E, Notarangelo S, Stuppia L, Vitacolonna E. Chrono-Nutrition: Circadian Rhythm and Personalized Nutrition. Int J Mol Sci. 2023 Jan 29;24(3):2571. doi: 10.3390/ijms24032571. PMID: 36768893; PMCID: PMC9916946.
- Garwolińska D., Namieśnik J., Kot-Wasik A.a, and Hewelt-Belka W. Journal of Agricultural and Food Chemistry 2018 66 (45), 11881-11896 DOI: 10.1021/acs.jafc.8b04031
- Geddes DT, Gridneva Z, Perrella SL, Mitoulas LR, Kent JC, Stinson LF, Lai CT, Sakalidis V, Twigger AJ, Hartmann PE. 25 Years of Research in Human Lactation: From Discovery to Translation. Nutrients. 2021 Aug 31;13(9):3071. doi: 10.3390/nu13093071. PMID: 34578947; PMCID: PMC8465002.
- George AD, Burugupalli S, Paul S, Mansell T, Burgner D, Meikle PJ. The Role of Human Milk Lipids and Lipid Metabolites in Protecting the Infant against Non-Communicable Disease. Int J Mol Sci. 2022 Jul 6;23(14):7490. doi: 10.3390/ijms23147490. PMID: 35886839; PMCID: PMC9315603.
- Gidrewicz, D.A., Fenton, T.R. A systematic review and meta-analysis of the nutrient content of preterm and term breast milk. BMC Pediatr 14, 216 (2014). https://doi.org/10.1186/1471-2431-14-216
- Gila-Diaz, A.; Arribas, S.M.; Algara, A.; Martín-Cabrejas, M.A.; López de Pablo, Á.L.; Sáenz de Pipaón, M.; Ramiro-Cortijo, D. A Review of Bioactive Factors in Human Breastmilk: A Focus on Prematurity. *Nutrients* 2019, *11*, 1307. https://doi.org/10.3390/nu11061307
- Gillman MW. Early infancy - a critical period for development of obesity. J Dev Orig Health Dis. 2010 Oct;1(5):292-9. doi: 10.1017/S2040174410000358. PMID: 25141932; PMCID: PMC4643648.

- Gluckman PD, Hanson MA, Pinal C. The developmental origins of adult disease. Matern Child Nutr. 2005 Jul;1(3):130-41. doi: 10.1111/j.1740-8709.2005.00020.x. PMID: 16881892; PMCID: PMC6860944.
- Grice EA, Segre JA. The human microbiome: our second genome. Annu Rev Genomics Hum Genet (2012) 13:151–70. doi:10.1146/annurev-genom- 090711-163814
- Hahn-Holbrook, J., Saxbe, D., Bixby, C. et al. Human milk as "chrononutrition": implications for child health and development. Pediatr Res 85, 936–942 (2019). https://doi.org/10.1038/s41390-019-0368-x
- Hooton, F., Menichetti, G. & Barabási, A.-L. Exploring food contents in scientific literature with FoodMine. Scientific Reports 10, https://doi.org/10.1038/s41598-020-73105-0 (2020).
- Horta BL, de Sousa BA, de Mola CL. Breastfeeding and neurodevelopmental outcomes. Curr Opin Clin Nutr Metab Care. 2018 May;21(3):174-178. doi: 10.1097/MCO.0000000000000453. PMID: 29389723.
- Horta BL, Rollins N, Dias M, Garcez V, Pérez-Escamilla R. Systematic review and meta-analysis of breastfeeding and later overweight or obesity expands on previous study for World Health Organization. Acta Paediatr 2022; published online June 21. https:// doi.org/10.1111/apa.16460.
- Innis SM. Impact of maternal diet on human milk composition and neurological development of infants. Am J Clin Nutr. (2014). 99:734– 41S. doi: 10.3945/ajcn.113.072595
- International Network of Food Data Systems, Food and Agriculture Organization (INFOODS/FAO). Food composition challenges (2023a). Available online https://www.fao.org/infoods/infoods/food-composition-challenges/en/ (accessed on 20 Mar 2023).
- International Network of Food Data Systems, Food and Agriculture Organization (INFOODS/FAO). (2023b). Available online https://www.fao.org/infoods/infoods/ en/ (accessed on 20 Mar 2023).

- Kapsokefalou M, Roe M, Turrini A, Costa HS, Martinez-Victoria E, Marletta L, Berry R, Finglas P. Food Composition at Present: New Challenges. Nutrients. 2019 Jul 25;11(8):1714. doi: 10.3390/nu11081714. PMID: 31349634; PMCID: PMC6723776.
- Keikha M, Bahreynian M, Saleki M, Kelishadi R. Macro- and micronutrients of human milk composition: are they related to maternal diet? A comprehensive systematic review. Breastfeed Med. (2017) 12:517–27. doi: 10.1089/bfm.2017.0048
- Koletzko B, Brands B, Grote V, Kirchberg FF, Prell C, Rzehak P, et al. Long-Term Health Impact of Early Nutrition: The Power of Programming. Annals of nutrition & metabolism. 2017; 70:161–69. [PubMed: 28683464]
- Koletzko B, Godfrey KM, Poston L, Szajewska H, van Goudoever JB, de Waard M, Brands B, Grivell RM, Deussen AR, Dodd JM, Patro-Golab B, Zalewski BM; EarlyNutrition Project Systematic Review Group. Nutrition During Pregnancy, Lactation and Early Childhood and its Implications for Maternal and Long-Term Child Health: The Early Nutrition Project Recommendations. Ann Nutr Metab. 2019;74(2):93-106. doi: 10.1159/000496471. Epub 2019 Jan 23. PMID: 30673669; PMCID: PMC6397768.
- Konig J (1878). Chemie der Menschlichen Nahrungs- und Genussmittel. Springer: Berlin.
- Kugananthan S, Gridneva Z, Lai CT, Hepworth AR, Mark PJ, Kakulas F, et al. Associations between maternal body composition and appetite hormones and macronutrients in human milk. Nutrients. (2017) 9:252. doi: 10.3390/nu9030252
- Lev HM, Ovental A, Mandel D, Mimouni FB, Marom R, Lubetzky R. Major losses of fat, carbohydrates and energy content of preterm human milk frozen at−80 degrees C. J Perinatol. (2014) 34:396–8. doi: 10.1038/jp.2014.8
- Lu M, Jiang J, Wu K, Li D. Epidermal growth factor and transforming growth factor-alpha in human milk of different lactation stages and different regions and their relationship with maternal diet. Food Funct. (2018) 9:1199‐204. doi: 10.1039/C7FO00770A

- Makela J, Linderborg K, Niinikoski H, Yang B, Lagstrom H. Breast milk fatty acid composition differs between overweight and normal weight women: the STEPS study. Eur J Nutr. (2013) 52:727–35. doi: 10.1007/s00394-012-0378-5
- Martin CR, Ling PR, Blackburn GL. Review of Infant Feeding: Key Features of Breast Milk and Infant Formula. Nutrients. 2016 May 11;8(5):279. doi: 10.3390/nu8050279. PMID: 27187450; PMCID: PMC4882692.
- Micha R, Coates J, Leclercq C, Charrondiere UR, Mozaffarian D. Global dietary surveillance: data gaps and challenges. Food Nutr Bull. (2018) 39:175–205. doi: 10.1177/0379572117752986
- Monteiro PO, Victora CG. Rapid growth in infancy and childhood and obesity in later life – a systematic review. Obes Rev. 2005; 6:143–154. [PubMed: 15836465]
- Nagele, P. (2003). Misuse of standard error of the mean (sem) when reporting variability of a sample. A critical evaluation of four anaesthesia journals. British Journal of Anaesthesia, 514–516. https://doi.org/10.1093/bja/aeg087
- Nunes CA, Alvarenga VO, de Souza Sant'Ana A, Santos JS, Granato D. The use of statistical software in food science and technology: Advantages, limitations and misuses. Food Res Int. 2015 Sep;75:270-280. doi: 10.1016/j.foodres.2015.06.011. Epub 2015 Jun 11. PMID: 28454957.
- Ong KK, Loos RJ. Rapid infancy weight gain and subsequent obesity: systematic reviews and hopeful suggestions. Acta Paediatr. 2006; 95:904–908. [PubMed: 16882560]
- Painter RC, de Rooij SR, Bossuyt PM, Simmers TA, Osmond C, Barker DJ, Bleker OP, Roseboom TJ. Early onset of coronary artery disease after prenatal exposure to the Dutch famine. Am J Clin Nutr. 2006 Aug;84(2):322-7; quiz 466-7. doi: 10.1093/ajcn/84.1.322. PMID: 16895878.
- Pacza, T., Baranyi, J., Martins, L.M., 2022. MilkyBase, a database of human milk composition as a function of maternal-, infant- and measurement conditions. doi:10.6084/m9.figshare.c.6160191.v1
- Peter D. Gluckman; Mark A. Hanson; Catherine Pinal (2005). The developmental origins of adult disease. , 1(3), 130–141. doi:10.1111/j.1740-8709.2005.00020.x

- Powe CE, Knott CD, Conklin-Brittain N. Infant sex predicts breast milk energy content. Am J Hum Biol. (2010) 22:50–4. doi: 10.1002/ajhb.20941
- Ramos-Lopez O, Milagro FI, Riezu-Boj JI, Martinez JA. Epigenetic signatures underlying inflammation: an interplay of nutrition, physical activity, metabolic diseases, and environmental factors for personalized nutrition. Inflamm Res. 2021 Jan;70(1):29-49. doi: 10.1007/s00011-020-01425-y. Epub 2020 Nov 24. PMID: 33231704; PMCID: PMC7684853.
- Rey-Mariño A, Francino MP. Nutrition, Gut Microbiota, and Allergy Development in Infants. Nutrients. 2022 Oct 15;14(20):4316. doi: 10.3390/nu14204316. PMID: 36297000; PMCID: PMC9609088.
- Ristevski, B., Chen, M.. "Big Data Analytics in Medicine and Healthcare" Journal of Integrative Bioinformatics, vol. 15, no. 3, 2018, pp. 20170030. https://doi.org/10.1515/jib-2017-0030
- ROBINSON S, FALL C. Infant nutrition and later health: a review of current evidence. Nutrients. 2012 Aug;4(8):859-74. doi: 10.3390/nu4080859. Epub 2012 Jul 26. PMID: 23016121; PMCID: PMC3448076.
- Rumbold P, McCullogh N, Boldon R, Haskell-Ramsay C, James L, Stevenson E, Green B. The potential nutrition-, physical- and health-related benefits of cow's milk for primary-school-aged children. Nutr Res Rev. 2022 Jun;35(1):50-69. doi: 10.1017/S095442242100007X. Epub 2021 Apr 27. PMID: 33902780.
- Sánchez, C. L. et al. Evolution of the circadian profile of human milk amino acids during breastfeeding. J. Appl. Biomed. 11, 59–70 (2013).
- Shenhav, L, and M. B. Azad. 2022. Using community ecology theory and computational microbiome methods to study human milk as a biological system. mSystems 7 (1):e0113221. doi: 10.1128/msystems. 01132-21.
- Sissener NH, Suarez RK, Hoppeler HH. Are we what we eat? Changes to the feed fatty acid composition of farmed salmon and its effects through the food chain. J Exp Biol. (2018) 221 (Suppl_1):jeb161521. doi: 10.1242/jeb.161521

- Thakkar SK, Giuffrida F, Cristina CH, De Castro CA, Mukherjee R, Tran LA, et al. Dynamics of human milk nutrient composition of women from Singapore with a special focus on lipids. Am J Hum Biol. (2013) 25:770–9. doi: 10.1002/ajhb.22446
- Toure, V., Flobak, Å., Niarakis, A., Vercruysse, S., & Kuiper, M. (2020). The status of causality in biological databases: Data resources and data retrieval possibilities to support logical modeling. Briefings in Bioinformatics, bbaa390. https://doi.org/10.1093/bib/bbaa390
- United States Department of Agriculture (USDA). (1992). Weights, Measures, and Conversion Factors for Agricultural Commodities and Their Product. Available online: https://www.ers.usda.gov/webdocs/publications/41880/33132_ah697_002.pdf (accessed on 22 April 2023).
- United States Department of Agriculture (USDA). (2023). FoodData Central. Available online: https://fdc.nal.usda.gov/ (accessed on 20 Mar 2023).
- Urwin HJ, Miles EA, Noakes PS, Kremmyda LS, Vlachava M, Diaper ND, et al. Salmon consumption during pregnancy alters fatty acid composition and secretory IgA concentration in human breast milk. J Nutr. (2012) 142:1603–10. doi: 10.3945/jn.112.160804
- USAID Advancing Nutrition. Advancements in Understanding Breastmilk: Learning from the BEGIN Project GNCP Webinar. [Internet]; [cited 2023 Mar 18]. Available from: https://www.advancingnutrition.org/events/2022/03/23/gncp-webinar-advancements-understanding-breastmilk-learning-begin-project
- Victora CG, Bahl R, Barros AJ, et al. Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect. Lancet 2016; 387: 475–90.
- Voortman T, Garcia AH, Braun KVE, Thakkar SK, Tielemans MJ, Stroobant W, et al. Effects ofMaternal Nutrition on Quantity and Nutritional Quality of Breast Milk: Systematic Review, inWCPGHAN. Montreal, QC(2016)
- Wiedmeier JE, Joss-Moore LA, Lane RH, Neu J. Early postnatal nutrition and programming of the preterm neonate. Nutr Rev. 2011 Feb;69(2):76-82. doi: 10.1111/j.1753-4887.2010.00370.x. PMID: 21294741.

- World Health Organization [WHO]. Department of Nutrition for Health and Development; Department of Child and Adolescent Health and Development. The optimal duration of exclusive breastfeeding: report of an Expert Consultation. Geneva, Switzerland. (2001).
- Yi, D.Y.; Kim, S.Y. Human Breast Milk Composition and Function in Human Health: From Nutritional Components to Microbiome and MicroRNAs. *Nutrients* 2021, *13*, 3094. https://doi.org/10.3390/nu13093094

7.2 CANDIDATE'S PUBLICATIONS

UNIVERSITY of DEBRECEN

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

Registry number: DEENK/157/2023.PL
Subject: PhD Publication List

Candidate: Mayara Lopes Martins
Doctoral School: Doctoral School of Nutrition and Food Sciences
MTMT ID: 10084850

List of publications related to the dissertation

1. **Martins, M. L.**, Pacza, T., Müller, K. E., Baranyi, J.: A computational approach to nutrition science reveals the dynamics of the protein content of human milk.
   Innovative Food Science & Emerging Technologies. 82, 1-5, 2022.
   DOI: http://dx.doi.org/10.1016/j.ifset.2022.103167
   IF: 7.104 (2021)
2. Pacza, T., **Martins, M. L.**, Rockaya, M., Müller, K. E., Chatterjee, A., Barabási, A. L., Baranyi, J.: MilkyBase, a database of human milk composition as a function of maternal-, infant- and measurement conditions.
   Sci Data. 9 (1), 1-7, 2022.
   DOI: http://dx.doi.org/10.1038/s41597-022-01663-1
   IF: 8.501 (2021)

Total IF of journals (all publications): 15,605
Total IF of journals (publications related to the dissertation): 15,605

The Candidate's publication data submitted to the iDEa Tudóstér have been validated by DEENK on the basis of the Journal Citation Report (Impact Factor) database.

15 May, 2023

DEBRECENI EGYETEM Egyetemi és Nemzeti Könyvtár

8. KEYWORDS

Food composition

Food database

Computational nutrition

Data science

Human milk

Big Data

9. ACKNOWLEDGEMENT

First, this thesis is a proof of faith in my God, Jesus Christ, who is my whole inspiration in life. This work was possible only because of His work..

I would like to thank my supervisors József Baranyi and Katalin E. Müller, whom I dearly call bosses, for their support over these years in Hungary! This was an incredible journey, thanks for all your lessons, patience, opportunities, friendship, and wisdom. You are the best bosses, I admire you so much, hope to have you in my life as dear friends.

To my dear friend and research partner Tünde Pacza for all the hard work and friendship over this PhD course, how blessed I am for having you by my side, this papers are our babies! Thank you so much!

To my family in Brazil who give me support and love. To my mother and father who always prioritized education in order to prepare their daughters for the world, I love you so much and hope to do the same with my children. In particular, to my mother, for believing in and encouraging me to pursue my dreams, you are my whole model of a woman, Mom.

To my husband-to-be, Daniel, who was the person who gave me this crazy idea to pursue my Ph.D. in Europe, what an unbelievable idea that, later, became a reality! This thesis is for our family!

To my dear research community in the Nutrition Institute at the University of Debrecen for having me as a collaborator and believing in the MilkyBase project. I especially thank Judit for all the support over these years, you are the heart of this institute!

Lastly, I would like to thank the Stipendium Hungaricum Scholarship Programme for granting my scholarship in Hungary.

10. APPENDIX

The present dissertation is based on the following publications: