Università degli Studi di Padova

VARIATIONAL PRINCIPLES AND OPTIMALITY IN BIOLOGICAL SYSTEMS

VICE COORDINATORE:
PROF. CINZIA SADA

SUPERVISORE DELLA RICERCA:
PROF. AMOS MARITAN

CANDIDATO:
LOREN KOÇILLARI

Abstract

The aim of this thesis is to investigate the signatures of evolutionary optimization in biological systems, such as in proteins, human behaviours and transport tissues in vascular plants (xylems), by means of the Pareto optimality analysis and the calculus of variations.

In the first part of this thesis, we address multi-objective optimization problems with tradeoffs through the Pareto optimality analysis ([132],[69]), according which the best tradeoff solutions correspond to the optimal species, enclosed onto low-dimensional geometrical polytopes, defined as Pareto optimal fronts, in the space of physical traits, called morphospace. Chapter 3 is devoted to the Pareto optimality analysis in the *Escherichia coli* proteome by projecting proteins onto the space of solubility and hydrophobicity. In chapter 4 we analyze the HCP dataset of cognitive and behavioral scores in 1206 humans, in order to identify any signature of Pareto optimization in the space of Delay Discounting Task (DDT), which measures the tendency for people to prefer smaller, immediate monetary rewards over larger, delayed rewards.

The second part of this thesis is devoted to solving an optimization problem regarding xylems, which are the internal conduits in angiosperms that deliver water and other nutrients from roots to petioles in plants. Based on the optimization criteria of minimizing the energy dissipated in a fluid flow, we propose in chapter 5 a biophysical model with the goal of explaining the underlying physical mechanism that affects the structure of xylem conduits in vascular plants, which results in tapered xylem profiles [104, 105, 117, 164]. We address this optimization problem by formulating the model in the context of the *calculus of variations*.

The results of these investigations, besides providing quantitative support to previous theories of natural selection, demonstrate how processes of optimization can be identified in different biological systems by applying statistical methods such as the Pareto optimality and the variational one, showing the relevance of employing these statistical approaches to various biological systems.

Riassunto

Lo scopo di questa tesi è quello di identificare le impronte che l'evoluzione ha avuto nei sistemi biologici, come ad esempio nelle proteine, nei comportamenti umani e nei tessuti trasportatori delle piante vascolari (xilemi), attraverso un'analisi di ottimizzazione di Pareto ed il calcolo delle variazioni.

Nella prima parte della tesi, affrontiamo l'ottimizzazione di problemi multi-obiettivo con competizione, attraverso l'analisi di ottimizzazione di Pareto, in base alla quale le migliori soluzioni di compromesso corrispondono alle specie ottimali, le quali vengono racchiuse in politopi geometrici, definiti come fronti ottimali di Pareto, nello spazio dei tratti fisici. Il capitolo 3 è dedicato all'analisi dell'ottimizzazione di Pareto nel proteoma dell'Escherichia coli, proiettando le proteine nello spazio della solubilitá ed idrofobicitá. Nel capitolo 4 analizziamo il set di dati HCP cognitivi e comportamentali in 1206 umani, al fine di identificare qualsiasi traccia di ottimizzazione alla Pareto nello spazio del "Delay Discounting Task" (DDT), che misura la tendenza per le persone a preferire ritorni economici piú piccoli e immediati rispetto a ricompense di premi piú grandi e ritardati.

La seconda parte di questa tesi è dedicata alla risoluzione di un problema di ottimizzazione riguardante gli xilemi, che sono i condotti interni degli angiospermi e forniscono con acqua ed altri nutrienti le piante, dalle radici ai piccioli. Basandosi sui criteri di ottimizzazione per minimizzare l'energia dissipata in un flusso di fluido, nel capitolo 5 proponiamo un modello biofisico con l'obiettivo di spiegare il meccanismo fisico sottostante che influenza la struttura di condotti dello xilema nelle piante vascolari, che si traducono in profili di xilema affusolati. Affrontiamo questo problema di ottimizzazione formulando il modello nel contesto del calcolo delle variazioni.

I risultati di queste indagini, oltre a fornire supporto quantitativo sulle precedenti teorie sulla selezione naturale, dimostra come i processi dell'ottimizzazione possono essere identificati in diversi sistemi biologici applicando metodi statistici come l'ottimalitá di Pareto e il variazionale

66 uno, mostrando la rilevanza di impiegare questi approcci statistici a vari
67 sistemi biologici.

Publications

Published:

[76] **Loren Koçillari**, Piero Fariselli, Antonio Trovato, Flavio Seno, & Amos Maritan, (2018). *Signature of Pareto optimization in the Escherichia coli proteome.* Scientific reports, **8(1)**, 9141. (**I.F = 4.122**)

Accepted for publication:

[31] Giorgia Cona*, **Loren Koçillari***, Alessandro Palombit, Alessandra Bertoldo, Amos Maritan, Maurizio Corbetta *"Archetypes of human cognition defined by time preference for reward and their brain correlates: an evolutionary trade-off approach"* [for *Neuroimage* Journal], * = 1st co-author

Accepted for publication:

[114] Rodrigo Rocha, **Loren Koçillari**, Samir Suweis, Maurizio Corbetta, Amos Maritan *"Homeostatic plasticity and emergence of functional networks in a whole-brain model at criticality"* [for *Scientific Reports* Journal]

In Preparation:

Loren Koçillari, Rodrigo Rocha, Alberto Lovison, Samir Suweis, Mark Olson, Tommaso Anfodillo, Jaynath Banavar, Franco Cardin and Amos Maritan *A model of xylem tapering in vascular plants* [in submission]

In Preparation:

Loren Koçillari, Amos Maritan, Flavio Seno, Mariella Rasotto and An-

Table of contents

Chapter 1

Introduction

The common thread that permeates this thesis is the idea that nature exerts a selective pressure for optimizing structures and functions in biological systems in order for them to best adapt to their ecosystem ([101], [40]). In the course of evolution, organisms carry out multiple tasks to strive for survival, which may lead to complex tradeoffs, meaning that the performance levels of all tasks cannot be concurrently optimized. To unravel how tradeoffs affect the phenotype selection we employed a statistical approach, developed in a recent paper by Shoval et al., [132], based on the Pareto optimality theory, devised initially to solve multi-objective optimization problems with competing objectives in economics and engineering, to identify evolutionary tradeoffs in biological systems.

According to Pareto optimality, optimal phenotypes (different species, individuals within a species, circuits, bacteria, proteins, etc.) that correspond to the best possible tradeoff solutions among different physical traits, such as the body mass, longevity, brain size etc, should be enclosed into low-dimensional geometrical polytopes, such as a segment, a triangle, a pentagon, etc., also referred to as the Pareto optimal fronts, in the space of traits, called morphospace. Without any tradeoff, phenotypes would be instead distributed in an uncorrelated cloud of points in the morphospace.

In the first three chapters, we highlight and discuss our findings concerning the signatures of Pareto optimality in biological systems. In chapter 2 we set the terminology and notations of Pareto optimality and define the fundamental concepts of multi-objective optimization, dominance and Pareto fronts in the objective space and in the morphospace ([87], [89]).

Chapters 3 and 4 are devoted to the application of Pareto optimal analysis to the biological systems, where the fitness, which is defined as an increasing function of the performance functions at all tasks, is harder to disclose. In chapter 3, we will present the first original result

which supports the emergence of signatures of Pareto optimization in the *Escherichia coli* proteome, by tuning the degree of hydrophobicity necessary for the proteins to fold correctly and that of solubility in order to perform their biological functions. In chapter 4, on the other hand, we show original findings in the context of the Human Connectome Project (HCP) dataset, by investigating cognitive and behavioral scores in 1206 humans through Pareto optimality.

The second part of this thesis is devoted to give a biophysical explanation of the tapering phenomenon of xylem conduits in vascular plants. Existing models of the tapering of xylem conduits ([130], [157], [123]) assume that xylem profiles have acquired a tapering degree in order to optimally convey water and essential nutrients to all parts of the trees ([104, 105, 117, 164]). Following this line of thought, we propose in chapter 5 a hydraulic optimal model, based on the optimization criteria of minimizing the energy dissipated in a fluid flow, which is due to the Hagen-Poiseuille resistance term. We address this optimization problem by formulating it in the context of the *calculus of variations*, where we define the main functional made up of the Hagen-Poiseuille resistance term and a Lagrange multiplier.

Finally, in chapter 6 we summarize all findings and discuss some further prospects.

Chapter 2

Pareto optimality in biological systems

All biological systems, or phenotypes, must efficiently perform multiple tasks to strive for survival. In some instances, the performance levels cannot be concurrently optimized for all tasks, so that the competition between them affects phenotype selection. Consequently, organisms evolve and adapt themselves to the environment through a precise trade-off. In order to accomplish this complex decision making, species are needed to solve an implicit multi-objective optimization problem (MOO).

To fully disclose the properties of this complex multi-objective optimization problem, scientists have employed the Pareto optimal analysis [17, 28, 36, 89, 92]. Basically, a solution of the multi-objective optimization problem is called *Pareto optimal*, if there does not exist any feasible (possible) solution which would increase any performance without inducing a concurrently decrease of at least another performance. Solving a multi-objective optimization problem often results, even in the simplest case of two competing objectives, in a continuum and infinite set of Pareto optimal solutions, named Pareto fronts (see Figure 4.1).

Since the Pareto optimal solutions are all equivalent, a *decision maker* (DM) is required to introduce further information to choose for the preferred solution of the problem. In biology for instance, the decision maker is the niche itself, which, under the pressure of evolutionary selection, prefers those species that are endowed with traits that best adapt to the environment. Without any extra preferential criteria, solution cannot be sorted and thus they are all equivalent between each other. In other words, they stand as the best compromises in performing competing tasks.

Historically, Francis Ysidro Edgeworth in 1881 ([43]) and Vilfredo Pareto [100] in 1906 have been the pioneers that formulated the framework
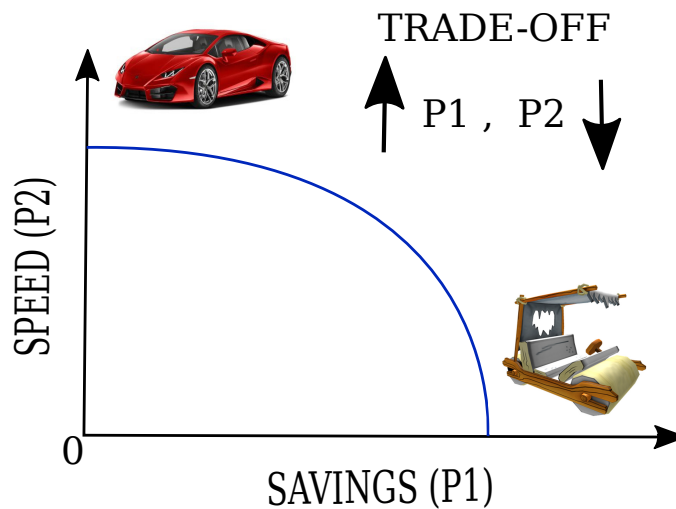
Fig. 2.1 Example of a Pareto optimal front. Here we sketch a multi-objective optimization problem in designing a car, between two tasks which present a tradeoff. The Pareto front is the continuous blue line, which represents the set of optimal designs. The optimal solution should be selected from the front by giving an additional information about the weights of each task in the final decision.

of multi-objective optimization problems with competing objectives. In 1951, Kuhn and Tucker posed the necessary and sufficient conditions for the Pareto efficiency [80]. For a detailed history of MOOs in the objective space consult Stadler and Dauer ([139]). Since then, a plethora of computational algorithms have been implemented in order to find Pareto fronts [87]. The first algorithms worked only for convex objectives, however, in the mid 1980s David Schaffer devised a more efficient algorithm to overcame this limit, called the vector evaluated genetic algorithm (VEGA), which was the first implementation of a real multi-objective evolutionary algorithm (EA) [124]. A remarkable advantage by employing EAs is that they generate multiple Pareto solutions in a single run [17].

To work properly, these methods require an explicit fitness function. It is the main function of the model, which accounts for any tradeoffs and the weights of all objectives. It is the starting point for both analytical and computational derivations to be made in order to infer the shape of Pareto fronts. It often occurs that this function is difficult, if not impossible, to mathematically disclose [73], especially for biological systems.

In a recent study of Shoval et al. [69, 132] however, they performed a study of Pareto optimality by translating the analysis from the objective or task space to the trait space, also referred to as the morphospace. It corresponds to the space of the quantitative traits that can be experimentally measured, like the mass, longevity, height, solubility, hydrophobicity,

delay discounting area etc., which constitute the phenotype of a given biological system. The choice of any subset of these traits depends both on the availability of datasets and on intuition of traits that could lead to tradeoffs, as suggested from experience.

In [132] they provide a compelling theorem that links Pareto optimal fronts with convex regions in the space of traits (see Appendix A for more details). The theorem is based on the following two postulates: *i)* species which are specialized in a given task, also called *archetypes*, cluster in the vertices of the convex-hulls and *ii)* the performance functions are maximal for a given task at the corresponding archetypes and decrease with distance from the archetype. The vertices of such distributions play the crucial role in inferring the tasks in tradeoff.

With the aid of this theorem, Pareto optimal fronts can be found even if it doesn't exist an explicit expression for both the fitness function and task performances. In [132] authors show that Pareto fronts emerge as low-dimensional convex-hulls in the morphospace, such as lines, triangles, tetrahedrons etc., depending on the number of competing objectives.

This is a more appropriate framework for a Pareto analysis in biology and nowadays, this method has been successfully applied to unravel signatures of evolutionary optimization in animal morphology ([132]), animal behavior ([50]), cancer ([69]), ammonite shells ([146]), bacterial and single cells gene expression ([148]; [77]), biological circuits ([142]), and more recently to the structure of polymorphisms ([129]), and to Escherichia coli proteome ([76]).

Pareto optimality can also be used to solve multi-objective problems in human-made systems, in order to find those optimal designs that attain the best compromise of the cost-efficiency ratio. For instance, planning to build a new house requires to find a balance between the costs of the construction and its final achievable comfort. Thus, a decision maker is often lead with a multitude, and possibly infinite equivalent optimal choices. It can resolve this problem by putting further criteria on the weights of each cost ([47]).

2.1 Basic Definitions

Single-objective optimization problems aim to find the minimum of a given function $f_0(x)$. The problem can be stated as follows ([16], [27]):

$$minimize \quad f_0(x) \tag{2.1}$$

$$subject\ to \quad g_i(x) \leq b_i, \quad i = 1, ..., m. \tag{2.2}$$

where $f_o : \mathbf{R}^n \to \mathbf{R}$ is the single-objective function, $x = (x_1, ..., x_n) \in S$ are the decision vectors in the feasible space S, and $g_i : \mathbf{R}^n \to \mathbf{R}$ denote the m constraint functions of the feasible space. For simplicity functions are convex meaning that the following inequality is satisfied:

$$f_i(\beta x^1 + (1 - \beta)x^2) \leq \beta f_i(x^1) + (1 - \beta)f_i(x^2) \tag{2.3}$$

for all $0 \leq \beta \leq 1$, $\sum_{i=1}^{m} \beta_i = 1$ and $\beta_i \geq 0$. If the feasible space $S \subset R^n$ is convex then:

$$\beta x^1 + (1 - \beta)x^2 \in S \tag{2.4}$$

for all $0 \leq \beta \leq 1$. The substantial difference between single and multi-objective functions is that for single-objective problems there is a single optimal solution, while for multi-objectives there is an infinite number of optimal solutions.

When we ask to simultaneously optimize k (≥ 2) conflicting objective functions $f_1(\mathbf{x}), f_2(\mathbf{x}), ..., f_k(\mathbf{x})$ we face a *multi-objective optimization problem* that can be mathematically defined as follows ([165], [166]), [17]:

$$
\begin{aligned}
minimize \quad & \mathbf{y} = \mathbf{f(x)} = (f_1(\mathbf{x}), f_2(\mathbf{x}), ..., f_k(\mathbf{x})) \quad k \geq 2 \\
subject\ to \quad & \mathbf{e(x)} = (e_1(\mathbf{x}), e_2(\mathbf{x}), ..., e_m(\mathbf{x})) \leq 0 \\
and \quad & \mathbf{x} = (x_1, x_2, ..., x_n) \in \mathbf{S} \\
& \mathbf{y} = (y_1, y_2, ..., y_k) \in \mathbf{Y}
\end{aligned}
\tag{2.5}
$$

where k (≥ 2) is the number of the $f_i : \mathbf{R}^n \to \mathbf{R}$ competing objective functions and $\mathbf{x} = (x_1, x_2, ..., x_n)^T$ are the decision vectors that belong to the *feasible region* S, while m is the number of the constrain functions $\mathbf{e(x)}$. Objective functions are images of decision vectors $Z = \mathbf{f}(S)$, where S is the feasible decision space and Z the feasible objective space. A small region of the objective space constitutes the *Pareto front* $P(S)$, namely the set of optimal solutions $\mathbf{z} = \mathbf{f}_i(\mathbf{x})$, which by definition have the property that none of their components could be improved without the worsening of at least another component.

A useful concept for the Pareto optimality is related to the *dominance*. We say that the decision vector $\mathbf{a}$ *dominates* another vector $\mathbf{b}$ if the following conditions are satisfied [166]:

$$\forall i \in \{1, 2, ..., n\} : f_i(\mathbf{a}) \geq f_i(\mathbf{b}) \wedge \tag{2.6}$$

$$\exists j \in \{1, 2, ..., n\} : f_i(\mathbf{a}) > f_i(\mathbf{b}) \tag{2.7}$$

343 Therefore, decision vectors $\mathbf{x} \in \mathbf{X}_f \in S$ are defined as *Pareto optimal* or
344 non-dominated iff:

$$\nexists\, \mathbf{a} \in \mathbf{A} : \mathbf{a} \succ \mathbf{x} \tag{2.8}$$

346 and accordingly, the set of objective vectors is denoted as Pareto front if
347 the corresponding decision vectors are Pareto optimal.

348   In case of minimization problems the lower bound of the Pareto front
349 is the *ideal objective vector*, denoted with $\mathbf{z}^* \in \mathbf{R}^k$, whose components can
350 be obtained by minimizing each objective function separately. Mathemat-
351 ically, it can be expressed as $z^* = f^* = (f_1^*, f_2^*, ..., f_M^*)^T$, where $x^{*(m)}$ is
352 the minimum decision vector solution and f_m^* is the minimum objective
353 solution. The ideal vector is the optimal solution of the multi-objective
354 optimization problem when objectives are not competing.

355

356   2.2   Pareto optimality in the morphospace

357 According to natural selection, biological systems coevolve to maximize
358 their fitness function, resulting in optimal phenotypes. However, when
359 facing complex environments, systems carry out multiple tasks, and all of
360 these tasks contribute to fitness. Hence a fundamental trade-off: As systems
361 cannot achieve optimal performance in all tasks, becoming specialists in
362 one set of tasks necessarily leads to a reduction of performance in a different
363 set of tasks.

364   The starting point of the Pareto Optimality approach is to define the
365 space of traits, or morphospace, where traits represent physical features
366 such as body mass, longevity, brain size etc, and species are usually data
367 points in the morphospace. The Pareto Optimality theory predicts that
368 if traits are likely to show trade-offs, then phenotypes will be enclosed
369 into a well-defined geometrical domain of this morphospace called poly-
370 tope (e.g., a segment, a triangle, a pentagon or other low dimensional
371 polygons/polyhedra...). This polytope will include the phenotypes that
372 have found the best possible trade-off solutions among different traits, and
373 will represent the Pareto front solution (see Figure 4.14). In the absence of
374 trade-offs, phenotypes will be instead distributed in an uncorrelated cloud
375 of points in the morphospace.
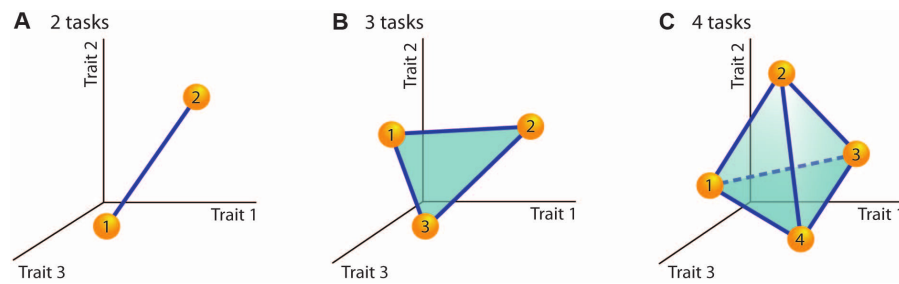
Fig. 2.2 Pareto fronts in morphospace. Here we show in the morphospace examples of Pareto fronts with an increasing number of vertices resulting in a segment (2 tasks), a triangle (3 tasks) and a tetrahedron (4 tasks) (Figure adapted from [132]).

The position of a given phenotype inside the Pareto front distribution is informative of its evolutionary strategy. Specifically, the vertices of the polytope contain the archetypes, namely the phenotypes that have traits leading to the maximal performance in one of the tasks and minimal performance in the competing tasks. Other key biological traits related to that task will be then maximally expressed or 'enriched' near that archetype, and minimally enriched near the other archetypes. Phenotypes that fall in the middle of the polytope are generalists, i.e. showing average performance in those tasks that define the trait space. In the case of two competing tasks, the phenotypes fall on a line segment in the morphospace, whereas for three tasks the phenotypes fall into a triangle. Four tasks would result in a tetrahedron distribution, and so on. Notably, this analysis is data-driven since it is the distribution of the data to indicate which traits are indicator of tradeoffs and what is the number of competing tasks, which correspond to the number of vertices/archetypes in polytopes.

An example of the application of Pareto optimality is the study (**??**). The authors found that species of mammals and birds fall within a triangular Pareto front distribution when they are projected in a morphospace created by the variables longevity and mass. The vertices of this triangle represent three archetypes. Specifically, large animals with high longevity (whales being the archetype); small animals with high longevity (bats); and, small animals with low longevity (mice). All other species, including humans, fall in between. Importantly, through enrichment analysis, it is possible to show that these traits are related to (enrich on) other traits that account for their evolutionary fitness. For instance, small animals with low longevity tend to have high fertility and tend to be preys (mice); conversely, small animals with high longevity have lower fertility, but also tend to be predator (bats).

2.2.1 Theoretical framework for biological systems

Consider a biological system denoted by n physical traits ν_i, $i = 1, ..., n$, which is described as a data point in the morphospace and assume that it concurrently performs k competing tasks. The k performances $P_i(\nu)$ for each task are expressed as functions of the physical traits ν_i. To each phenotype ν is assigned a fitness function $F(P_1(\nu), ..., P_k(\nu))$, which is defined as an increasing function of the k competing performances.

As defined previously, a Pareto optimal solution is associated to a given phenotype ν for which it doesn't exist any other feasible different phenotype ν' that is better at all tasks than ν. The set of all Pareto optimal phenotypes provide the Pareto front. For two tasks, Pareto fronts are line-segments connecting both archetypes and data points in the line-segment are found to be optimal phenotypes of the multi-objective tradeoff problem. For three tasks, Pareto optimal fronts are triangles, while for k-tasks, in principle we should get (k-1)-dimensional simplexes with k vertices, such as tetrahedrons, etc. Pareto optimal fronts could explain the long-standing observation that the morphospace is mostly void and phenotypes typically cluster in small regions ([88], [108],[109], [125]). Based on the principle of natural selection, indeed evolutionary pressures have wipe out the morphospace from species that are sub optimal, leaving only the optimal species inside the Pareto fronts.

According to the theorem shown in Appendix A, it has been proved the relationship among Pareto fronts and convex-hulls in the morphospace. The vertices of these polytopes play a crucial role in the theory, since they are the place where archetypal species sit. Each archetype is defined as the specialized organism that optimally performs a single task at the expenses of the performances of other competing tasks, which are optimally performed by the remaining archetypal species, located in the other vertices of the polytope. The performance of a given task decreases monotonically from the corresponding archetypes toward the center of the convex-hull as follows:

$$P_i(\nu) = P_i((\nu - \nu_i^*)^T M (\nu - \nu_i^*)), \qquad (2.9)$$

where M is a positive-definite matrix denoting the metric of the space. We will consider now on that $M = \mathbf{I}$, which correspond to the euclidean metric.

For Euclidean metrics Pareto fronts are enclosed by straight lines. In that case, the performance function $P_i(\nu)$ is a decreasing function of the Euclidean distance from the corresponding archetype ν_i^*, resulting with circular performance contours 4.15. By connecting the dots where two

443 adjacent circular contours are tangent we get the straight edges of Pareto
444 fronts. In case of more general metrics, the contours of the performance
445 functions would acquire different shapes, and the set of tangent dots will
446 typically result in curved lines [128] (see Figure 4.15). In addition, when
447 performance is maximized in a whole region of archetypes, the Pareto
448 front is the straight line connecting the closest point between the regions
449 of archetypes. For more detailed cases, see [128].
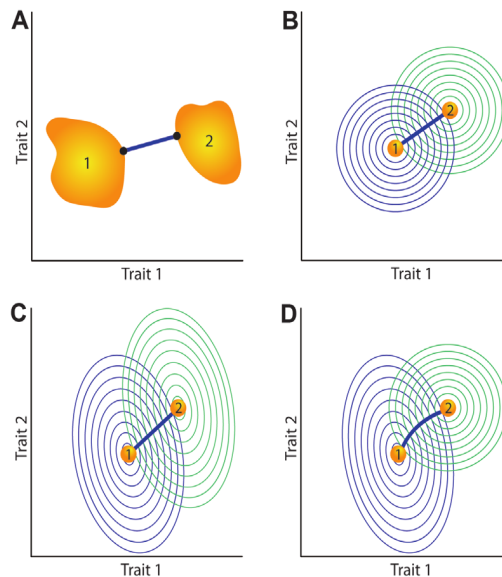
Fig. 2.3 Parti algorithm in the Morphospace. Relaxing some assumptions result
in a curved Pareto fronts in the morphospace. In **a)** it is shown the Pareto front in
case of regions of archetypes instead of single-point archetypes, **b)** a Pareto front
resulting with straight lines in case of Euclidean metric, **c)** and **d)** cases when the
assumption of the Euclidean metric is relaxed (Adapted from [132])

450 ## 2.2.2   Classical examples of Pareto optimality in morphospace

451 In their seminal paper [132], Shoval et al. provided several examples of
452 Pareto fronts in animal morphology for Darwin finches, leaf-cutter ants
453 and microbats, and the gene expression of *Escherichia coli* bacteria.
454   As a first example (Figure 2.4A), they analyzed the dataset of Grant
455 et al. [62] of Darwin's finches and detected a statistically significant tri-
456 angular shaped distribution in the space of body mass and beak shape.
457 Species at the vertices of the triangle were inferred to correspond to three
458 archetypal Darwin finches that feed with totally orthogonal diets, which
459 are supposed to be in tradeoff, namely that it is not possible for a given
460 finches to concurrently feed with the same performance at all diets (see
461 Supplementary Materials of [132]).
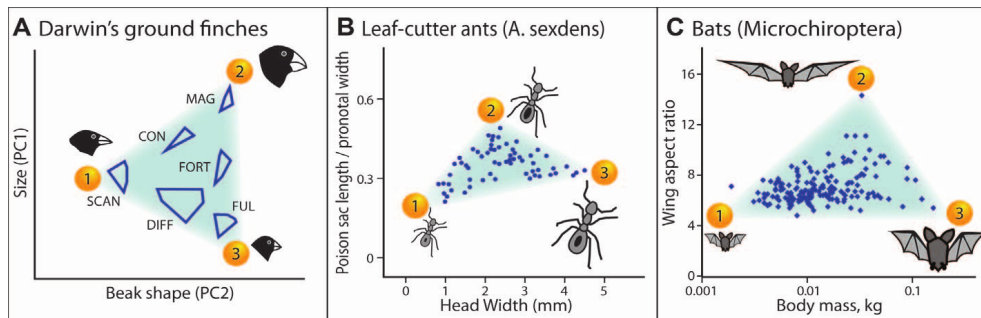
Fig. 2.4 Examples of Pareto fronts in morphospace. In a) a Pareto front for Darwin's ground finches in the trait space of beak shape and size after performing a PCA, b) a triangular-hull Pareto front for the leaf-cutter ants in the space of head width vs a normalized poison sac length, c) a triangular-hull Pareto front in bats in the space of body mass and wing aspect ratio (Adapted from [132]).

As a second example (Figure 2.4B), authors analyzed the dataset E. O. Wilson's [158] on leaf-cutter ants and found another statistically significant triangular-hull in the space of traits such as the head width and poison sac length. They proved that the triangle was a Pareto front by inferring the archetypal ants for each vertex, namely ants that are specialized in : 1) nursing/gardening, 2) foraging outside the nest, and 3) soldiering.

They identified a triangular-hull Pareto front also for the microbats study (Michrochiroptera) of Norberg and Rayner [97] (see Figure 2.4C), in the trait space of the body mass and wind aspect ratio. The three archetypes that correspond to each vertex are interpreted to be associated to specialized microbats in 1) eating insects in vegetations, 2) in the air above the vegetation, and 3) large prey in vegetation.

A Pareto front was found by analyzing the gene expression in the *Escherichia coli* bacteria [163]. They inferred two competing tasks such as rapid growth, mostly provided by the ribosomal genes, and survival, which is mainly provided by the oxidative stress response proteins.

2.3   Pareto optimality in the objective space

As a complementary argument we summarize in this section the idea behind the MOOs algorithm when applied in the objective space. The most frequently used algorithm to attempt a solution of multi-objective optimization problems is the *weighted sum method*. It can be stated as

follows ([51], [161], [90]):

$$minimize \quad y = \sum_{i=1}^{k} w_i f_i(\mathbf{x}) \tag{2.10}$$

$$subject\ to \quad \mathbf{x} \in \mathbf{X}_f \tag{2.11}$$

where $w_i \geq 0 \ \forall i = 1, ..., k$ and $\sum_{i=1}^{k} w_i = 1$. The weighted sum method aims to link to each objective a weighting parameter w_i and linearly sum the parametrized objectives into a single fitness functions. As a result of this process, the multi-objective problems are converted in single-objective problems, which are simpler to be solved. The major weakness of this method is related to the fact that it is unable to find any Pareto optimal solutions for non-convex objective functions. In order to overcome this limitation, it has been developed the so called $\epsilon-$constraint method ([68]), which minimizes only one objective and transforms the other objectives into constraint functions with an ϵ upper bound (see [90]).

A common drawback of these methods is related to their highly demanding computational efforts in order to search the Pareto-optimal solutions in the objective space. However, more efficient algorithms, called evolutionary algorithms (EA), have been established in order to handle the computational limits and search for the whole Pareto optimal front within a single run of simulation [27]. EA algorithms have been first implemented by Schaffer in his pivotal work ([124]), and since then, five classical EA approaches have been developed (for a quantitative comparison of their efficiency see ([165]), which substantially differ in the definition of the fitness function ([166], [47]).

Once the fitness function is clearly stated, EAs have the particular advantage to capture several Pareto optimal solutions simultaneously in a single computational run. Remarkable applications of Pareto optimality in the objective space, by employing classical and evolutionary methods have been applied, ranging from optimal protocols ([135]) in thermodynamics, the design of low-thrust spacecraft trajectories in aeronautics ([34]), to optimal complex networks ([126]), multilayer network growth ([122]), and language networks ([127]). The major limitation of evolutionary algorithms is that they always need to get assigned a specific fitness functions, and thus become ineffective in the majority of cases in biological systems, where the fitness function is only known to exist, but we don't have an explicit expression.
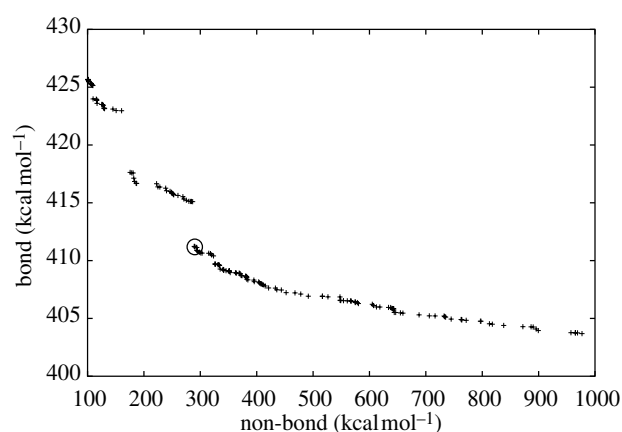
Fig. 2.5 Pareto optimal set in protein morphology. Here we show an example of a Pareto optimal front that emerges as the solution of the folding of proteins in optimal shapes, subjected to the tradeoff objectives such as tension, torsion etc. in the objective space (Adapted from [35])

2.4   Summary

In this chapter we defined in section 2.1 multi-objective optimization problems and introduced the basic concepts of Pareto optimality such as Pareto fronts and dominance and laid the basic terminology. In section 2.2 we posed the rules for studying Pareto fronts by means of the Parti algorithm, even if we don't have a fitness function. This comes at the cost of shifting analysis from the objective to the trait space. We furthermore discussed some examples of Pareto fronts in biology, such as in Darwin finches, ants and bats (microbats), as found by Parti in finding Pareto fronts in biology, which emerge as low-dimensional polytopes in the space of physical traits. In section 2.3 we furnished some literature for the MOOs in the objective space.

In the following chapters of this first part of the thesis we will apply the computational algorithm [69], to investigate signatures of Pareto fronts in the solubility and hydrophobicity space of the proteome of *Escherichia coli* bacteria (see chapter 3), and in human behavioral tasks (see chapter 4).

Chapter 3

Signature of Pareto optimization in the *Escherichia coli* proteome

Pareto polytopes have been shown to enclose the variation of phenotypic traits for organisms of the same species that adapt to different environmental niches, or the variation of gene expression patterns for cells of the same organism that adapt to different tissues (or pathological conditions in the case of tumor cells). In this chapter, we extend the Pareto front analysis to a further downward step toward shorter scales of the proteome of the *Escherichia coli* bacteria. Proteins have coevolved with cellular environments to improve or preserve their functions, maintaining at the same time the degree of hydrophobicity necessary to fold correctly and enough solubility to perform their biological roles.

Here, we study the variation in protein physico-chemical features of solubility-hydrophobicity in the *Escherichia coli* proteome using a Pareto front analysis. We choose the *E.coli* since it is a simple prototype organism which has been widely studied and, furthermore, its genome is extensively annotated.

From the Taguchi's database [95], we extracted the following three continuous characteristics: experimental solubility, experimental yield, and predicted isoelectric point (pI). All quantities were available only for a subset of 3,172 proteins. We added, as a further fundamental continuous trait, an overall measure of protein hydrophobicity, which was obtained by summing up the hydrophobicity values of all its residues according to the Kyte-Doolittle scale [81].

We find evidence that *E.coli* proteins were selected by trading off the performances of various competing tasks and we infer those tasks. Indeed, in section 4.4 we report the results of the Pareto analysis, indicating the emergence of a triangular-hull Pareto optimal front in the space of solubility

and hydrophobicity, whose vertices correspond to archetypal proteins specialized in distinct tasks, such as 1) *regulatory processes*, 2) *membrane transport*, 3) *outer-membrane pore formation, catalysis, and binding*.

In section 3.2 we will set and generalize the theoretical framework of of the state-of-the-art Pareto optimality analysis, in order to connect specific sub-cellular environments with the competing tasks performed by the proteins located in these regions.

In section 3.3 we further show that the vertices are enriched also with proteins that occupy different subcellular compartments, namely, cytoplasmic, inner membrane, outer membrane, and outer membrane bounded periplasmic space. The combination of various enriching features offers an interpretation of how bacteria use the physico-chemical properties of proteins, both to drive them into their final destination in the cell and to have their tasks accomplished.

In section 3.4 we will show that when the Pareto analysis is extended to include protein yield, a tetrahedron emerges as the convex hull representing the new front in $3D$ with the yield feature corresponding to the third principal component.

Finally, in section 5.4 we summarize our results and make some final remarks and discussions.

3.1 A triangle in the space of solubility vs hydrophobicity

Three of the above traits (i.e. the experimental solubility, experimental yield, and predicted isoelectric point (pI)) inherently convey competing chemical characteristics of polypeptide chains concerning both a water-like solvent and different cellular environments, such as the crowded cytoplasm and the interior of biological membranes. The yield, which is how many proteins are expressed by the 'in vitro' reconstituted translation system [95], adds a further characterization.

3.1.1 PCA analysis

With each protein represented by the set of continuous traits defined before, we apply a Principal Component Analysis (PCA) to reduce the dimensionality of the morphospace and search for Pareto polytopes. The PCA variance is mainly explained (about 95%) by two principal components that are substantially parallel to the hydrophobicity (PC1) and solubility
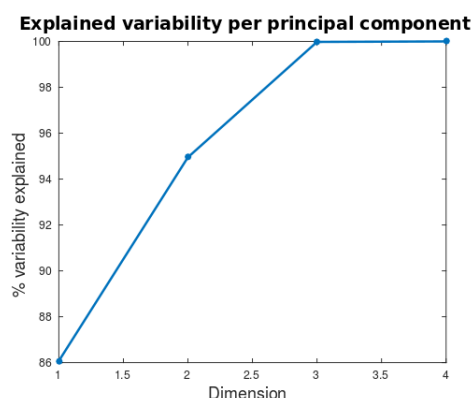
Fig. 3.1 PCA for the four dimensional space of continuous traits. The first component is better explained by the hydrophobicity, the second component by the solubility, whereas the third component by the protein yield (see Table 1). The first two traits, i.e. solubility and hydrophobicity, are able to explain around 95% of the overall variability. We achieve almost the total variability if we consider also the third principal component, but in this three dimensional morphospace the convex hull is affected by robustness caveats (see Section 3.2).

(PC2) trait, respectively (Table 3.2, Figure 3.1). This can be rationalized by considering that hydrophobicity is the dominant force implicated in the folding process of globular proteins [5, 20, 23, 41], whereas solubility is a property that emerges as a necessary feature to prevent protein aggregation [38, 144, 151], and, consequently, the onset of relevant maladies in humans [26]. Solubility also appears to be related to mRNA expression levels, at least for specific proteins [145]. The maintenance of protein solubility is also a fundamental aspect of protein homeostasis [38], being an essential requirement for protein functionality. Furthermore, proteins are evolutionarily selected to perform necessary and useful functions, so they must be stable (at least marginally) but also flexible enough to accomplish their tasks through relevant conformational changes.

If we z-score solubility-hydrophobicity-yield-pI traits before the PCA, we find that the variance changes, with the pI trait which this time has relevant loadings in the first two principal components. However, by projecting the data points in the first two principal components, as obtained from the z-scored traits, the resulting convex hull is not a triangle anymore, with a p-value>0.05, as evaluated from the t-ratio test.

3.1.2   PCHA analysis

We performed the archetypal analysis, introduced by Cutler and Breiman [36], whose goal is to find the best-fitting convex hull of the data in the

trait space, that is the solution of the minimization problem. This can be done computationally by the PCHA algorithm, developed by Morup et al. [92] and implemented in the Pareto Task Inference (ParTI) developed by Hart et al [69]. This algorithm allowed us to find the explained variance of the convex hull that encloses the data points, as a function of the number of vertices (see Figure 3.2). The positions of the vertices of the convex hull in the trait space were determined by employing the Sisal algorithm [12] which is analogous to PCHA but considers in a more flexible way the presence of outliers and the possibility that archetypes lie outside the convex hull [69]. See Table 3.1 for the archetype positions found using Sisal, after 100 iterations, and Figure 3.4 for the archetype positions using different types of algorithms. We also computed the errors in the positions of the archetypes by employing the so called bootstrapping method [69]. This relies on the generation of n-bootstrapped datasets with the same number of proteins $(3, 172)$ as the original dataset, and on computing from each new dataset the corresponding archetype positions. We generated 10^4 bootstrapped datasets, and we computed their center of mass and the standard deviations of archetype positions. Errors are depicted as ellipsoids in Figure 3.3.
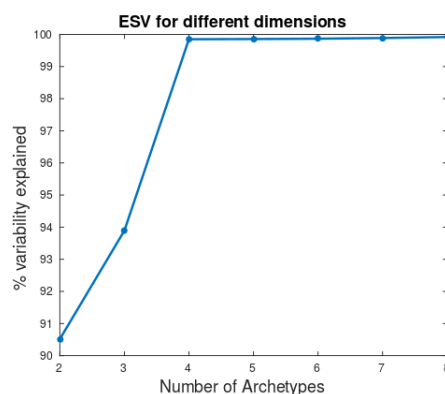
Fig. 3.2 Number of archetypes. Explained variance [92] of the data points as a function of the number of archetypes. In our analysis, we considered only the first three archetypes, which account for 94% of the total variance.
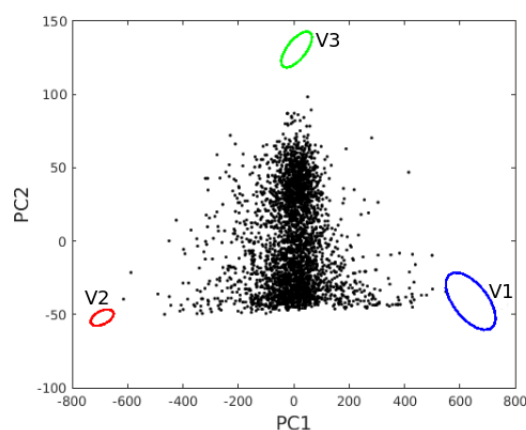
Fig. 3.3 Archetype positions. Error distribution of the coordinates of the vertices of the triangle as obtained by the Sisal algorithm[12] performing 10^4 bootstrapped datasets .

639

Arch (PCA) Position	Hydrophobicity (PC1)	Solubility (PC2)
Blue	639.2	-41.0
Red	-691.6	-52.2
Green	10.9	130.5

640

Arch (Orig) Position	Hydrophobicity	Solubility
Blue	572.4	1.5
Red	-751.7	1.1
Green	7.3	193.9

Table S 3.1 Position of the three archetypes as found with Sisal. The positions of the three vertices in the principal component plane are shown in the top table, whereas the same positions in the solubility -hydrophobicity plane are shown in

641 the bottom table.
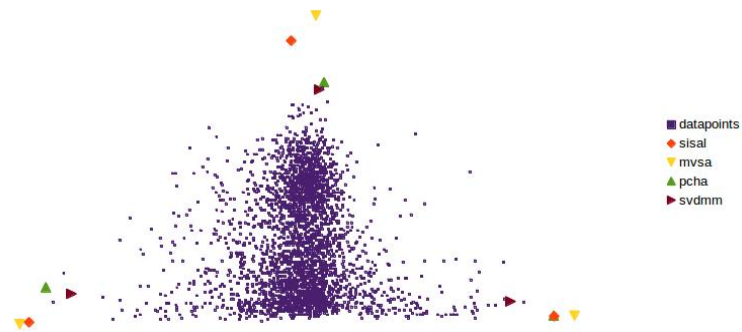
Fig. 3.4 Archetype coordinates. Archetype coordinates evaluated with four different methods such as Sisal, PCHA, MVSA, SVDMM. They give equivalent results.
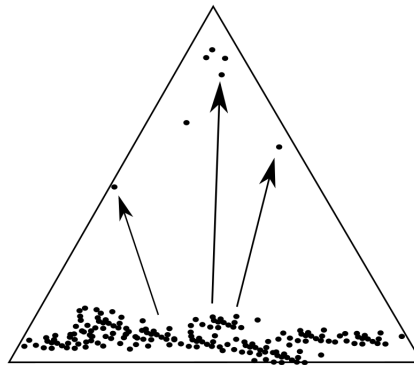
Fig. 3.5 Robustness of the Pareto front. PCHA analysis does not necessarily imply that the data are well distributed on a convex hull. Sometimes Pareto analysis cannot be applied, for example in cases where the outliers dominate the statistics and triangles appear even when the majority of points clusters only in specific regions of the convex hull and a few outliers are responsible for adding other vertices.
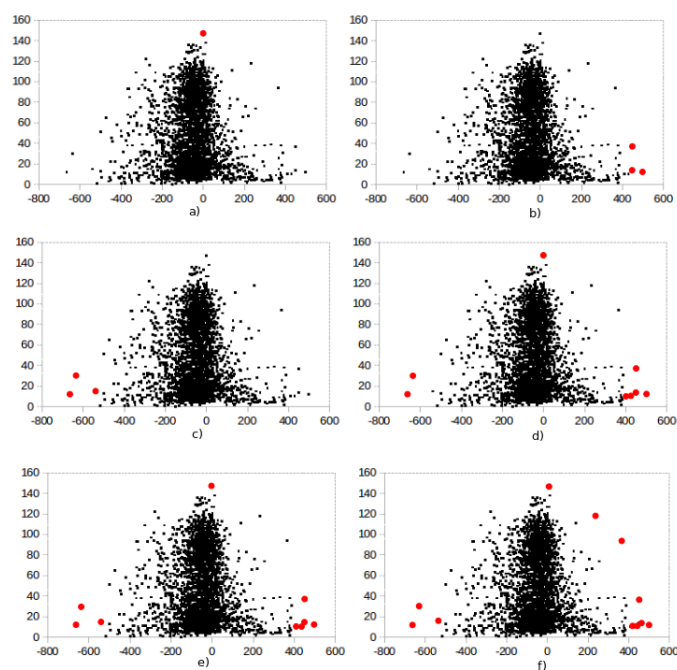
Fig. 3.6 Robustness of the triangle in the solubility vs hydrophobicity plane. We computed the p-value, after removing the proteins in red, for each case. For a) p-value= 0.5%, b) p-value=0.4% , c) p-value< 0.01%, d) p-value=0.06%, e) p-value= 0,04%, f) p-value< 0.01%

3.1.3 Statistical robustness of the Pareto front

In the solubility-hydrophobicity space, the *E.coli* proteins lie inside a triangle, a clear hallmark of Pareto optimality (Figure 3.8). The statistical significance of the detected Pareto front is assessed using the p-value [132], which is based on the t-ratio, defined as the ratio between the area of the triangular convex hull (in Figure 3.8), and the area of the minimum triangle in which the convex hull can be embedded. The t-ratio of the experimental data points is then compared to the t-ratios of 10^4 null-models, generated by randomizing pairs of solubility and hydrophobicity values from the original data, i.e., by taking the same cumulative distribution function (CDF), along single axes, as in the original dataset. The resulting p-value is lower than $5 * 10^{-3}$ and in literature, p-values lower than 5% are accepted as highly significant. Pareto analysis however, can be hampered when the results are heavily influenced by the presence of some outliers (see Figure S6). Statistically speaking, the results must be, as much as possible, outlier-independent. More practically, the deletion of a small number of data points in the above analysis must not affect archetype identification and the p-value of the detected polytope. We generated 10^4 null-models for all of the six possible combinations of the four continuous traits, finding that the most robust triangles with the lowest p-values are projected in the hydrophobicity-solubility and hydrophobicity-yield planes (p-value of the order of 0.5%). In the remaining four cases the lowest p-value is higher than 5%. We further found that the triangle in the yield-hydrophobicity plane is strongly dependent on outliers, while the triangle in the solubility-hydrophobicity plane is very robust. In the former case, the p-value fluctuates in the range $0.5\% - 10\%$ when (up to 4) proteins with the highest yield are removed, while in the latter case the p-value is almost unaffected (see Figure S6).

the volume of the convex hull with a higher number of vertices that encloses the majority of the data points. The t-ratio is usually larger than 1, and the closer it is to 1, the better the polytope captures the shape of the data.
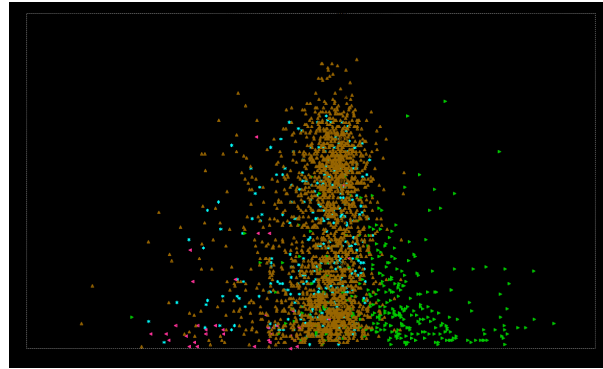
Fig. 3.7 The Pareto front. Data points in the space of solubility vs hydrophobicity. Proteins are coloured as follows. Green:Inner membrane, Yellow:Cytoplasmic, Light blue:Periplasmic-bounded outer membrane, Rose:Outer membrane.

3.2 Theoretical framework

We theoretically extended the state-of-the-art Pareto analysis [132], in order to connect specific sub-cellular environments with the competing tasks performed by the proteins located in these regions. We made the following assumptions:

(*i*) The bacterium environments are characterized by specific concentrations, $(\rho_1, \rho_2, \ldots, \rho_n) \equiv \rho$, of n chemicals (water, lipids, etc.). As one moves from one place to another, ρ varies with continuity at the mesoscopic scale. This is a formal representation of the fact that, even though bacterial cells lack membrane-bounded organelles, they are intricately organized, with different chemical concentrations in different locations [32, 59, 120].

(*ii*) Each protein can perform k possible tasks/activities, and to each of them (the j-th task) we may associate a specific performance P_j, as measured by the amount of biological activity of j-th type, $j = 1, \ldots, k$. The j-th task is performed at its best in the environment characterized by $\rho^{(j)}$, i.e. P_j is maximal at a specific value of ρ (e.g. transport is better carried out where there is a high concentration of chemicals that need to be transported from one membrane side to the other). The environment with $\rho = \rho^{(j)}$ will be called the j-th environment. As a consequence, the performances are in trade-off, since the k environments where each of them can be maxized are mutually exclusive (one could also assume that the environments are $k' < k$, since more than one performance can be maximal in the same environment).

(*iii*) The relevant traits are represented by a vector ν that targets the protein to the environment characterized by $\rho(\nu)$, in such a way that its *biological function* is maximally exploited. Thus the j-th performance is assumed to be a function of $\rho(\nu)$, $P_j(\rho(\nu))$.

(*iv*) The biological function of a protein is quantified by its *fitness* function, as follows:

$$F\left(P_1(\rho(\nu)), \ldots, P_k(\rho(\nu))\right) . \tag{3.1}$$

F is assumed to be an increasing function of all its arguments. According to (*iii*), we must maximize F with respect to ν in order to find where the protein characterized by F will be directed. The derivative of F with respect to the traits ν leads to the optimal solutions:

$$0 = \frac{\partial F}{\partial \nu_m} = \sum_{j=1}^{k} \frac{\partial F}{\partial P_j} \frac{\partial P_j(\rho)}{\partial \nu_m} . \tag{3.2}$$

From (*ii*) $P_j(\rho)$ is maximum at $\rho = \rho^{(j)}$. We make the simplifying hypothesis that $\rho^{(j)} \equiv \rho(\nu^{(j)})$ and, at the leading order in $\rho - \rho^{(j)}$,

$$P_j(\rho) = P_j(\rho^{(j)}) - (\rho - \rho^{(j)})^T g(\rho - \rho^{(j)}) , \qquad (3.3)$$

where g is some metric tensor and, at the leading order in $\nu - \nu^{(j)}$,

$$\rho(\nu) - \rho(\nu^{(j)}) = M(\nu - \nu^{(j)}) , \qquad (3.4)$$

with $M_{i,m} = (\partial \rho_i(\nu)/\partial \nu_m)_{\nu=\nu^{(j)}}$, independent of j. This leads to

$$0 = \sum_{j=1}^{k} \frac{\partial F}{\partial P_j} \hat{g}(\nu - \nu^{(j)}) , \qquad (3.5)$$

where $\hat{g} = M^T g M$ is the induced metric tensor in trait space. Thus, we are led to the condition for the optimal choice of ν,

$$\nu = \frac{\sum_{j=1}^{k} \nu^{(j)} \partial F/\partial P_j}{\sum_{j=1}^{k} \partial F/\partial P_j} , \qquad (3.6)$$

which means that the optimal ν lies in the convex hull in ν-space whose vertex are $\nu^{(j)}$, $j = 1, \ldots, k$. We then expect that a convex hull in the trait subspace is a signature of a Pareto optimization in the *E.coli* proteome.

3.3 Enrichment analysis

Each archetype/vertex must be enriched with at least one discrete or continuous feature characterizing the corresponding archetype. Density profiles of the features enriching a given vertex must attain their maximum value in the region (or bin) of the polytope containing that vertex, and then decrease monotonically with the distance from it. From enrichment analysis, Pareto optimality theory allows us to infer competing tasks for each vertex of the polytope (three tasks in our triangular case) from the attributes of the corresponding enriched features (continuous or discrete).

3.3.1 Enrichment analysis with continuous and discrete features

We performed enrichment analysis on discrete features, such as the subcellular localization annotations (6 annotations), obtained from the Taguchi's dataset, and the GO-annotations (702 annotations). GO-annotations were

obtained from the Gene Ontology dataset [3] which has the structure of a directed acyclic graph with nodes, called GO terms, which describe the molecular functions of each protein, their locations in the cell environment and the biological processes in which they are involved. Below, we will show how to build the complete table of discrete features for the enrichment analysis.

We treated the discrete features on the same footing as the continuous features, by assigning to data points the value 1 if they hold a given feature and 0 otherwise. For each vertex we associate a ranked vector of euclidean distances ordered from the nearest point to the furthest from the vertex. Data points are then clustered in bins, such that each bin has the same number of points. We compute the ratio of densities of the discrete feature in a given bin, with respect to the mean density among all data. The results, plotted versus the bin number (ordered from the nearest to the farthest from the archetype), are shown in Figure 3.10.

3.3.2   Statistical significance of enriched features

The statistical significance of the enriched features can be evaluated by computing a p-value test, based on the probability of finding a higher density of the feature in the first bin with respect to the other bins (see Supplementary Materials of [132]). We analyzed a large dataset of 708 discrete features. With such a big number, several enriched curves could appear just by chance. Thus, the p-values must be corrected for the possibility of "false-positive" p-values. A common approach employed to deal with these type of errors is the false discovery rate (FDR) [8].

The statistical significance of enriched features was tested also against the null-model, by reshuffling the values of a given feature. It is expected that only a few enrichments survive after a random reshuffling. For 10^3 random datasets, with 708 randomized features each, we found that only 50 out of 10^6 NULL-features are enriched by chance, with a threshold of 0,05 for the FDR.
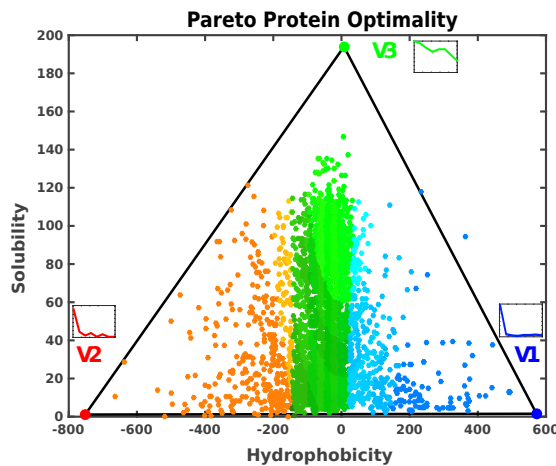
Fig. 3.8 Solubility-hydrophobicity triangle. We show a scatter plot of the 3,172 proteins of the *Escherichia coli* proteome. Each protein is represented as a point whose coordinates are the values of its hydrophobicity and solubility. The Pareto front is the triangular-hull that exhibits a low p-value of the order of $5 \cdot 10^{-3}$, confirming the statistical significance of the plotted distribution. Proteins whose points lie inside the triangle are the best compromise in the multi-objective optimization of the three tasks, which are better performed by the corresponding archetypes located at the three vertices. Points outside the triangle would have a better counterpart inside the triangle in at least one of the tasks. The RGB colors identify the distribution of the integral inner membrane (blue), outer membrane, and outer membrane bounded periplasmic (red) and cytoplasmic (green) proteins, which also characterize the vertices.

Table S 3.2 Principal components and their relative weights

| Table Of Loadings	PC1	PC2	PC3
Hydrophobicity	**0.9996**	0.0002	0.0275
Solubility	-0.0040	**0.9999**	0.1409
Yield	-0.027193	-0.1410	**0.9896**
Calculated pI	0.0037	-0.0069	-0.0095

Table S 3.3 Inferred tasks for each archetype in the *Escherichia coli* proteome, along with subcellular localization labels.

Archetype (Vertex)	Inferred tasks	Subcellular localization	Enriched GO-a
Blue (V1)	Transporting	Integral Membrane	Cation transmembrane t Active transmembrane t Anion transmembran
Red (V2)	Polysaccharyde Binding Catalysis	Outer Membrane and outer membrane bounded periplasmic	Po Polysaccharide metabo Hydrol Molecular functio
Green (V3)	Regulation	Cytoplasm	Regulation of the metab Regulation of biologi

3.3.3   Sub-cellular Localization Annotations

The process of targeting proteins towards the correct cellular compartments seems critical in the functionallity of prokaryotes and eukaryotes. Here, we are looking for optimization criteria which drive the localization of proteins inside the cells. As pointed out in the above section, Pareto optimization requires enriched features at the archetypes, so that we consider as discrete features the sub-cellular localization annotations as given by Taguchi [95]. Each protein is labelled with one out of eleven possible cellular component features: periplasmic, cytoplasmic, inner membrane, outer membrane beta barrel (see figures 1 and 3 in the main text), membrane anchored, inner membrane lipoprotein, outer membrane lipoprotein, membrane lipoprotein, membrane associated, perisplasmic with N-terminal Membrane Anchored and extracellular proteins. We selected for further analysis only the six features with an occurrence frequency higher than 15: periplasmic, cytoplasmic, inner membrane, outer membrane (see Figure 3 in the main text), membrane anchored, outer membrane lipoprotein.

We remind that in *Escherichia coli*, as in other gram-negative bacteria, the cytoplasm is surrounded by a multi-layered cell envelope that consists of the plasmatic or inner membrane, composed of a phospholipid bilayer, and a second external lipid bilayer, identified as the outer membrane. This second external membrane is asymmetric and has a different composition with respect to the inner membrane. Moreover, the outer membrane exposes lipopolysaccharide molecules to the external environment. The outer membrane, is the most protective barrier for the organism, and the lipidic layer, together with the outer membrane proteins and the lipopolysaccha-
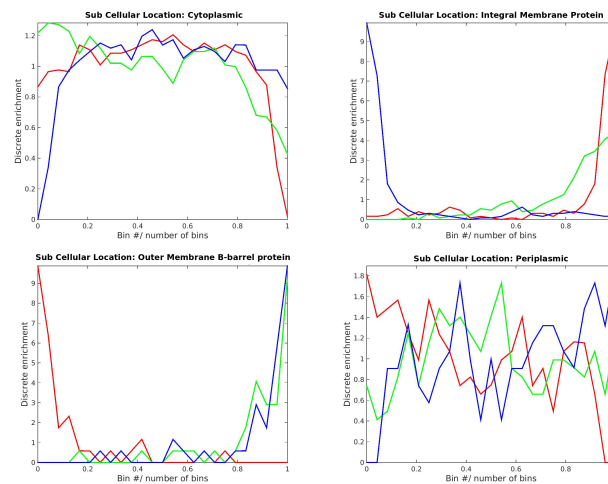
Fig. 3.9 Discrete enrichments of proteins annotated with sub-cellular compartmentalization. Data points are clustered in 25 bins with the same number of proteins according to their euclidean distance from one of the three archetypes. We booleanized the data (1 for proteins with the given feature, 0 otherwise) and for each of the 25 bins we computed the ratio between the fraction of proteins with the specified feature in the bin over the fraction with the same feature inside the whole triangle. This procedure is repeated for all the archetypes. The red and blue curves are almost specular since the triangle is approximately isosceles, with a slight shift toward the blue vertex.

ride, create the tactile organ of the gram-negative bacteria. Between the two membranes lies the periplasm, a crowded space that contains proteins, small molecules and a peptidoglycan mesh layer [99] The vertices with the lowest solubility values are mainly populated by membrane proteins (V1 and V2 in Table 3.3). Nonetheless, there is a clear-cut distinction between the two vertices. Vertex V1 has a very high hydrophobicity component, in the trait vector, and is enriched in inner membrane proteins (represented by blue points in Figure 3.8). Whereas vertex V2, which presents higher water-like propensity (i.e., low hydrophobicity), is enriched in outer-membrane and outer membrane bounded periplasmic proteins (red points in Figure 3.8). This sharp separation between membrane proteins (both with low solubilities) is striking, and it shows that the different values in their hydrophobicity component appear to be an essential ingredient in driving membrane proteins to their final destination. Vertex V3, which has a very high solubility, is enriched with proteins that occupy the cytoplasmic region (green points in Figure 3.8). Enrichment curves are rather smooth in the case of a small number of bins (5 − 10) while their roughness increase with a higher number of bins.

3.3.4   Enrichment analysis: GO annotations

The distribution of Gene Ontology annotations [3], considered as a function of the distance from the polytope vertices (the archetypes), unveils the competing tasks related to them. The Gene Ontology annotations of each protein hereafter referred to as GO-terms, are extended to include the parent GO-terms, to improve the robustness of protein annotations (see SI for further details). We consider the Gene Ontology dataset as given from http://geneontology.org, which consists on a total number of 4442 GO-terms. We booleanized this dataset by assigning to each protein the value 1, if they are annotated with the given term, and 0 otherwise. Then, we considered only those annotations with occurrencies higher than 15, resulting with a final table of 702 GO-terms. (Each protein can be annotated with more than one GO-term at the same time. We bin the space into equally populated regions [69, 143], and for any given annotation, we check whether the first bin is more enriched than the other bins. The statistical significance of the enriched terms is evaluated with a Benjamini-Hochberg procedure to take into account the problem of multiple hypothesis testing. Finally, the False Discovery Rate (FDR) with a threshold set to 0.05 is computed [8].

Based on this analysis, we find GO-annotations that are significantly enriched at each vertex. The vertex V1 (blue) is enriched in transmembrane transporters; in the vertex V2 (red) we observe enriched GO-terms for Porin activity, polysaccharide metabolic process, and hydrolase activity; the third vertex V3 (green) is enriched in molecular functions related to different kinds of regulation tasks. The enrichment densities of these features are shown in Figure 3.10 and listed in Table 3.3.

According to our mathematical derivation, the tasks found to enrich the triangle vertices are expected to be better performed in the distinct subcellular localizations that label the corresponding vertices. This finding is confirmed by the types of GO-terms, related to the molecular functions and biological processes, that enrich those vertices.
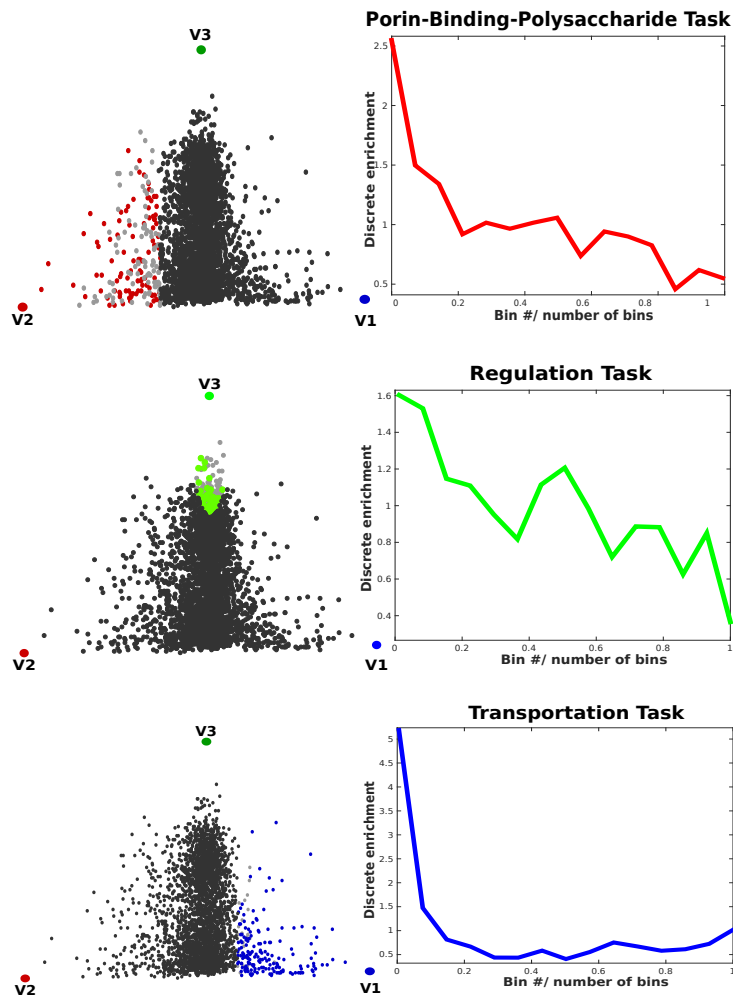
Fig. 3.10 Enrichments. Enrichment plots as a function of the distance from the corresponding archetype. Pareto optimality is defined such that the points closest to the vertices of the triangle must be maximally enriched in some features (they behave as specialists or "pure" types). All the tasks (GO-terms) that enrich each vertex are added together. Next to the enrichment plot, the proteins are mapped in the solubility-hydrophobicity plane. The colors highlight the enriched proteins belonging to the first bin. The vertices in the figures (V1, V2, and V3) label the protein subcellular localizations (as presented in Figure 3.8), namely, cytoplasmic proteins (green), integral inner membrane proteins (blue), outer membrane, and outer membrane bounded periplasmic proteins (red).

Archetype 1

The right vertex, the blue one, is enriched with inner membrane proteins, which are characterized by low solubility and high hydrophobicity. It is highly populated by proteins specialized in the *transportation process* such as: cation transmembrane transporter activity (GO:0008324), ion transport (GO:0006811), active transmembrane transporter activity (GO:0022804), ion transmembrane transport (GO:0034220),ion transmembrane transporter activity (GO:0015075), organic anion transport (GO:0015711), substrate-specific transmembrane transporter activity (GO:0022891). Further GO-terms that specify the inner membrane location are the following: single-organism transport (GO:0044765), intrinsic component of plasma membrane (GO:0031226), single-organism localization (GO:1902578), bacterial inner membrane (GO:0005886) (see Figures 3.11 and 3.12).
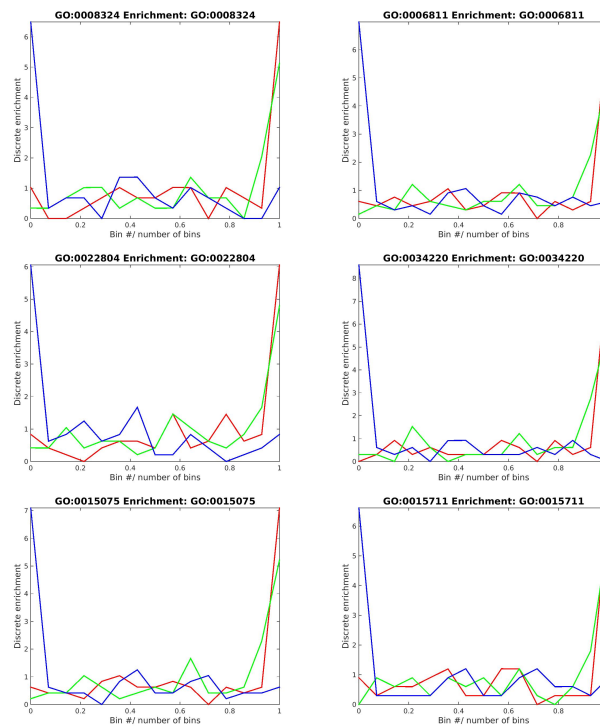
Fig. 3.11 Right Vertex Density enrichments are shown in the case of 15 bins and FDR<0.05.
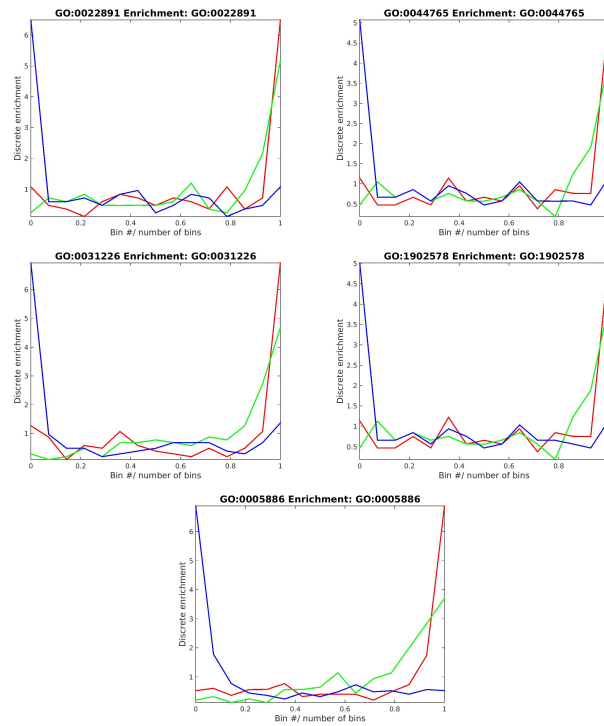
Fig. 3.12 Right Vertex Density enrichments are shown in the case of 15 bins and FDR<0.05.

Archetype 2

At the left vertex, the red one, we find outer-membrane and outer membrane-bounded periplasmic proteins, which are characterized by low solubility and low hydrophobicity. In this vertex, proteins are specialized in *wide-pore forming* from the intake of molecules, *catalysis, binding activity and polysaccharide metabolic processes*. The enriched GO-terms are the following: elemental activities, such as catalysis or binding (GO:0003674), polysaccharide metabolic process (GO:0005976), macromolecule catabolic process (GO:0009057), hydrolase activity (GO:0016787), external membrane of Gram-negative bacteria (GO:0019867), outer membrane-bounded periplasmic space (GO:0030288), cellular polysaccharide metabolic process (GO:0044264), external encapsulating structure part (GO:0044462), 4 iron, 4 sulfur cluster binding (GO:0051539) (see Figures 3.13 and 3.14).

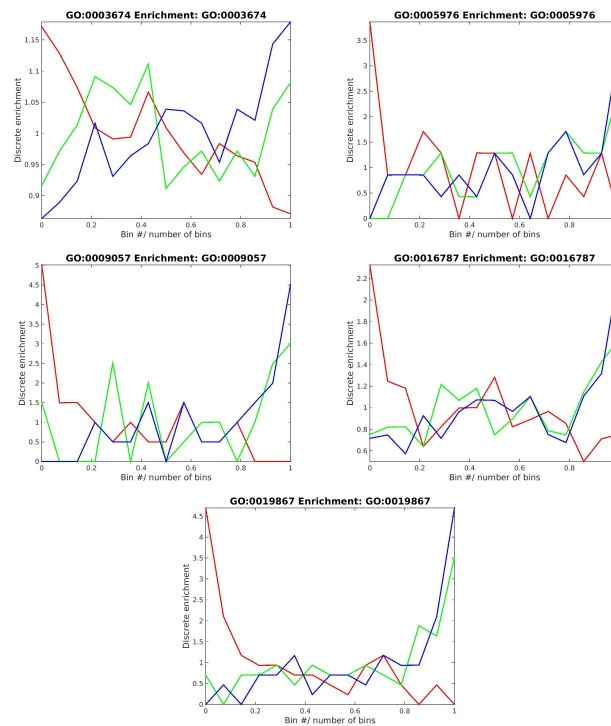**Fig. 3.13 Left Vertex** Density enrichments are shown in the case of 15 bins and FDR<0.05.

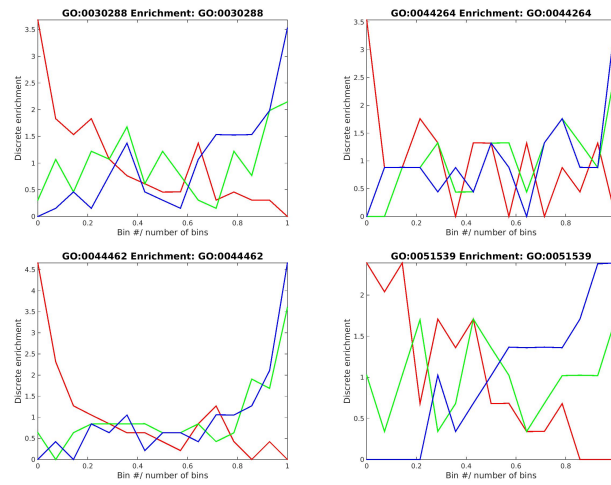**Fig. 3.14 Left Vertex** Density enrichments are shown in the case of 15 bins and FDR<0.05.

Archetype 3

As seen in the section above, cytoplasmic proteins, which are characterized by high solubility and low hydrophobicity, cluster at the top vertex. These proteins are specialized in *regulation processes*, as derived from the enrichment analysis of the GO terms. In the figure 9 below we have examples of enriched regulation processes, such as: regulation of biological processes (GO:0050789), regulation of metabolic processes (GO:0019222), biological regulation (GO:0065007) and regulation of primary metabolic processes (GO:0080090). The cytoplasmic characteristic of these proteins is supported also by the cellular component *cytosol component* (GO:0044445), see Figure 3.15.
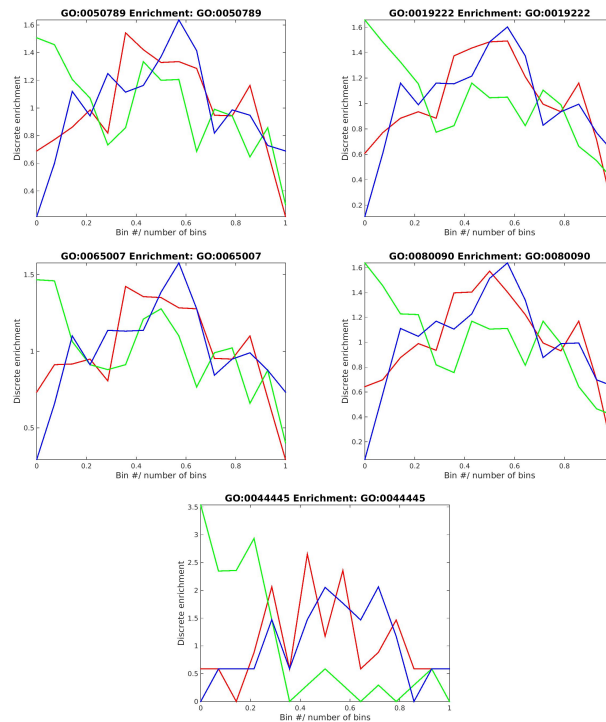
Fig. 3.15 Top Vertex Density enrichments are shown in the case of 15 bins and FDR<0.05.

We end this section by introducing three generic GO-labels which are useful to group the archetypal GO-annotations with each other in three main classes. Enrichment analysis performed on this new labels can be considered as an average analysis of the archetypal annotations.

GO-annotations associated to Archetype 1, (GO:0005886, GO:0008324, GO:0006811, GO:0022804, GO:0034220, GO:0015075, GO:0015711, GO:0022891, GO:0031226, GO:0044765, GO:1902578), are thus relabelled as "transportation", those associated to Archetype 2 (the red one), (GO:0003674, GO:0005976, GO:0009057, GO:0016787, GO:0019867, GO:0030288, GO:0044264, GO:0044462, GO:0051539), are relabelled as "porin-binding-polyssaccharyde", while those associated to Archetype 3 (the green one), (GO:0050789, GO:0019222, GO:0044445, GO:0065007, GO:0080090), are thus relabelled as "regulation". In the Figures 3.16 and 3.17 below, we plot the displacement of the proteins pertaining to the three classes:
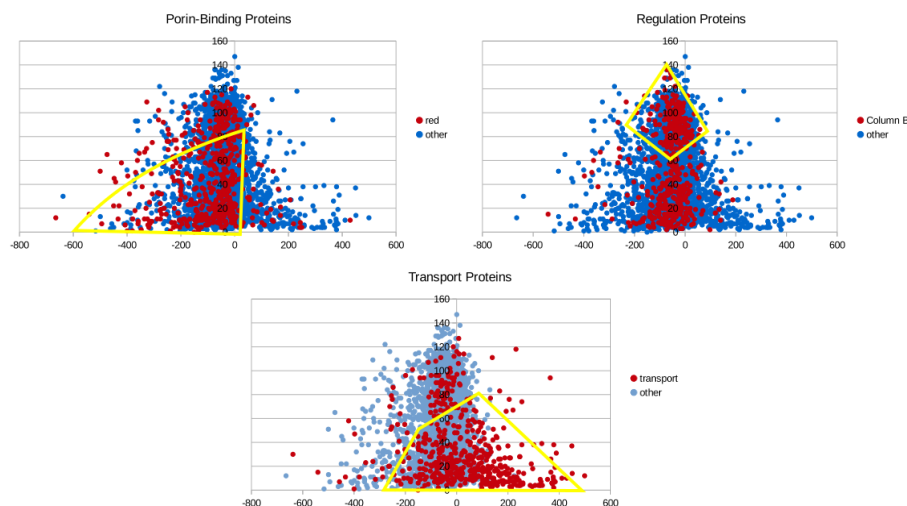
Fig. 3.16 Density of the archetypal feature Proteins labelled with regulation proteins, porin-binding-polyssaccharyde, transport proteins are plotted in the space of solubility vs hydrophobicity. We enclose with a yellow convex hull the specialized proteins.
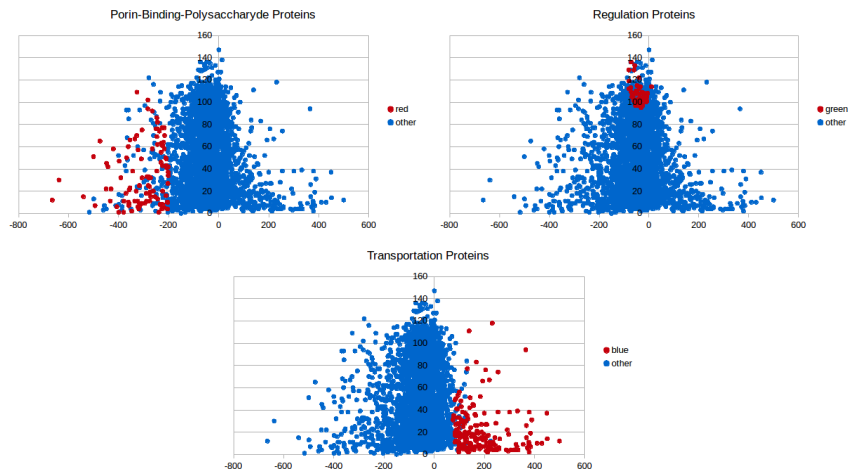
Fig. 3.17 Archetypal proteins in the 1st bin. Red points denote the proteins with the given feature in the bin nearest each vertex ($\approx$ 200 proteins).

Enrichment analysis performed on the three archetypal groups is shown below in the Figure 3.18:
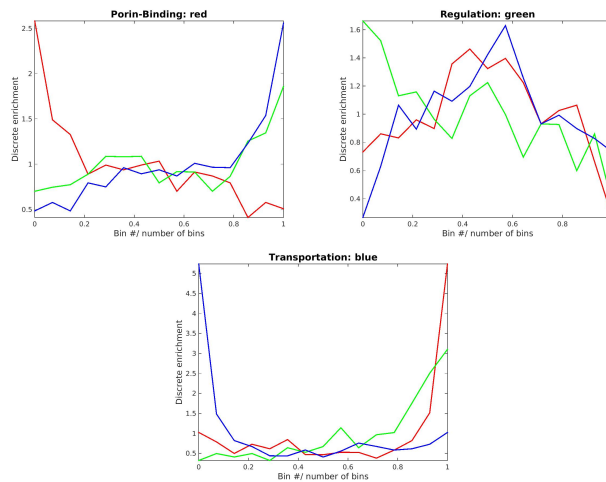
Fig. 3.18 Enrichment analysis of the three main groups We binned the dataset into 15 bins. In panel a) porin-binding-polyssaccharyde proteins, b) regulation proteins, c) transportation proteins.

Statistical fluctuations increase with the number of bins. In the case of 25 bins the three archetypal groups have the following enrichment patterns:
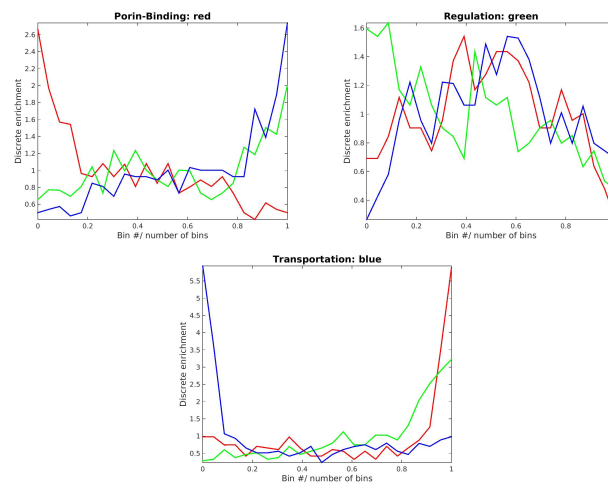
Fig. 3.19 Enrichment analysis of the three main groups We binned the dataset into 25 bins.

3.4 Evidence for a tetrahedron

If protein yield is added as a third trait to the Pareto front analysis, a statistically significant tetrahedron emerges as the convex hull enclosing all data. The tetrahedron base, in the hydrophobicity-solubility plane at the low yield, reproduces the already discussed triangle with vertices V1, V2 and V3 corresponding to different cellular compartments. The fourth tetrahedron vertex, V4, at high yield, is inferred to be related to archetypal proteins that are cytoplasmatic (as for vertex V3) but involved explicitly in tRNA/RNA metabolic processes. The above conclusion needs to be further validated, because of the low number of proteins found close to V4. The finding that proteins highly expressed by a cell-free translation system [95], based on translation factors, tRNAs and ribosomes, with no chaperons involved, can be associated to Pareto optimality through their functional role in tRNA/RNA metabolic processes is intriguing. In keeping with the general framework established in this work, whereby different tasks are associated with different environments, the presence of RNA molecules may be interpreted as defining a specific type of environment for the archetypal V4 protein.

When the Pareto analysis is extended to include protein yield, a tetrahedron emerges as the convex hull representing the new front in $3D$ (Figure 3.20). The yield feature, as derived from the Taguchi's dataset, corresponds to the third principal component (see Table 3.2). The tetrahedron encloses most of the data points, with a p-value smaller than 0.01%. Based on the Pareto theory, all the vertices of the tetrahedron must be enriched with at

least one feature per vertex, in order to infer the competing tasks for all the vertices. The triangular convex hull discussed above can be obtained from the tetrahedron by projecting it on the solubility-hydrophobicity plane, so that the enriched features found for triangle vertices can be associated to three of the tetrahedron vertices as well. The new vertex, V4, is characterized by proteins with a high yield component, low hydrophobicity, and low solubility. This vertex, similar to vertex V3, is enriched with cytoplasmic proteins; however, the tasks that characterize vertex V4 are different. According to our GO-terms analysis (see Figure 3.21), they are related to RNA processes such as tRNA metabolic process (GO:0006399), tRNA modification (GO:0006400 and GO:0009451) and ncRNA metabolic process (GO:0034660). This finding indicates that proteins involved in tRNA/RNA metabolic processes are also the ones that have higher expression levels in a cell-free translation system. However, in contrast to the two-dimensional triangular Pareto front, the found tetrahedron is not robust. When few data points with the highest yields are removed, the p-value increases from 10^{-4} to 10^{-1}, making the results of this analysis less reliable.
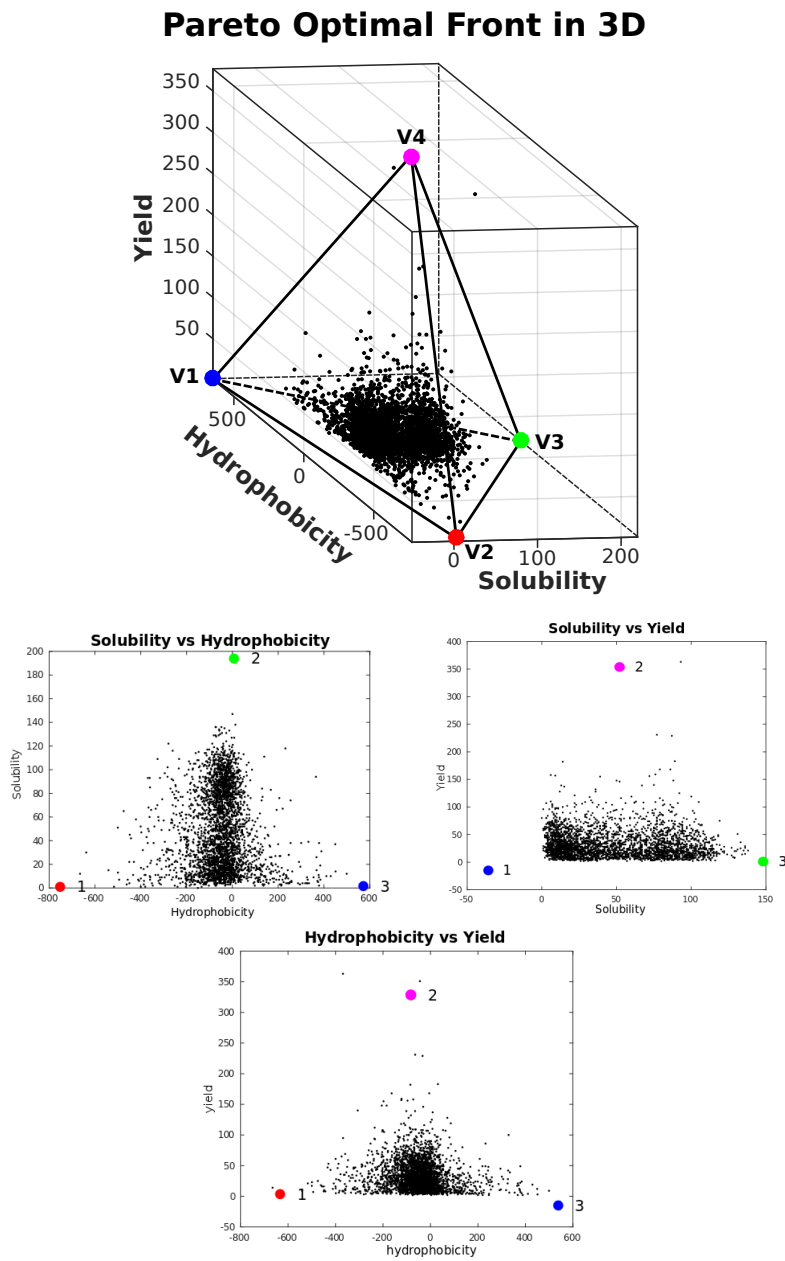
Fig. 3.20 Tetrahedron projections Tetrahedron in the hydrophobicity-solubility-yield space. The three vertices in the hydrophobicity-solubility plane, correspond to the archetypes identified in the previous section.

Arch (Orig) Position	Hydrophobicity	Solubility	Yield
Red	-755.26	-2.62	5.57
Purple	-128.65	48.15	378.9
Green	13.31	211.37	-0.09
Blue	636.04	-34.12	1.84

Arch (PCA) Position	Hydrophobicity (PC1)	Solubility (PC2)	Yield (PC3)
Red	-703.60	-47.41	-53.32
Purple	-87.59	-49.51	340.54
Green	63.94	165.46	-7.68
Blue	687.40	-77.69	-23.15

Table S 3.4 Coordinates of the four archetypes as found with Sisal. The coordinates of the four vertices in the solubility-hydrophobicity-yield space are shown in the top table, whereas the coordinates in the principal component space are shown in the bottom table.
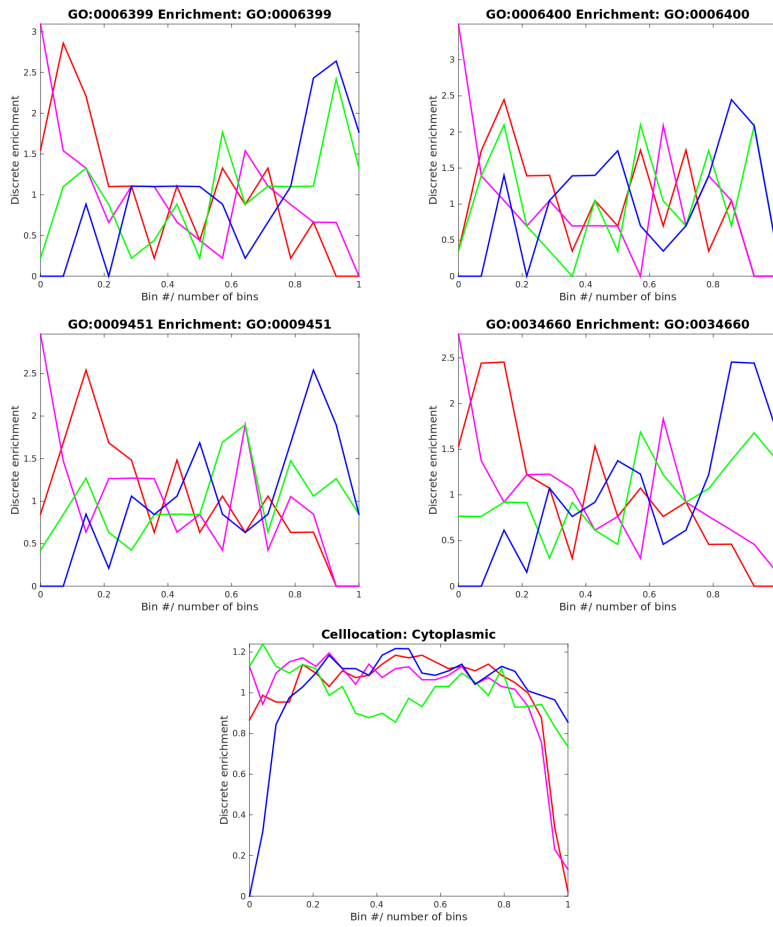
Fig. 3.21 Fourth Vertex enrichments Density enrichments are shown in the case of 15 bins and FDR<0.05. We show the subcellular location in the case of 25 bins.

3.5 Conclusion and Discussion

In this chapter, we extended the Pareto front analysis to the molecular level. We found evidence that *Escherichia coli* proteins were selected by trading off the performances of different competing tasks, and we inferred the latter ones. According to the Pareto interpretation, we suggest that *E.coli* seems to exploit solubility and hydrophobicity signals to drive the proteins in the cell compartments where they perform the required biological functions at their best. Finally, in the specific case of membrane proteins, which inherently have very low solubilities, our analysis can split apart outer and inner membrane proteins, using their different hydrophobicities.

According to the standard view, the basic physical properties considered here, hydrophobicity and solubility, were evolved in the first place to allow the foldability of proteins and to prevent them from aggregation. On top of that, our findings suggest the novel idea that the solubility-hydrophobicity signal encoded in the protein sequence can flag the final localization of the latter in the cell, and at the same time can hint at its biological function. According to the Pareto interpretation, the two traits have evolved to optimize three different performances simultaneously, each related to a separate cellular compartment.

Thus, the major result of our study is the crucial role played by subcellular compartments in the fitness of the *Escherichia coli* proteome, obtained by a direct mapping between the Pareto front vertices and the subcellular compartments (Figure 3.10, 3.22). It turns out that natural selection pushed the bacterium to optimality by tuning the solubility-hydrophobicity traits of all proteins, in such a way that each of them can reach the distinct environment where it can perform the required task at its best. On the other hand, protein biological tasks are eventually related to their interactions with metal ions, ligands, substrates, other proteins, or nucleic acids. Therefore, one could speculate that the specific solubility-hydrophobicity traits of each protein are needed to optimize the interactions associated with the related biological tasks.

The Pareto analysis shows that the protein performances are in a trade-off with each other and identifies archetypal tasks located closer to polytope vertices. From that, we can infer that the archetypal proteins found at vertex V1 of Figure 3.8(inner membrane) are specialized in the transport of organic and inorganic molecules. Archetypal proteins at vertex V2 (outer membrane and periplasmic space) are specialized in wide-pore forming from the intake of molecules, catalysis, binding activity and polysaccharide metabolic processes, while those at vertex V3 (cytoplasmic space) are

specialized in the regulation of different processes (Table 3.3). As noted before, the difference in solubility can be due to different structural classes [95]. Nonetheless, we found that membrane proteins, which have very low solubilities (also confirmed by experimental data [95]), can be split into outer and inner membranes through their hydrophobicities. Notably, the two membrane protein classes have very different structures, in spite of the fact that their measured solubilities are similar.

The problem of spatial protein distribution in bacteria is of paramount importance since the subcellular localization of proteins is crucial to provide the physiological context for their function, to achieve functional diversity and to economize protein design and synthesis [18]. Although bacterial cells (such as *E. coli*) lack internal membrane-bounded structures, they are not "bags of mostly randomly localized macromolecules" [59]. Instead, they are organized with different macromolecules that display complex subcellular localization patterns [18, 32, 120]. Different mechanisms drive proteins toward their final cell destination [18, 32, 120] through the cytoplasm and the subcellular localization of proteins in *E. coli* across the different membrane barriers, and one of the major achievements that our analysis offers is a significant breakthrough for the comprehension of this transport mechanism. With the Pareto front analysis, we find indications that Gram-negative bacteria exploit the solubility and the hydrophobicity of proteins to take them in the major compartments where they can perform the function needed for the organism at their best. This finding does not exhaust the complexity of the protein sorting, but it adds new clues. Among all known mechanisms and signals, the solubility-hydrophobicity balance of a protein could be exploited by the cell as a subcellular localization signal. According to our results, it appears that solubility and hydrophobicity values provide a signature to the protein's final destiny, and possibly an indication of the task that proteins perform at their best in that environment. This result, which was obtained from our Pareto analysis, should be experimentally validated in future research.
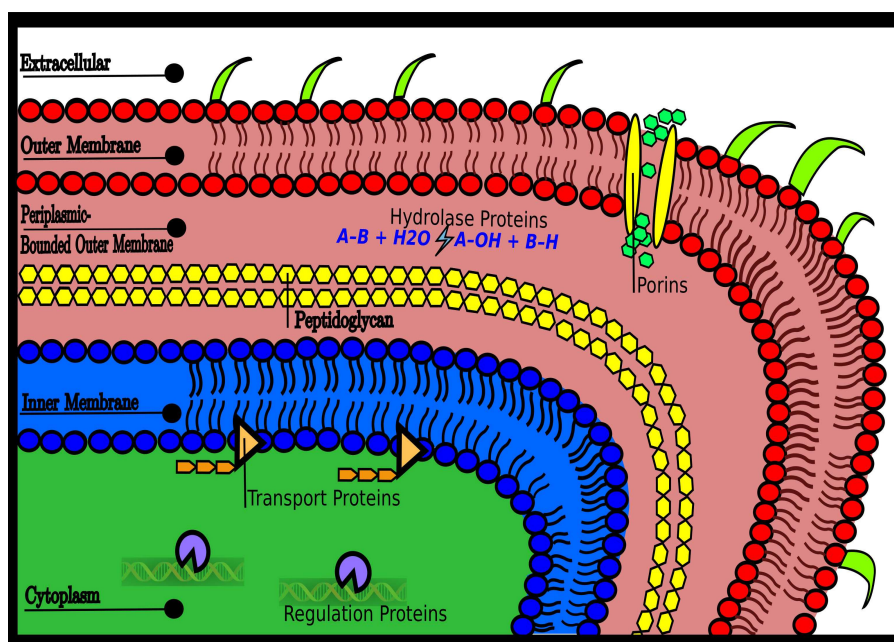
Fig. 3.22 Cell compartments and Pareto triangle. There is a direct mapping between the four different compartments of *Escherichia coli* (outer membrane and outer membrane bounded periplasmic proteins, inner membrane, and cytoplasm) and the proteins that populate the vertices of the Pareto front.

Chapter 4

Archetypes of human cognition defined by time preference for reward and their brain correlates: an evolutionary trade-off approach

4.1 Introduction

Biological systems carry out multiple tasks in their lifetime, which, in the course of evolution, may lead to trade-offs. In fact phenotypes (different species, individuals within a species, circuits, bacteria, proteins, etc.) cannot be optimal at all tasks, and, according to Pareto optimality theory, lay into a well-defined geometrical distribution (polygons and/or polyhedrons) in the space of traits. The vertices of this distribution contain archetypes, namely phenotypes that are specialists at one of the tasks, whereas phenotypes inside the geometrical distribution generalists.

In this chapter we test the predictions of Pareto optimality theory to human cognition and behavior by analyzing data from the Human Connectome Project (HCP) that includes a wealth of cognitive, personality, health, socio-economic status, and brain measures ([150], see also section 4.2).

The trade-offs in cognitive tasks are not a given. In fact, the well established theory of general intelligence, or g-factor, posits a positive correlation among a large number of cognitive tasks ([137]). While human intelligence may embrace more than sixty specific cognitive abilities, the g factor is common to all of them ([21]; [29]), explaining large amount of

variance (45–50%) across test scores in large samples of healthy subjects ([4]; [46]).

We asked if neuropsychological or behavioral scores distribute according to Pareto Optimality theory and focused on triangular shaped distribution. In section 4.4 we show that among all possible combinations of pairs of cognitive and behavioral traits of the dataset, the best fit to Pareto optimality is found when individuals were plotted in the trait-space of time preferences for reward, evaluated with the Delay Discounting Task (DDT). As we will exhaustively introduce in section 4.4.2, the DDT measures subjects' preference in choosing either immediate smaller rewards or delayed larger rewards. Time preference for reward was described by a triangular distribution in which each of the three vertices included individuals who used a particular strategy to discount reward.

These archetypes accounted for variability on many cognitive, personality, and socioeconomic status variables, as well as differences in brain structure and functional connectivity, with only a weak influence of genetics. Based on this enrichment analysis, we inferred the competing human evolutionary strategies. Furthermore, we identified differences among archetypes in brain structure (volume, gray matter, etc.), and function (resting state functional magnetic resonance imaging rs-fMRI connectivity). Finally, we explored the influence of genetics on archetype variability. Specifically, we asked if behavioral scores on the identified tasks were more concordant in monozygotic versus dizygotic twin pairs. In summary, time preference for reward reflects a core variable that biases human phenotypes via natural and cultural selection.

4.2 HCP Dataset

We analyzed the public data release of the WU-Minn Human Connectome Project (HCP) consortium ([150]), which includes 1206 healthy young adults, from families with both twins and non-twin siblings. The current sample was obtained from the March 2017 data release (1200 Participants; http://www.humanconnectome.org). The database consists of behavioural measures (e.g., cognitive, personality), socio-demographic measures, and high-resolution 3T MRI imaging data. Some data are restricted due to subject privacy (e.g. twin or smoking status etc). The HCP subjects include 168 Monozygotic twin pairs, and 103 Dizygotic twin pairs. The behavioral database consists of tests that are part of the NIH Toolbox battery and of several Non-Toolbox behavioral measures (see below). For each subject, we also obtained the brain volumes from

the Freesurfer software and analysed them by voxel-based morphometry. They consist of continuous features and are normalized with respect to intracranial volume.

The behavioral database consists of tests that are part of the NIH Toolbox battery and of several Non-Toolbox behavioral measures. They are collected in the following main domains:

1) **Demographics**: Gender, Age by Year, Race, Ethnicity, Handedness, Self-Reported demographics on education, income, relationship status from SSAGA.

2) **Health and Family History**: Body Mass Index, Blood Pressure, Parental Psychiatric or Neurological Illnesses.

3) **Alertness**: Cognitive Status, Sleep

4) **Cognition**: Episodic memory (Picture sequence and Verbal), Executive Function (Cognitive Flexibility and Inhibition), Fluid Intelligence, Language (Reading decoding and Vocabulary comprehension), Processing Speed, Self-regulation/Impulsivity (Delay Discounting), Spatial Orientation, Sustained Attention, Working Memory.

5) **Emotion**: Emotion recognition, Psychological Well-being, Social Relationships, Stress and Self-Efficacy.

6) **Motor**: Endurance, Locomotion, Dexterity, Strength.

7) **Personality**: Five Factor Model (NEO-FFI).

8) **Psychiatric and Life Function**: Achenbach Self-Report of Life function and Psychiatric Clinical Symptoms, Self-reported Psychiatric Clinical Symptom measures from SSAGA.

9) **Sensory**: Audition, Olfaction, Pain, Taste, Contrast Sensitivity, Color Vision, Visual Acuity.

10) **Substance Use**: Urine Drug Screen, Seven-day Alcohol and Tobacco Use Retrospective, Self-Reported Substance Use and Abuse measures from SSAGA.

4.3   Pareto Optimality Inference method

The Pareto Optimality analysis is based on the assumptions presented in chapter 3, where, instead of dealing with the proteins of the *Escherichia coli* bacteria, we considered each subject as a data point in the morphospace of the set of continuous traits ν, which correspond to measures of cognitive, personality, socio-demographic, and brain features.

We focused on identifying the best-shaped polytope that encloses the data points in the multi-dimensional space of traits starting from a triangular Pareto front distribution ([13]). In principle, other polygons or

polyhedrons in higher dimensional space might exist, but, based on prior evolutionary studies ([132]; [50]; [142]; [147]), and our study [76] presented in chapter 3, the initial focus was on triangular solutions. Clearly more work is needed to investigate polyhedrons in higher dimensional morphospaces, however this study is consistent with the theory that cognitive traits, as many other phenotypes in nature, are in trade-off.

As compared with other classical clustering methods (k-means, Gaussian Mixture models, Latent Class Analysis), Pareto Optimality approach differs as it identifies the vertices (rather than centroids) of a distribution. Clustering and Pareto analysis are indeed both able to find centroids, but in a complementary way, since the former is sensible to local density inside the distribution, while Pareto is mainly sensitive to the external shape (the external perimeter) of distributions, also called convex hulls (for further comparisons between the Pareto method and clustering methods see [69]). Pareto analysis and enrichment analysis, as described below in this section, were run using the software package ParTI: (https://www.weizmann.ac.il/mcb/UriAlon/download/ParTI).

The first step in our analysis was projecting for each pair of behavioral measures the 1206 participants' data points in a two-dimensional space. We considered measures related to each cognitive and performance domain (e.g., fluid intelligence, memory, spatial orienting, self-regulation, strength, dexterity etc. (see section 4.2 for details on the measures). After removing redundant, ordinal measures or measures with too few observations, we considered a subset of 25 traits and we combined them in pairs of cognitive and performance-related traits, resulting in 300 possible combinations.

As a second step, we checked if the distribution of points obtained for each combination of pairs of traits fits a triangular shape. The statistical significance of each potential triangle was tested with the triangularity test (the t-ratio test (see section 3.1.1 in chapter 3)). To further assess the validity of a triangular Pareto distribution, we measured the fraction of variance accounted for (across subjects) as a function of the number of vertices (2 to 6) of the possible polygons (see Figure 4.2).

This chapter focus on the best triangle in the morphospace of traits of the HCP dataset. This triangle includes individual scores on two measures of the Delay Discounting Task (DDT). The DDT measures the tendency to opt either for immediate smaller rewards or delayed larger rewards ([63]; [74]). This task assumes that the subjective value of a reward (e.g., money) is increasingly discounted from its nominal amount as a function of the delay until reward reception. Discounting is a pervasive phenomenon in decision making shared by humans and animals ([103]). The DDT is a

sensitive measure of the ability to wait for a reward (time preference) as well as impulsivity and self-control processes ([75]; [91]). In the context of Pareto Optimality, the vertices of this triangle contain individuals that use different strategies to discount reward in time. Interestingly, these groups enriched on a variety of other cognitive, behavioral, socio-economic, and health features, and differed on measures of brain structure and function. However, genetic influence was modest. Therefore, strategies for discounting reward represent phenotypes that have developed under evolutionary and/or cultural pressures to adapt to our environment.

4.4   A Pareto front in the delay discounting space (DDT)

For each participant, we took into account 25 continuous measures of the HCP (i.e., cognitive and behavioural scores), mapping them into the multi-dimensional space of traits (i.e., morphospace). The best triangular Pareto front solution was found in a two dimensional space that contains, for each subject, the values associated with the Area-under-the-curve (AUC) for $200 and AUC for $40,000, two measures of the DDT (Figure 4.1). Indeed, among all possible pairwise combinations of traits, the triangle defined by the two measures of the DDT was the only one to survive the permutation test on triangularity (over 1000 permutations) corrected for False Discovery Rate (FDR) ($p < 10^{-4}$). The Principal Convex Hull-Archetypal analysis (PCHA) showed that the triangle was the best polygon to enclose all the data points among planes with 2-6 vertices. In fact, a triangle shape distribution ($n = 3$ vertices) explained the majority of variance ($> 99.5\%$ variance), and increasing the number of vertices did not improve the amount of variance accounted for (Figure 4.2).

4.4.1   Validation of Pareto Front Solution

Even though the triangularity test examines the statistical significance of the obtained Pareto front solution against a null distribution through per-mutation tests, we also ran additional validation analyses. In one analysis, we performed a split-half replication: we ran the Pareto analysis separately on two random independent smaller samples of the HCP data set (n = 559 and n = 560 subjects, respectively), taking into account all 300 possible combinations of pairs of the 25 traits. This was done to ensure that the Pareto Front solution obtained from Pareto Optimality Inference method
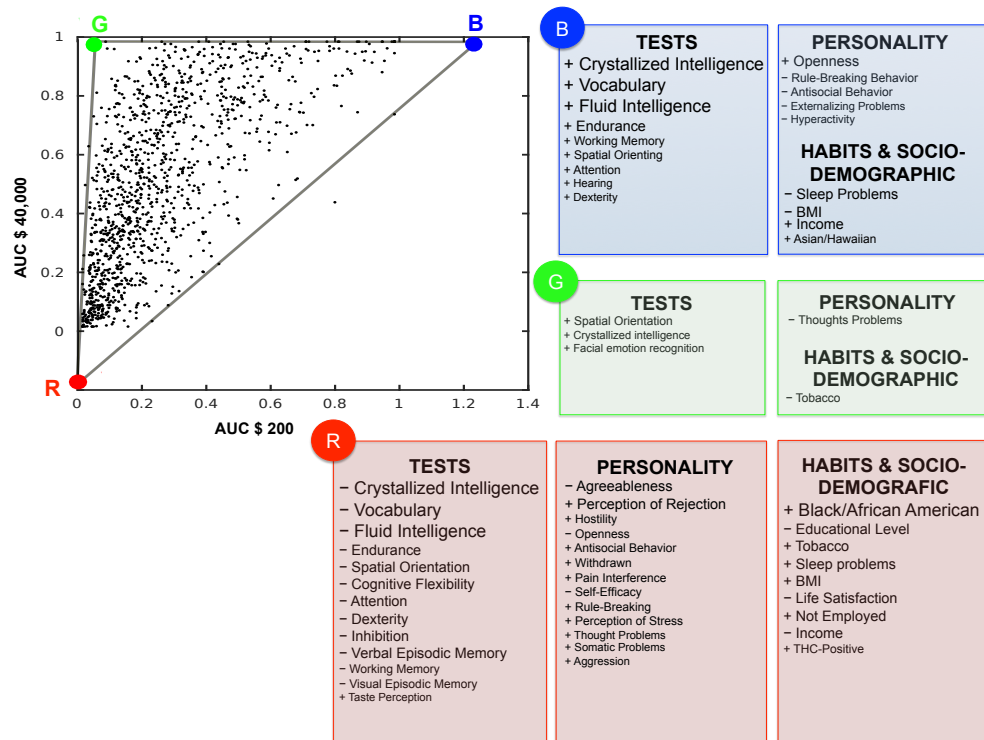
Fig. 4.1 Pareto distribution (triangular polytope) in a space of AUC $\$200$ **(x-axis) versus AUC** $\$40,000$ **(y-axis)..** The AUCs (Area-Under-the-Curve) are two measures of the Delay Discounting Task. The distribution of AUC scores is triangular hence fitting Pareto optimality theory. The three vertices of the triangle (labelled as Blue, Green, Red) contain individuals who adopt three different strategies for time preferences for reward (archetypes). These strategies co-vary with cognitive, sensory and physical abilities, personality traits, measures of substance use, and socio-demographic variables, which were identified by an enrichment analysis (see also Figure 4.13 and Table 4.8). The size of the font corresponds to the relative significance of each trait (larger font, lower p-value).
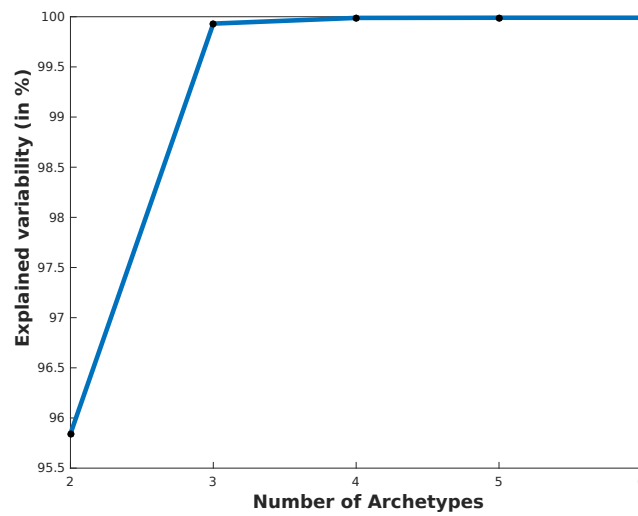
Fig. 4.2 Optimal number of vertices of the Pareto Front. This figure shows that three is the optimal number of vertices that explaines the largest amount of variance (in percentage) of the data point, which are plotted in the two dimensional space of AUC $200 and AUC $40,000. We made the analysis by varying the number of vertices from two to six. The vertices were found by using PCHA algorithm, as developed by [93]. The slope of the blue curve describes the increment of the explained variance as increasing the number of vertices. It results that three is a stationary point, after which the explained variance increases negligibly.
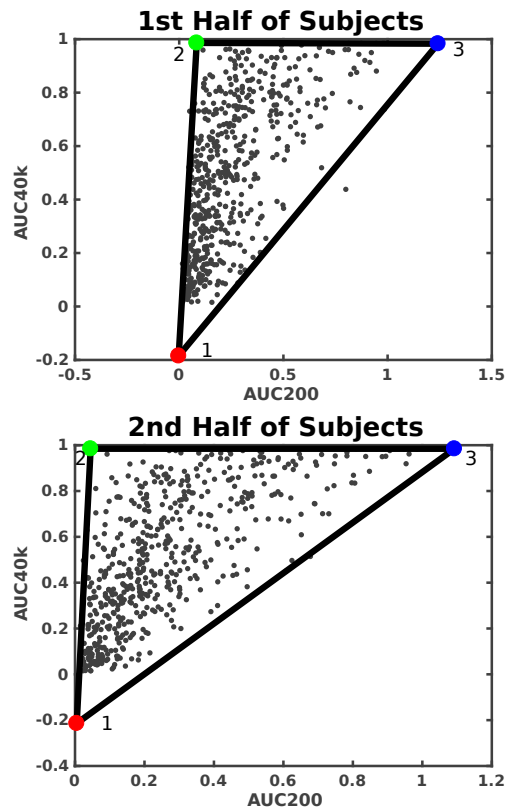
Fig. 4.3 Robustness of the Pareto Front-Test 1. This figure shows the robustness of our triangular front when we randomly split in two sub-samples of equal population the original sample of 1206 subjects and then we made for both the triangularity test. It results that the p-values are $< 10^{-4}$.

was robust, i.e. significant in two independent samples. The only significant triangle that emerged in both groups was that defined by the DDT measures (for both sub-samples: $p < 10^{-4}$, after FDR correction) (Figure 4.3).

We also asked whether the obtained Pareto front solution was robust to gender and race. In one analysis, two samples of subjects were created based on gender: Males (549 subjects) vs. Females (649 subjects) ($p < 10^{-4}$ independently for male and female subjects). In the second analysis, three groups of subjects were compared: Asian-Nat. Hawaiian-Other Pacific (n=67 subjects) vs. Black or African American (n=192 subjects) vs. White (n=883 subjects), with p-values such as: 1) $p = 5 \cdot 10^{-2}$ for Asian-Native Hawaiian or Other Pacific populations; $p = 10^{-4}$ for White subjects; $p = 0.2$ for Black or African American individual (Figure 4.4). In summary, the Pareto front for the DDT was highly significant, and robust over race, gender, and independent samples of subjects.
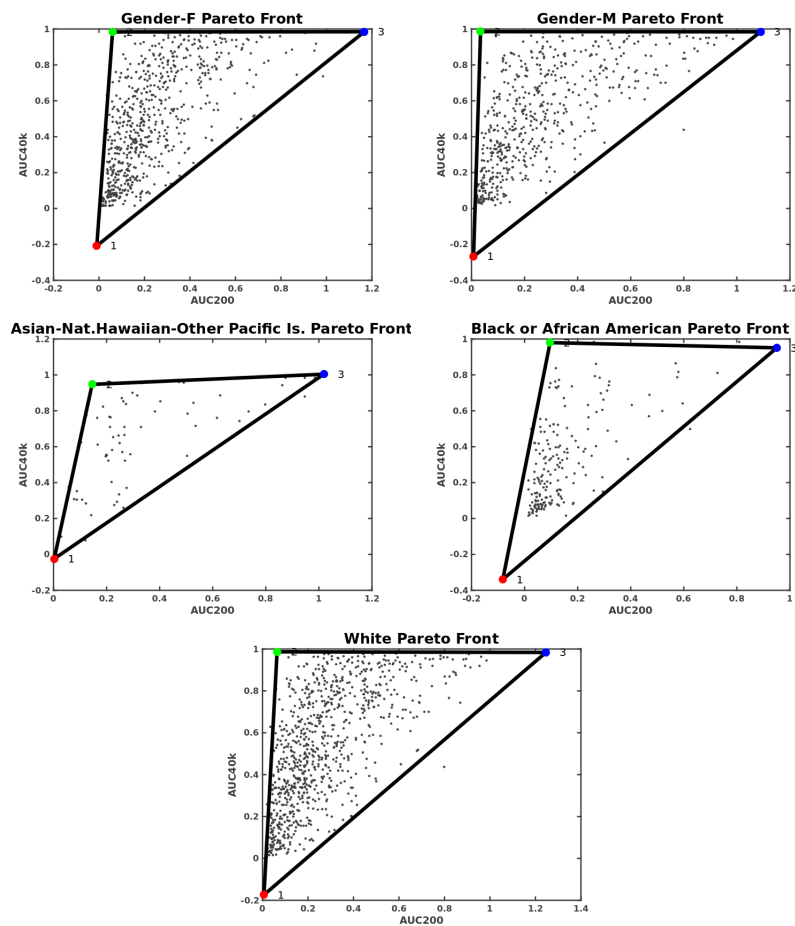
Fig. 4.4 Robustness of the Pareto Front-Test 2. This figure shows the robustness of our triangular front. We considered many sub-samples of the data points (1206 subjects) and made for each of them the triangularity test. We analyzed separately samples of only female/male subjects and the different race (Asian-Nat.Hawaiian-Other Pacific, Black or African American, White). It results that the triangular shape is robust to gender and race labels, meaning that the properties of the triangle are not related to them.

4.4.2   The Delay Discounting Task (DDT)

Since the best Pareto front solution was observed in the morphospace created by two measures of the Delay Discounting Task (DDT), here we briefly describe the DDT. All subsequent analyses (enrichment (in section 4.5), structural and functional brain features (in section 4.6), heritability (in section 4.7)) will be carried out on the distribution of data points derived from the combination of two measures of the DDT.

In the DDT, participants were asked to choose between two options on each trial: a smaller amount of money to be given immediately vs. a larger amount of money given at a later point in time. Participants made choices for each of 6 possible delays (1 month, 6 months, 1 year, 3 years, 5 years, and 10 years), and for two 'reference' delayed amounts that were kept constant ($200 and $40,000). The amount available immediately was instead adjusted after each choice in order to determine the amount judged subjectively as equivalent to the delayed amount. If the participant choose the immediate amount, then the immediate amount was reduced on the next trial, whereas if he/she choose the delayed amount, then the immediate amount was increased. For each combination of amount of delayed reward and time delay, participants were asked to make 5 choices, and the value that would have been used for the immediate amount in the 6th choice was used as the indifference point. The indifference point represents the point where an individual is equally likely to choose a smaller reward earlier (e.g., $50 immediately) versus a larger reward later on (e.g., $200 in 1 month). The Area under the curve (AUC) for each of the two reference amounts ($200 and $40,000) was computed based on the indifference points and ranges from 0 (maximum discounting) to 1 (no discounting) ([94]).

The AUC measures of the DDT are considered a reliable indicator of self-control in cases of lower discounting rate (i.e. preference for larger delayed rewards), and impulsive behavior in cases of higher discounting rate (i.e. preference for smaller earlier rewards)([75]; [91]). Although the rewards are hypothetical, there is a good correspondence with real rewards ([82]). Based on the processes involved in the DDT, the three vertices ('archetypes') of the Pareto front triangle identify three optimal strategies to deal with discounting reward in time.

4.5   Enrichment analysis of the Archetypes

According to the Pareto Optimality theory, the vertices of the triangle identify specialists that express different traits to the maximum (or minimum) extent, and that according to the theory are in trade-off. If Pareto theory is correct, then, other traits (i.e., enriched features) should be maximal or minimal in those specialists, and performance on those traits should decline (or rise) as a function of the distance from that archetype.

To identify traits that enrich, we first divided the distribution of individual scores in bins and then analyzed, for each trait, the change of the mean value of that trait across the bins of the polytope, normalized with respect to the mean value of the given trait for the whole distribution. For simplicity, we binned the Pareto front three times, each time starting from one of the three vertices, into n bins. To make the analysis statistically valid in terms of sample size, we constrained each bin to contain the same number of participants. This procedure was repeated systematically by varying the number of bins between 8 and 15. A higher number of bins leads to higher statistical fluctuations in the density analysis. Features could be discrete or continuous. For continuous variables, we computed the ratio among the mean value at all bins and the mean value of the entire triangle. We plotted this ratio as a function of the n-th bin. For discrete features, we first booleanized them (i.e. a value 1 was given if the participant had the given feature, 0 otherwise), then we treated them as continuous variables.

Enriched features were validated if they passed the p-value test based on the hyper-geometrical distribution ([69]) and corrected for FDR test. This test measures the probability that the mean value of a trait is maximal/minimal in the bin closest to a given vertex. The robustness of the enrichment was assessed by performing a null-test, namely a random permutation of the values of the traits among the different bins. Features belonging to four main domains were separately analyzed:

1. Cognitive, Physical and Sensory traits (1119 subjects and 46 measures);

2. Discrete traits of Personality, affective behaviour, substance abuse, socio-demographic features (1123 subjects, 40 measures);

3. Continuous traits of Personality, affective behaviour, substance abuse, socio-demographic background (1123 subjects, 70 continuous measures);

4. Structural brain measures (1105 subjects and 56 measures).

Structural brain measures (n=56) included volume of cortical gray matter, white matter, and volume of anatomical regions in the right and left hemisphere (e.g. right and left hippocampus, thalamus, etc..) segmented in Free Surfer. Before running the enrichment analysis, the measures were first normalized per intracranial volume.

The three vertices of the DDT triangle (see Figure 4.1), colored in Blue, Red, and Green, identify archetypes, namely 'specialists', i.e. subjects who adopt unique strategies to deal with the discounting task, while subjects in the middle of the triangle are 'generalists'.

In the following we will show that the Blue archetype corresponds to individuals with stable preference for larger rewards that are delayed in time, independently of the amount. The Red archetype identifies individuals with stable preferences for smaller immediate rewards. The Green archetype includes individuals who prefer delayed rewards when the amount is very large (i.e., $40,000$), but prefer taking sooner for smaller amounts (200).

4.5.1 Cognitive, Physical and Sensory traits

We carried out the enrichment analysis on 46 features reflecting cognitive, physical, and sensory abilities from 1119 participants, with a complete data set. We found that near the Blue archetype, several cognitive features enriched including crystallized and fluid intelligence, vocabulary knowledge, working memory, spatial orientation, and attention (Figures 4.1-4.13; Table 4.8; Figure 4.5-4.6).

For all these measures, individuals close to the Blue archetype showed the highest scores, hence they were superior in these domains. Also measures of sensory and physical abilities enriched near/at the Blue archetype, with those subjects showing the highest levels of hearing function, submaximal cardiovascular endurance, and manual dexterity.

When focusing on the Green archetype, individuals near this vertex scored high on measures of cognitive flexibility, crystallized intelligence and spatial orientation, and were fastest in recognizing facial emotions.

Finally, individuals closest to the Red archetype showed the lowest levels of performance on crystallized and fluid intelligence, vocabulary and spatial orientation, cognitive flexibility, attention and inhibition, working memory, verbal and visual episodic memory. These individuals also manifested the lowest performance on endurance and dexterity tasks. However, they scored highest on taste perception, i.e. they had a stronger perceived intensity to gustatory stimuli. Therefore, individuals near the Red archetype

showed an overall lower g factor. Notably, many of the cognitive, physical, sensory traits (excluding taste perception) reached a minimum near the Red archetype, and increased rapidly with distance from that archetype.
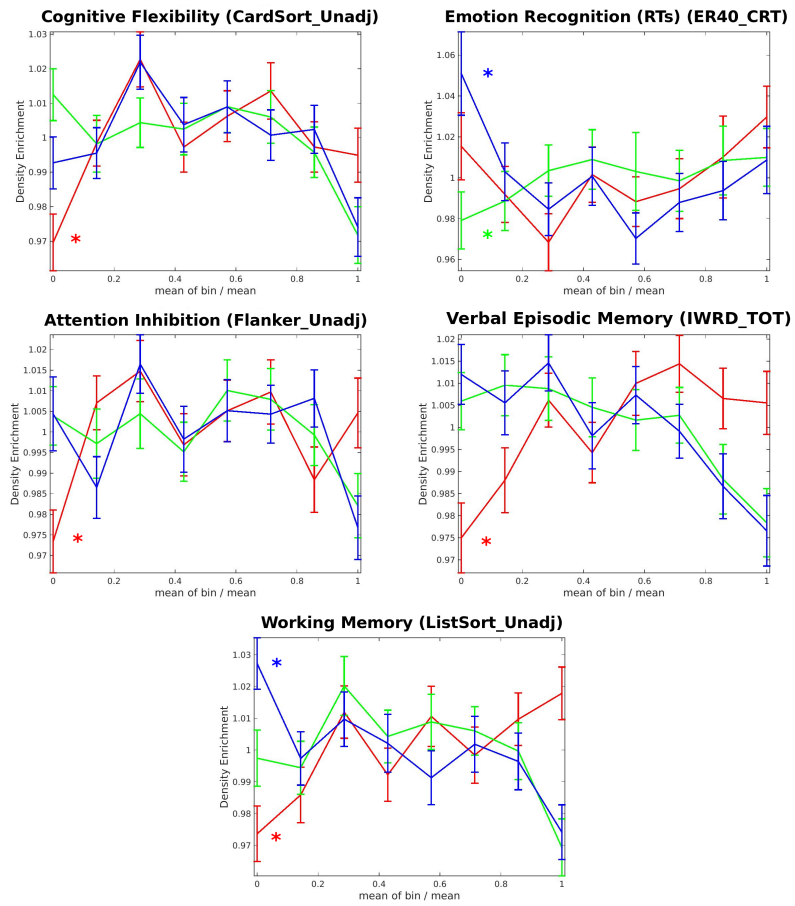
Cognitive

Fig. 4.5 Cognitive 1
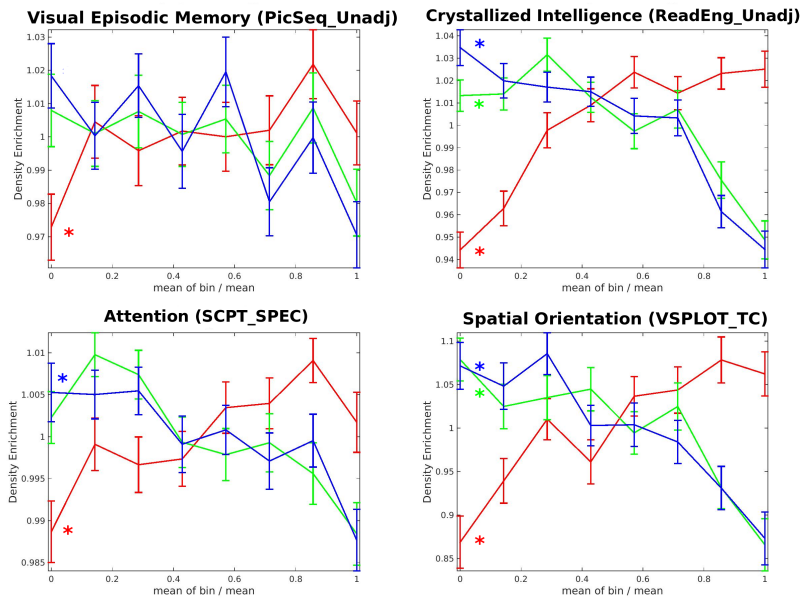
Fig. 4.6 Cognitive 2

Physical and Sensory
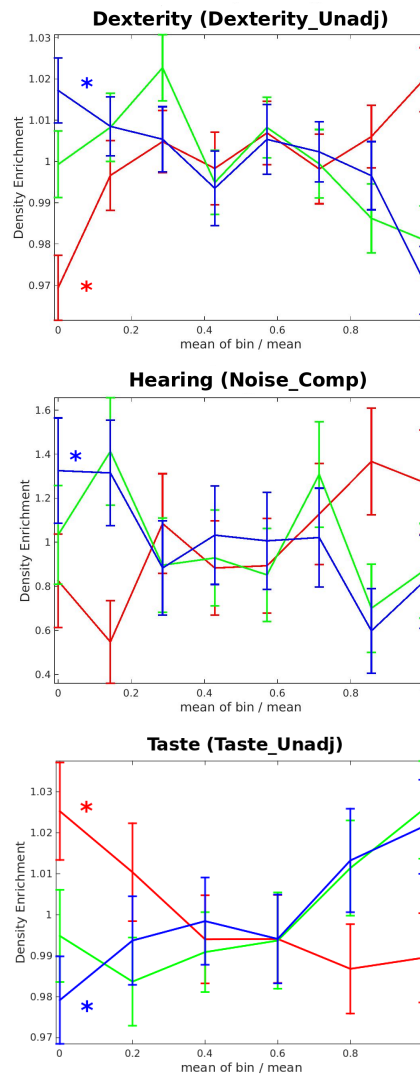
Fig. 4.7 Enrichment analysis of Cognitive, Physical and Sensory traits. We plotted all the enriched features of the Cognitive, Physical and Sensory traits, which result from the density analysis of the archetypes, in the case of 8 bins.

4.5.2 Personality, Substance use, socio-demographic traits

Data from 1123 participants were analyzed. Two analyses were performed separately on 70 continuous and 40 discrete measures (however, for clarity they will be described jointly). The enrichment analysis was carried out on measures clustered into:

1) self-reported measures reflecting behavioral, social, and emotional problems, adaptive functioning, and substance use (e.g., ASR and DSM-oriented measures);

Archetype	Experimental Measures	Features	Average difference (p-value)	First bin
B	ReadEng_Unadj	Crystallized Intelligence	2.5873E-12	max
B	PicVocab_Unadj	Crystallized Intelligence	1.1805E-10	max
B	PMAT_ACR	Fluid Intelligence	2.9223E-10	max
B	NEOFAC_O	Openness	0.00000285	max
B	PSQI_Score	Sleep problems	0.00001369	min
B	BMI	Body Mass Index	0.0000747	min
B	Endurance_Unadj	Endurance	0.00021863	max
B	SAGA_Income: 8	High Income	0.00032978	max
B	ASR_Rule_Raw	Rule-Breaking Behaviour	0.0012588	min
B	ListSort_Unadj	Working Memory	0.001287	max
B	Race:Asian/Hawaiian/Oth Pacific	Race	0.0021611	max
B	DSM_Antis_Pct	Antisocial Behaviour	0.002465	min
B	ER40_CRT	Emotion Recognition (RTs)	0.0056397	max
B	SCPT_SPEC	Attention	0.0063268	max
B	VSPLOT_TC	Spatial Orientation	0.0077114	max
B	Noise_Comp	Hearing	0.010801	max
B	Dexterity_Unadj	Dexterity*	0.010861	max
B	ASR_Extn_Raw	Externalizing	0.013512	min
B	DSM_Hype_Raw	Hyperactivity	0.017876	min
B	Taste_Unadj	Taste*	0.037597	min
G	VSPLOT_TC	Spatial Orientation	0.0040994	max
G	ASR_Thot_Pct	Thought problems	0.016071	min
G	Avg_Weeday_Any_Tobacco_7days	Tobacco	0.017359	min
G	ReadEng_Ageadj	Crystallized Intelligence	0.031099	max
G	ER40_CRT	Emotion Recognition (RTs)*	0.13533	min
R	ReadEng_Ageadj	Crystallized Intelligence	2.5873E-12	min
R	Race: Black/African American	Race	4.0364E-11	max
R	PicVocab_Unadj	Crystallized Intelligence	1.1805E-10	min
R	PMAT_ACR	Fluid Intelligence	2.9223E-10	min
R	Endurance_Unadj	Endurance	3.8829E-07	min
R	SAGA_Education: 12	Low Education	1.0987E-06	max
R	VSPLOT_TC	Spatial Orientation	2.3749E-06	min
R	SAGA_TB_Still_Smoking	Cigarette Smoking	7.5111E-06	max
R	Avg_Weeday_Any_Tobacc	Tobacco	0.00001353	max

Fig. 4.8 Enrichment analysis of the archetypes. The first column represents the label of each archetype (B = Blue archetype; G = Green archetype; R = Red archetype). The second and the third columns describe the measure and the corresponding trait enriched, respectively. The resulting p-value is shown in the fourth column and it is specified, in the last column, if the value of each trait is maximum or minimum in the bin close to a given archetype. The asterisk indicates traits that are significantly enriched using a 6-bins analysis.

2) substance use and physiological variables (e.g., quality of sleep, smoking);

3) socio-demographic features (i.e., educational level, race, income) (Figures 4.1-4.13; Table 4.8; Figure 4.9-4.11).

Individuals closest to the Blue archetype resulted more open to experiences, defined as an appreciation for art, creativity, intellectual curiosity, and preference for variety and novelty. They also reported the lowest scores on scales related to sleep problems, rule-breaking and antisocial behavior, hyperactivity and externalizing behaviors (such as impulsivity and aggression). Finally, they had the lowest Body Mass Index (BMI), a measure of body fat.

Individuals close to the Green archetype were characterized by minimum scores in thought problems (i.e., hallucinations, strange thoughts and behaviors, obsessive-compulsive behavior, self-harm and suicide attempts), and by the lowest number of cigarette smoked per day (or other tobacco-related substances (Table 4.8; Figures 4.9-4.10-4.11)).

Finally, near the Red archetype, several features enriched with maximum scores in scales reflecting aggressive, hostile, antisocial and rule-breaking behavior, withdrawn behavior and anxiety. Furthermore, individuals closest to the Red archetype reported the lowest life satisfaction, highest perception of stress, most feelings of social rejection, most somatic complaints, most problems related to intrusive thoughts, greatest interference of pain perception in daily life, and poorest sleep quality. Near this archetype, we also observed the highest number of smokers, individuals reporting to smoke the most cigarettes per day, and cannabis users as indicated by the number of positive cases to the THC drug test on the day of the experiment (Figures 4.1-4.13). Notably, BMI (obesity) was also maximal in the bin next to the Red archetype, and steeply declined with distance from that archetype.

Examining socio-demographic variables, individuals close to the Blue archetype had the highest income whereas individuals close to the Red archetype had the lowest income, lower educational level, and were most frequently unemployed. Finally, when considering enrichment on the variable race, Black or African-American individuals were more numerous near the Red archetype, whereas Asian (and Hawaiian or other Pacific Islanders) individuals were more concentrated in the bin closest to the Blue archetype (Figure 4.13). The variable race was one of the strongest enriched features (p= 4.06×10^{-11}). Therefore, it is important to ask whether a triangular distribution for the DDT scores existed separately in each race. As shown above (Figure 4.4), a Pareto optimal distribution was found

in each racial group, i.e. when considering separately White, Asian and Hawaiian individuals, or Blacks. In Black subjects, however, the distribution was also triangular, but no longer significant, compatible with the results of the enrichment analysis (see Figure 4.4).

   In summary, this enrichment analysis shows that stronger (Blue archetype) and more flexible (Green archetype) self-control, as indexed by the DDT scores, are associated with higher fitness on cognitive, behavioral, socio-economic, and health variables, while weaker self-control is associated with lower scores. Importantly, Blue and Green archetype subjects scored highest on different domains, suggesting different cognitive profiles (Figure 4.1 and Table 4.8).

Personality
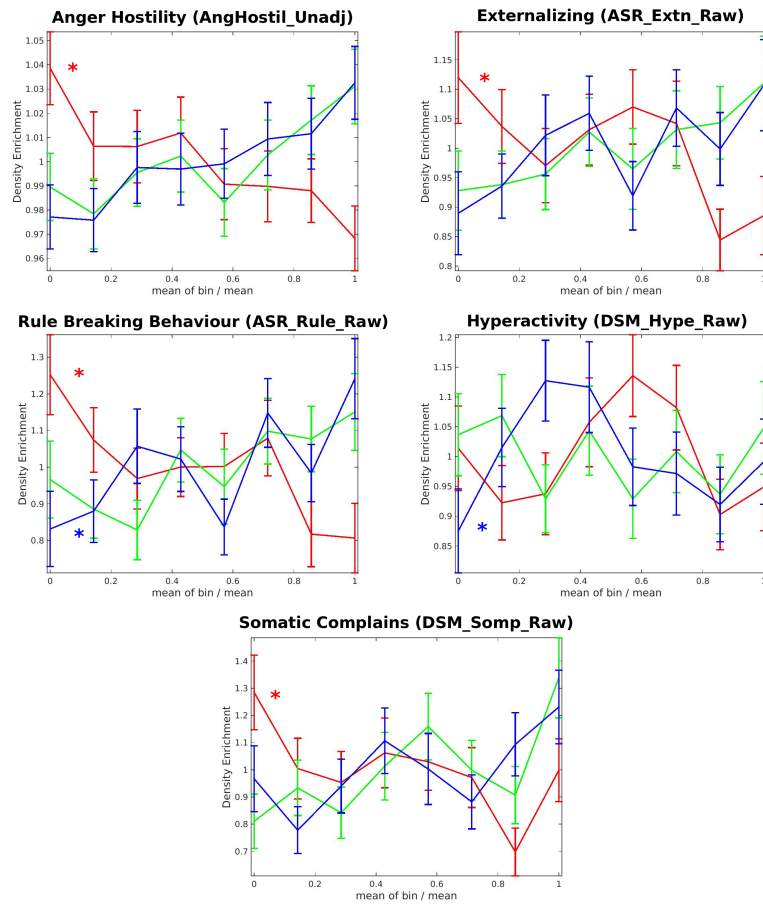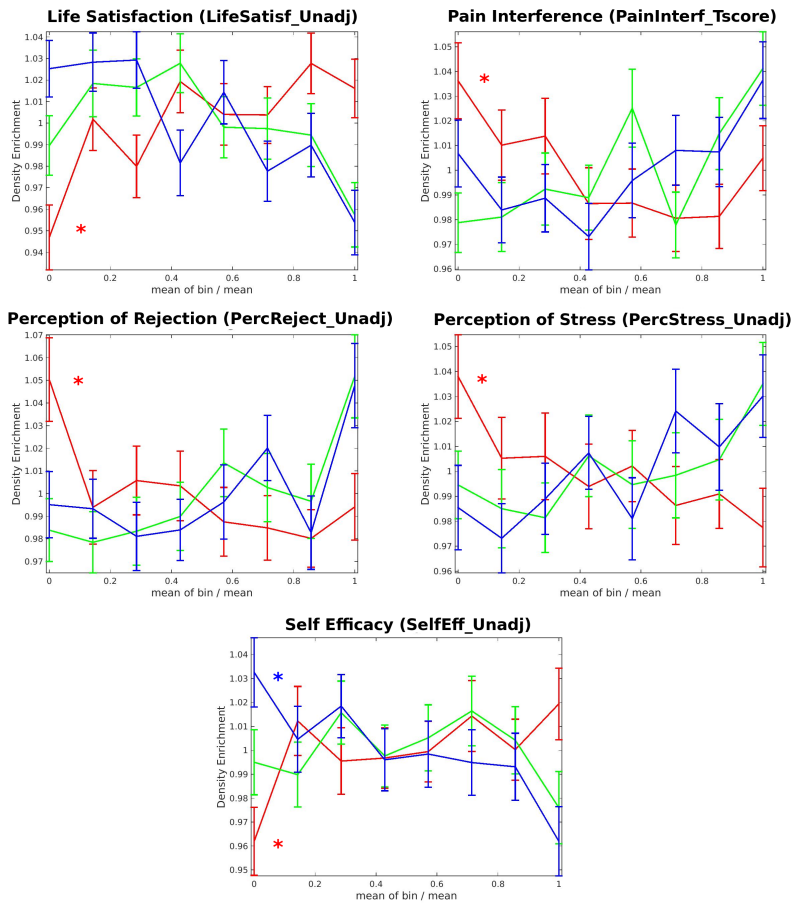

Fig. 4.9 Personality 1

Fig. 4.10 Personality 2
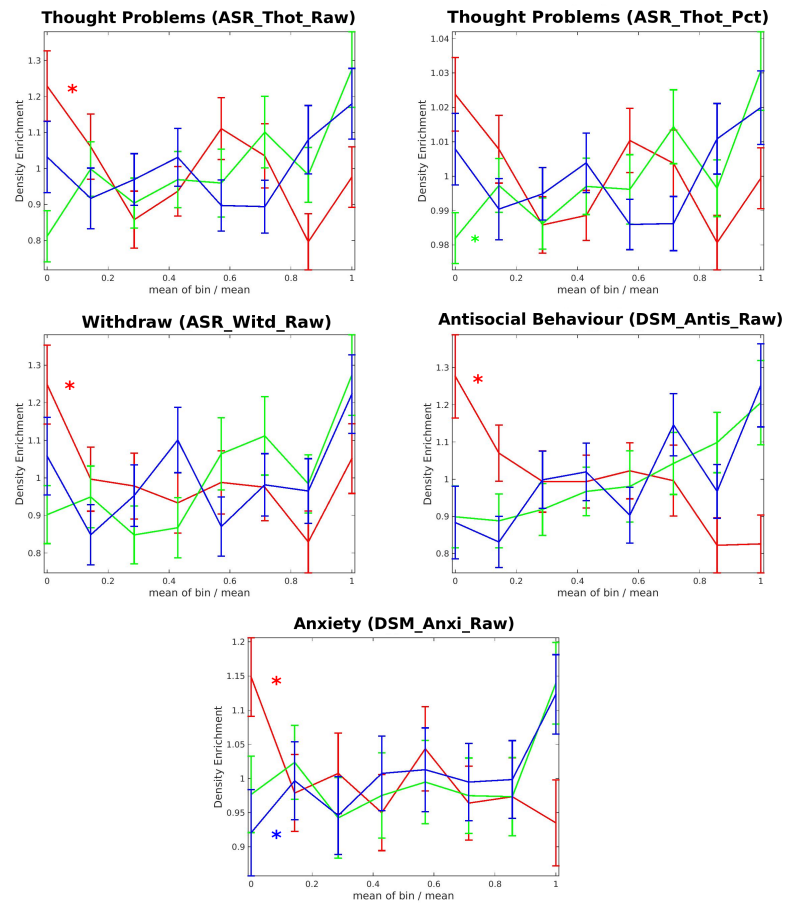
Fig. 4.11 Personality 3

1363 **Substance use**

Avg_Weekend_Any_Tobacco_7days

Cigarette Smoking (SAGA_TB_Still_Smoking

THC-Positive (THC:True)
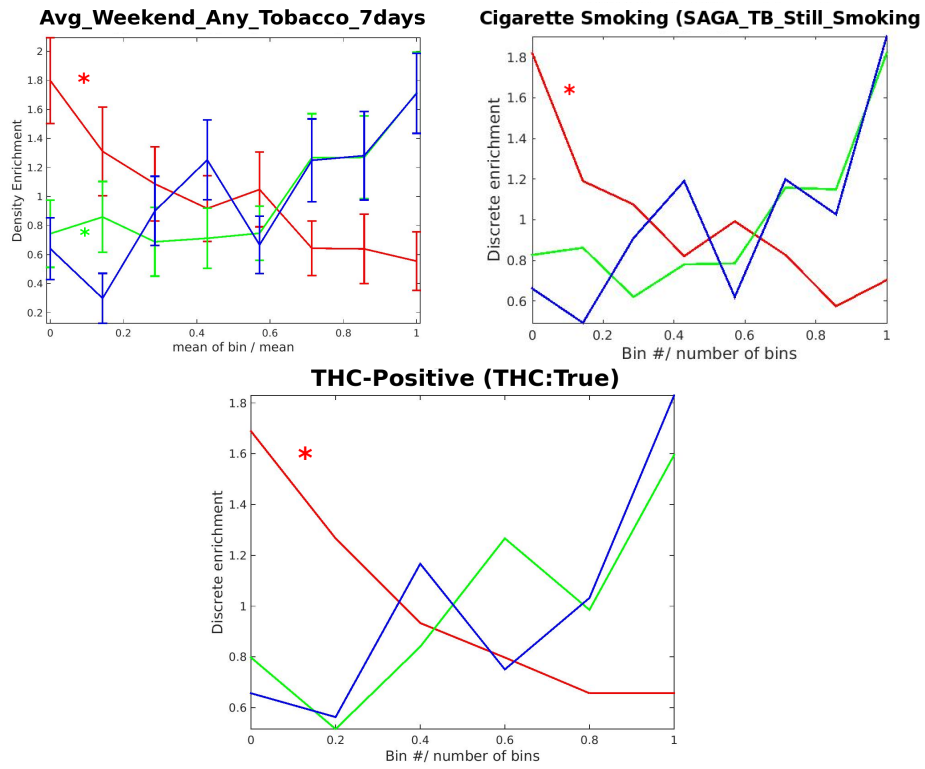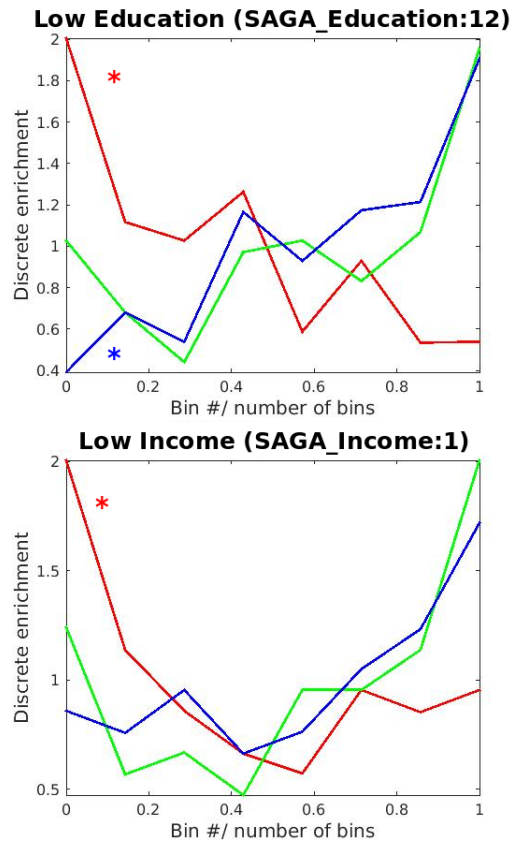
Socio-demographic

Fig. 4.12 Enrichment analysis of personality, Substance use and socio-demographic traits. We plotted all the enriched features of the personality, Substance use and socio-demographic group, which result from the density analysis of the archetypes, in the case of 8 bins.

4.6   Structural variables

We examined 56 measures related to mean volume of both white and gray matter, both in specific anatomical brain regions, and in the total cortical and subcortical gray and white matter level, normalized per intracranial volume. Measures were collected from a total of 1105 participants. Only total cortical gray matter volume was shown to be significantly enriched near the archetypes.

Total cortical gray matter volume was highly enriched near the Blue archetype reaching a maximum value near that archetype (Figure 4.14). To compare total gray matter volume as function of archetype, we ran an ANOVA restricted to individuals close to each of the three vertices (100 participants per group). This analysis showed a significant effect of
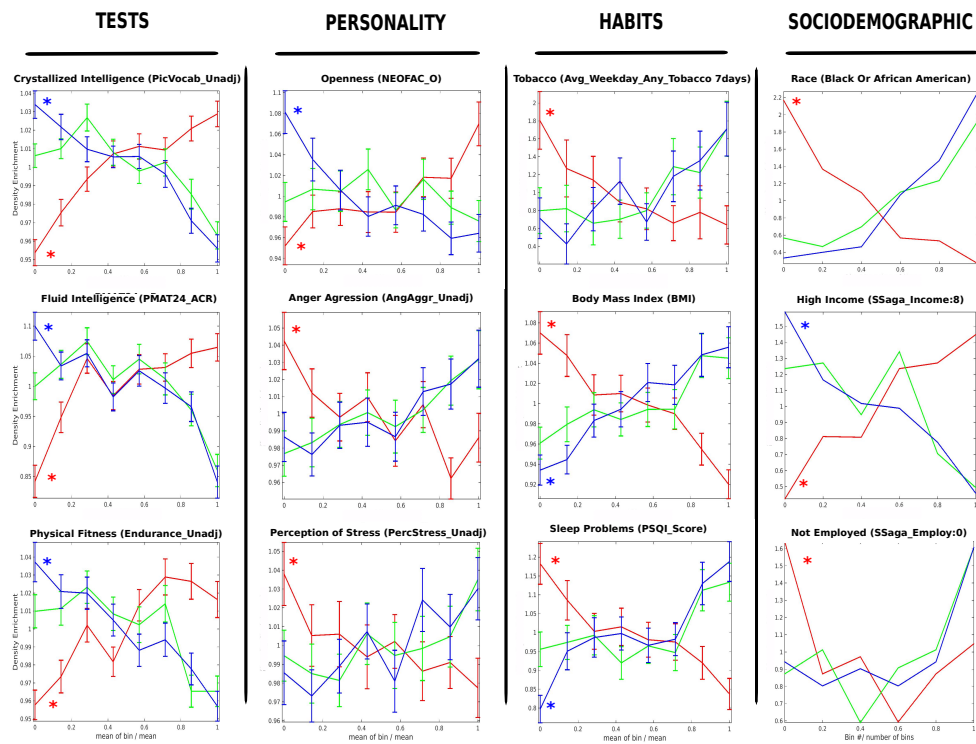
Fig. 4.13 Enrichment of different features near each archetype. Individuals were binned to equal sized bins according to distance from each archetype. The average value in the bin is normalized by the average value in the whole front distribution. The error bars are computed only for continuous measures. The enrichment analysis included cognitive tests, personality scales, substance use and socio-demographic features. Curves for features that enrich significantly near an archetype are marked with an asterisk.

archetype $[F(2, 297) = 7.9; p < .001; = .05]$, with the Blue archetype being characterized by larger cortical gray matter volume as compared to both Red and Green archetypes (p < .05; Bonferroni correction) (Figure 4.14). No difference was instead observed between Red and Green archetypes (p > .05). In summary, stronger self-control (Blue archetype) was associated with larger gray matter volume. Importantly, Blue and Green archetype subjects showed a different profile.

4.6.1 Resting-state Functional Connectivity analysis

To characterize differences in functional connectivity among different archetypes of significant Pareto front solution, we analyzed resting state functional connectivity (FC) from R-fMRI as available in the HCP data set. Subjects. Three-hundred healthy subjects (172 F, age: 29 ± 3y) were selected from the 1200-subject release HCP dataset, considering, for each archetype, 100 subjects with minimal Euclidean distance from each archetype vertex of the Pareto distribution. This sample size was selected because it was similar to the average sample size of the binning analysis for feature enrichment.

Imaging Data. The HCP imaging protocol included up to four 15-minute runs of resting state fMRI (60 min total), divided in two imaging sessions (TR = 720ms, isotropic voxel-size 2 mm) and structural images, made available as data packages with pre-defined processing options, for more details refer to the study by [55]. In this analysis, we employed minimally pre-processed fMRI time series from surface space defined and registered by means of a Multi-modal surface alignment method (MSM-All, ([112])) with minimal smoothing (surface and volume based 2mm spatial smoothing) and de-trending. Moreover, FIX-ICA ([121]) denoised data was employed as available from HCP public repository to reduce motion-related confounds ([86]).

Data Processing. Available denoised rs-fMRI time-series were signal averaged based on the functional parcels defined from the [58] for cortical regions, and a volume based segmentation ([45]) for subcortical regions (Cerebellum, Putamen, Pallidum, Ventral Diencephalon, Thalamus, Caudate, Amygdala, Hippocampus, and Accumbens in each hemisphere and Brainstem).

Parcellated rs-fMRI time series were Pearson cross-correlated and Fisher r-to-z transformed, with r the estimated Pearson linear correlation coefficient at edge-level ([70]) to obtain for each subject and run a FC matrix across 352 brain regions ([133]). We discarded rs-fMRI runs that included

more than 30% of motion corrupted volumes. Framewise Displacement (FD) was employed to identify the motion-corrupted volumes as it indexes bulk head movements across consecutive volumes ([106]) from the volume realignment parameters (motion correction). Since the available rs-fMRI data were previously pre-processed with FIX-ICA denoising, we relaxed the threshold for motion-corrupted volumes to $FD > 0.5$ mm as compared to previous suggestions of $FD > 0.15 - 0.2$ mm ([106]). After removal of motion-corrupted runs, all subjects had at least two valid fMRI runs. Correlation values in corresponding edges were finally averaged across valid runs to obtain a single FC matrix per subject.

The subjects included in the sample were not found to be significantly different in terms of motion content as function of the archetype. Inter-run and inter-subject global variability was removed by normalization ([52]).

ROI analysis on DDT and reward. Importantly, we performed a region of interest (ROI) analysis in the three groups of subjects based on a-priori hypotheses of cortical and subcortical regions recruited during the DDT and associated with reward processing ([84]; [83]; [155]). The selected ROIs were: Ventromedial prefrontal cortex (vmPFC), orbitofrontal gyrus (OFG), middle frontal gyrus (MFG), dorsomedial prefrontal cortex (dmPFC), dorsolateral prefrontal cortex (dlPFC), superior frontal gyrus (SFG), anterior prefrontal cortex (aPFC), anterior cingulate cortex (ACC), posterior cingulate cortex (PCC), anterior internal capsule (aIC), hippocampus (Hip), parahippocampus (Parahip), Striatum, Caudatum, Putamen, Accumbens, Globus Pallidus, Thalamus, and Amygdala. These ROIs were mapped onto the cortical/subcortical parcels/regions of the Gordon-Lauman atlas according to a visual overlap criterion at the group level. The selected ROIs overlapped with 63 parcels of the 352-parcels of the Gordon-Lauman atlas extended to subcortical regions. Therefore, the initial 352x352 FC matrix was reduced to a 63x63 matrix. In general, each ROI included multiple adjacent parcels with very similar functional connectivity profiles. To enhance the statistical robustness and the interpretability of comparisons across archetypes, we averaged the correlation values of adjacent parcels within anatomically defined ROIs based on Destrieux Atlas ([45]) and across hemispheres (left and right homologous parcels were averaged). This led to a reduction of the correlation matrix from 63x63 parcels to 18x18 ROIs corresponding to the functional ROIs identified above from the literature. To check that this anatomical selection was not introducing biases, we ran a hierarchical clustering on the FC profiles of the 63 parcels (Ward hierarchical method, [153]). The tree was cut to yield the same number of clusters as the anatomical areas of interest

(i.e. 18). We found a high point-wise agreement (high Rand's index of 0.922; [72]) between the clusters and the anatomical grouping criteria. Analysis and statistical comparisons. We carried out a Ward hierarchical clustering between coupled archetypes based on Euclidean distance similarity of connectivity profiles (i.e. FC rows, or columns by symmetry) similar to [96]. This analysis consists in the hierarchical clustering of FC matrices to identify the node clustering structure of one group of subjects (e.g. those belonging to one archetype) and use this structure to reshape the FC representation of another group of subjects (those belonging to the other archetype). In this way, differential hierarchical organization between FC in different groups of subjects will be visually clarified. As we did not find any significant difference in the FC hierarchical organization among the three archetypes, the reported analysis is based on clustering of FC matrices based on all subjects across the three groups. Next, we tested for differences among groups using a 1-way Analysis of Variance (1w-ANOVA) with bootstrap sampling for statistic evaluation on pair-wise ROI FC (Fisher-transformed Pearson correlations) testing the null hypothesis of equal connectivity between the three archetypes (see [159], for a similar approach). An FDR method was applied to correct for not independent multiple comparisons testing conditions. Post-hoc tests were run by means of one-tailed paired two-sample t-test with bootstrap sampling to investigate the directionality of connectivity by archetypes couples. FDR correction was again employed and restricted according to a Bonferroni strategy over the number of performed post-hoc tests.

Software and tools. Processing of rs-fMRI data, available as Neuroimaging Informatics Technology Initiative volumes (NIFTI) or Connectivity File Based Data (CIFTI) files was done with Connectome Workbench ([86]) and CARET (Van Essen Laboratory, Washington University) as well as surface visualization and representation of relevant brain areas. Statistical comparisons and further analysis were performed in MATLAB (R2016b; MathWorks, Natick, MA).

4.6.2   Brain functional connectivity

To explore differences in functional organization we compared resting state FC to/from ROIs recruited during the DDT and associated with reward processing ([84]; [83]; [155]) mapped onto the Gordon Laumann functional atlas of the human cerebral cortex ([58]). This analysis was run in three samples of subjects (each n = 100) who were closest to each archetype on the DDT. The three samples were matched in gender frequency (per-

centage of females: Red = 63%; Green = 52%; and, Blue = 57%) (Chi-square test, p > 0.1 for each paired comparison), and age (Average age: Red = 28.9 years old; Green = 28.6 years old; Blue = 29.6 years old) [F(2,299) = 1.99, p > 0.1], variables known to influence functional connectivity. The subjects were the same as those utilized in the structural MRI assessment.

A paired hierarchical analysis of connectivity profiles showed two main clusters: one cluster cortical involving regions in medial prefrontal and parietal cortex plus hippocampus, para-hippocampus, and amygdala; the other cluster subcortical-cortical including basal ganglia, thalamus, and lateral prefrontal cortex (Figure 4.15A).

The cortical cluster (violet in Figure 4.15A) includes areas belonging to the fronto-parietal network (FPN) and the default mode network (DMN), typically involved in control- and regulatory processes. The subcortical cluster (orange in Figure 4.15A) includes regions more strictly related to reward processes. To examine functional connectivity differences across archetypes, we ran a 1-way bootstrap-ANOVA with 0.05 significance level (FDR corrected for multiple comparison across 18 ROIs x17/2 tests). Figure 4.15B shows edges where FC significantly differed between archetypes: red vs. blue post-hoc comparisons under the diagonal, and blue vs. green above the diagonal of the matrix.

Interestingly, there were significant differences in ROI connectivity between clusters (Figure 4.15B), specifically between prefrontal and cingulate regions, involved in control and regulation, and subcortical regions involved in reward. In contrast, there was no significant difference in ROI connectivity within each cluster. In particular, subjects of the Blue archetype, as compared to subjects of the Red and Green archetypes, showed increased FC: 1) between amygdala and posterior cingulate cortex (PCC), thalamus, caudate nucleus and putamen; 2) between caudate nucleus and ventromedial Prefrontal Cortex (vmPFC), anterior cingulate cortex (ACC), PCC, amygdala and ventral diencephalic structures (e.g., substantia nigra, hypothalamus, thalamus); and 3) between anterior prefrontal cortex (aPFC) and vmPFC (Figure 4.15B). All these connections, except those involving the amygdala, were also stronger in subjects of the Green archetype as compared to subjects of the Red archetype. The Red archetype showed stronger FC between superior frontal gyrus (SFG) and ACC and hippocampus, as compared to the other two archetypes. In summary, stronger (blue archetype) and more flexible (green archetype) self control was associated with stronger FC between reward/emotion related regions (e.g. amygdala, caudate) and control related regions.
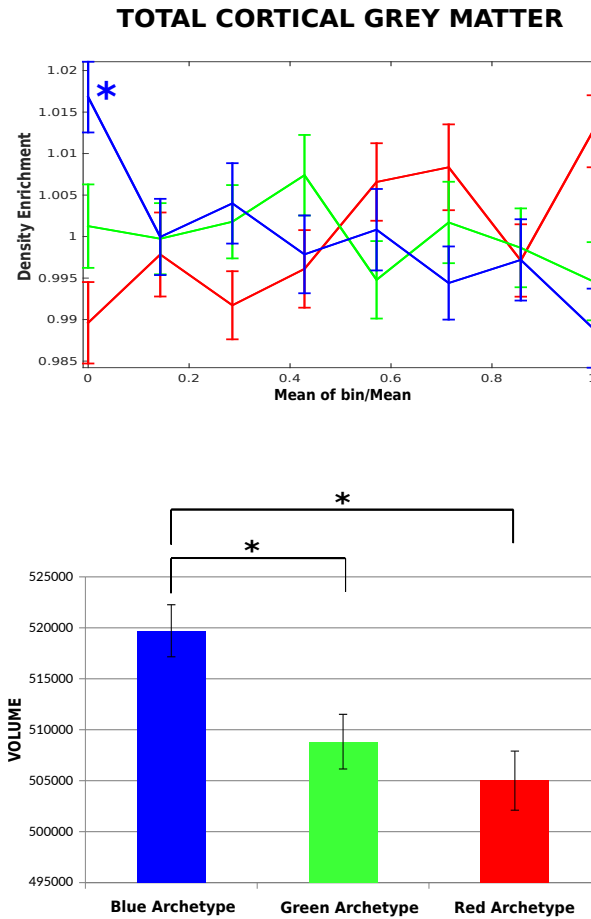
Fig. 4.14 Total cortical grey matter volume varies as a function of archetype. The enrichment analysis (left panel) shows that total grey matter volume is enriched for the Blue archetype. The histograms (right panel) indicate mean volume in the sub-groups of participants (n = 100 for each group) that are closest to the three archetypes. Total cortical gray matter volume is maximal for individuals next to the Blue archetype, intermediate next to the Green archetype, and minimum next to the Red archetype. Asterisks highlight significant differences. Bars indicate standard error. .

4.7 Analysis of heritability

Finally, we sought to investigate the heritability of time preferences for rewards by assessing possible differences in intra-class correlations (r) for the AUC \$200 and AUC \$40,000 between pairs of monozygotic twins ($MZ; n = 130$) and dizygotic twins (DZ; $n = 138$) by means of Fisher's z test. Then, we calculated the heritability (h2) index on the basis of the difference in the MZ–DZ correlations for AUC \$ 200 and AUC \$40,000, applying the Falconer's formula (see the study by [39] for a similar approach).

4.7.1 Twin correlations and heritability

In the last analysis, we explored the genetic influence on time preferences for rewards by assessing possible differences in intra-class correlations (r) for the AUC $200 and AUC 40000$ between pairs of MZ twins and DZ twins by means of Fisher's z test. The correlation value did not significantly differ between MZ and DZ pairs, either for the AUC $200 (MZ r = 0.30 versus DZ r = 0.32; z = −0.208 p = 0.48), or the AUC 40,000$ (MZ r = 0.51 versus DZ r = 0.40; z = 1.158 p = 0.124). The difference in MZ–DZ correlation for AUC $40,000 was 0.11, indicating a broad heritability (h2) of only 0.22. For AUC$ 200, this calculation was even meaningless as the value for DZ twins was higher than the value for MZ twins. Therefore, MZ twins were not substantially more similar in delay discounting than DZ twins. The heritability (h2) value indicates that there is not a strong genetic dominance of this trait, as genetic dominance can be inferred for DZ twin correlations that are about ¼ MZ twin correlations.
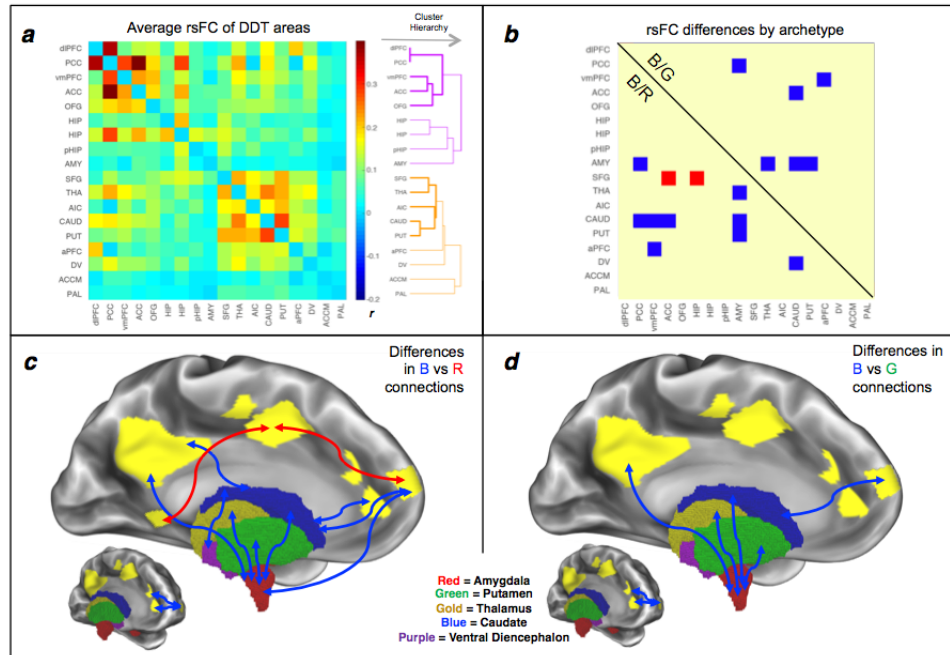
Fig. 4.15 Resting-state functional connectivity differences between archetypes. (a) Average rsFC matrix between regions of interest involved in reward and delay discounting task. The FC matrix is divided in two clusters based on a hierarchical cluster analysis (the colour indicates the same functional module membership; the thickness of the line represents the similarity of FC weighted by the connectivity significance). **(b)** Differences in rsFC among the three archetypes as identified by post-hoc comparisons. The lower triangular part compares Blue (B) versus Red (R) archetypes; the upper triangular part contrasted B archetype versus Green (G) archetype. The color of the squares indicates the edges showing stronger rsFC (p < .05, FDR corrected) for one archetype over the other. The 'c and d' panels depict the topography of significantly different connections. Connections are coloured according to the archetype that shows stronger connectivity level, separately for B/R comparison (c panel) and B/G comparison (d panel). Cortical regions are displayed in yellow, while subcortical regions are displayed according to the color legend.

4.8   Conclusion and Discussion

In the present chapter we have applied Pareto Optimality theory to human cognition and behavioral data to find trade-offs and archetypes that represent potentially different evolutionary strategies in cognitive development. In the HCP dataset that measures in a large sample of healthy subjects, cognitive, sensory and physical abilities, personality traits, substance use, and socio-demographic variables, the strongest Pareto Front solution was found when we projected scores from two measures of the DDT that measures time preferences for reward, an index of self-control and regulation. This Pareto Front triangular distribution was robust in independent samples of subjects. The archetypes defined different strategies for time preference for reward that enriched on different cognitive functions, but also physical, emotional, personality, and socio-economic variables. The archetypes also differ in total gray matter volume, and functional connectivity between subcortical reward and cortical control regulatory regions. Finally, archetypes were weakly affected by genetics.

In this section, we discuss the difference between Pareto Optimality and g-factor accounts of cognitive variability, potential evolutionary pressures that led to different strategies in time preference for reward, and underlying neural correlates, which provide insights into evolution, cognition, neuroscience, psychology and economy.

4.8.1   Pareto Optimality vs. g-factor theories of individual variability in cognition

We focused on the first one related to the DDT scores that appeared to be the most robust. This experiment was not designed to pitch Pareto Optimality vs. g-factor theories, but to evaluate the presence of Pareto fronts and their potential significance in human cognition and behavior. The results clearly support that there is more than bivariate relationships in human cognition, and time preference for reward appears a powerful variable that shapes many other cognitive, behavioral, and brain variables.

4.8.2   Time preferences for reward: evolutionary perspective

The evolutionary foundation of time preference for rewards has attracted the interest of economists and biologists for many years ([115]). The study of delay discounting and time preferences for reward originated from

animal work (e.g., [107]). This body of research has shown that animals discount rewards hyperbolically ([64]), and that birds and roditors discount delayed rewards significantly more steeply than humans ([1]). Interestingly, bonobos and chimpanzees - our closest living relatives - show a degree of patience not present in other species, and chimpanzees are even more willing to wait for food than humans. Overall these studies support the evolutionary importance of discounting rewards as time-sensitive decisions are important for foraging and mating in their natural environment (see [60]).

In this chapter, we have shown that measurements of time preferences for reward in humans distribute according to a triangular Pareto front which, which according to the theory, indicates that this trait is under evolutionary pressure. The archetypes identified by the analysis correlate with other cognitive, physical, emotional, and socio-economic variables that should provide those specialist individuals with relative advantages from an evolutionary standpoint.

People close to the Blue archetype enrich on features that are typically considered positive and desirable qualities, at least in a highly structured and modern environment. For example, being intelligent, agreeable, and open, as well as physically fit, could increase the likelihood to find a mate, as well as earning a high income could increase the offspring quality, via better nourishment and/or investment in education.

Likewise, people near the Green archetype flexibly changes the strategy according to the reward amount, suggesting, as compared to the two archetypes, a greater flexibility in adapting their behavior to environmental pressures. Also, these Green archetype individuals are best at recognizing facial expressions, which may help them in understanding others' feelings and needs.

The evolutionary advantage of people near the Red archetype is less intuitive, but it may be explained as follows. Firstly, there may be 'evolutionary mismatch' between the environment in which we currently live and the environment in which we evolved. Therefore, a behavior that was adaptive hundreds of thousands to hundreds of years ago becomes inappropriate into our current environment ([113]). In some circumstances, for example, children and adolescents showing aggressive and externalizing behaviors become dominant and respected in their peer groups, whereas in other cases become unpopular or rejected ([49]). Hence it is conceivable that the strategy of taking immediately irrespective of the rewards might have been more advantageous in the past to achieve social status and dominance. Secondly, according to life history theory, time preferences

are influenced by resource scarcity, mortality and uncertainty in the environment ([65]). Delay discounting rate was found to be steepest under stressful conditions in people with low socio-educational background or poor health, all conditions in which individuals close to the Red archetype report to live ([24]; [65]). Finally, natural selection would favor individuals who made reproductive efforts sooner. In this regard, although the HCP dataset does not include such information, we expect that individuals close to the Red archetype were more likely to have their first child sooner and have a larger number of offspring. This speculation is supported by data showing that a steeper discounting rate in teenagers and young adults is associated with a range of sexual behaviors, including earlier first experience with sexual intercourse and past or current pregnancy ([25]). Furthermore, if discounting rate is influenced by the expected future fitness, then living in relatively adverse circumstances (e.g., elevated risk of mortality, high stress levels, resource scarcity) makes individuals more prone to activate reproductive effort immediately ([37]), as also apparent in other species (e.g. wasps, [116]).

As for the nature vs. nurture question: are archetypes in time preferences for reward genetically or environmentally determined? The absence of significant differences between MZ and DZ correlations and the low heritability (h2) value indicate a weak genetic influence. Yet, genetic and cultural selection are not mutually exclusive. Heritability of time preferences is indeed not constant across lifespan. It is higher during late childhood/adolescence ([2]) and several studies found genetic polymorphisms being associated with differences in time preferences ([14]; [44]). By contrast, heritability has less contribution in adulthood (age range of HCP participants: 22-35 years), when other factors, such as environmental stressors and/or cultural factors, could have an impact on individuals' time preferences to some extent. A sensitive issue is the impact of evolutionary vs. socioeconomic factors in explaining the high proportion of Black and African American individuals near the Red vertex. Adverse health and socioeconomic conditions, as consistently revealed by the large amount of data collected through the NSAL (The National Survey of American Life: *http* : *//www.rcgd.isr.umich.edu/prba/nsal.htm#overview*), may favor strategies that emphasize short term rewards. At the moment, however, the present findings cannot clearly disentangle biological and cultural factors.

4.8.3 Archetypes for time preference for reward: brain and cognitive associations

Our study demonstrates that archetypes for time preference for reward also differ in brain structure and functional connectivity. The Blue archetype has larger cortical gray matter volume respect to the other two archetypes, consistent with previously reported associations between brain volume and intelligence ([111]), or self-control, a critical function in the DDT ([85]). Interestingly, in [85] the evolution of self-control was linked to absolute brain size across 36 different species ([85]).

The three archetypes also differed in the functional connectivity profiles of brain regions associated with the DDT ([83]; [84]; [155]). Individuals with more self-control showed stronger functional connections at rest between cortical prefrontal, cingulate, and parietal regions involved in control and regulation, and subcortical regions involved in reward and emotions. Importantly, functional connectivity differences between archetypes occurred in the projections that connected different modules. In previous work, stronger functional connections between modules or networks were observed when subjects went from rest to an attention demanding task, consistently with increased interactions (e.g. [136]). So we can interpret our results suggesting that individuals with more self-control have more communication between regulatory control regions and reward regions.

These data are also consistent with a number of dual-system models of decision-making (e.g., [7]; [9]). These models state that decision-making underlies a relative balance of activation between two neurobiological systems ([9]). An evolutionarily older impulsive system that includes limbic and paralimbic regions (amygdala, ventral pallidum, striatum, nucleus accumbens) values immediate rewards. By contrast, a more recently evolved control system that includes PFC and ACC is important for the inhibition/regulation of the impulsive system and the associated evaluation of delayed rewards. Our findings support these ideas showing that the ability of delaying a reward is associated with stronger functional coupling between regulatory cortical and reward subcortical regions, specifically amygdala and caudate.

A key area of the reward system is the amygdala, whose functional connections with putamen, caudate, and aPFC in our data (Figure. 4.15C-D) were strongly modulated by archetype, stronger in the Blue than Red and Green archetypes. The amygdala is classically regarded the core region for the regulation of emotions regulation ([33]), and a hub of emotion related networks ([102]). In line with our results, altered amygdala-centered

connectivity was found in drug addicts ([141]) who show steeper discounting rates and lower self-regulation ([11]). Interestingly, [136] reported altered resting-state functional amygdala-centered connectivity in cigarette smokers during early nicotine withdrawal. The ability of self-control and postpone a reward may be the result of a stronger functional connections to/from the caudate nucleus. Fronto-striatal circuitry is implicated in inhibitory control ([53]), with the caudate nucleus associated to behavioral control and goal-directed actions ([61]). Importantly, [57] documented that connections between dorsal caudate and frontal regions facilitate self-control. The increased FC between caudate and PFC regions in subjects able to exert stronger self-control is consistent with these findings. Conversely, alterations of cortico-striatal connectivity has been linked to disruption of self control. Several studies have reported alteration of functional connectivity between ACC and striatum in cigarette smokers ([71]; [82]), as well as altered activation of these regions in cannabis users ([160]). [71] have proposed that rsFC between dACC and striatum may represent a circuit-level biomarker for nicotine addiction.

The Red archetype showed stronger functional connections between ACC and superior frontal regions. Although at a first sight this result appears counterintuitive, it is, however, consistent with a study that found stronger functional coupling in ACC-frontal circuits to be predictive of a poorer DDT performance in drug addiction, even if it is important to acknowledge that the study involved a different population, namely cocaine users ([19]).

Finally, from a psychological perspective, although the present study cannot make any conclusion about causal relationships, it provides the most comprehensive overview of the associations between time preference and other individuals' attributes. We observed that people tendency to choose more immediate or more delayed rewards is a crucial trait that can explain individual differences not only in cognitive abilities, but also personality traits, substance use and dysfunctional behaviors, as well as socio-demographic features. Notably, in line with previous studies, we found that a stable preference for immediate smaller rewards seems to predict a constellation of behavioral and real-life problems, including hostile, antisocial, rule-breaking and withdrawal behaviors (e.g., [48]), anxiety ([119]), problems of intrusive thoughts ([134]), sleep problems, high levels of stress and high BMI (e.g., [22]), somatic symptoms and pain interference with daily living ([149]), and perception of rejection, low levels of life satisfaction and self-efficacy, and substance addiction (e.g., [11]). Taken together, our findings support the idea that steeper discounting rates are

1747 associated with a range of impulse-control disorders and unhealthy behaviors ([10]; [110], for reviews). Therefore, time preference appears to be a
1749 promising candidate endophenotype for multiple dysfunctional behaviors
1750 and might represent a therapeutic target for treating these disease states.

Chapter 5

Variational principle for xylem's tapering in vascular plants.

5.1 Introduction

In chapters 3 and 4 we applied the Pareto optimality analysis to search for Pareto optimal fronts in the morphospace to fully disclose the role of evolutionary pressures in biological systems which face complex multi-objective optimization problem. Based on the optimization criteria of minimizing the energy dissipated in a fluid flow, we propose in this chapter a biophysical model with the goal to explain the underlying physical mechanism that affects the structure of xylem conduits in vascular plants, which result in tapered xylem profiles [104, 105, 117, 164].

Xylem conduits are the fundamental constituents in trees which convey water and nutrients by means of a negative pressure gradient from roots to leaves, and their conductance measures the degree in the efficiency of water transportation ([56], [164], [138], [66],[67]). The concept of resistance of the fluid flow, which is by definition the inverse of the conductance, can be approximately accounted by the Hagen-Poiseuille equation [140, 154, 164]. This physical law is only valid in the idealistic case of long cylindrical pipes of constant cross section, being proportional to the length of the pipe and inversely proportional to the fourth power of the radius.

Existing optimality models of the tapering of xylem conduits ([130],[157], [123]) assume that xylem profiles have acquired a tapering degree in order to optimally convey water and essential nutrients to all parts of the trees ([104, 105, 117, 164]).

We propose here an optimality model by aiming to minimize the hydrodynamic resistance of the sap flow inside the xylem conduits, in the context of the *calculus of variations*. The variational approach presents

several theoretical advantages, in what it is mathematically well established and it has been widely applied in physics.

In section 5.1.1 we make a brief review of the literature for two prominent models, such as the pipe and WBE models. In section 5.2 we present our model by first defining an initial functional, that accounts for the Hagen-Poiseuille resistance term and in section 5.9 we show how the tapered xylem profile can be derived. In section 5.3 we validate of our results by sampling 72 vascular plants species of the angiosperm family and fitted them with the parametric curve. As compared to the WBE-model, it comes out that our theory accurately describes the tapering seen in xylems by means of a universal function, which is a key property of the model. In addition, the model allows us to correlate the heights of the trees with the xylem's radius at the stump of the trees. Finally, in section 5.4 we discuss our results.

5.1.1 Classical hydraulic models

A great theoretical effort has been devoted to model the xylems' profiles in order to shed light on the tapering mechanisms in vascular plants ([30] [98]), see Figure 5.1. A distinctive approach in hydraulic models is related to the incorporation of the tapering effect in xylems. In the *pipe model* ([130],[131]), authors conceived xylems of as thin cylinder with a constant diameter along the stem. Recently, West, Brown and Enquist conceptualized an optimality hydraulic model [157] (WBE-model), and generalized in [123], where xylems widen tip-to-base with a power law scaling, in order to minimize the hydrodynamic resistance cost of the sap flow inside xylems.

Based on the underlying assumptions of [157], xylems' architecture results in a fractal-like transportation network, which is structured in several branch levels. Each branch is composed by an identical number of xylem segments, and is connected in series with the branches of previous and further levels. This model is mainly an idealized representation of the xylems' architecture since they totally ignore the tapering of the radius in xylem conduits within segments in a given branch.

The WBE model is based on four simplifying axioms. The first axiom regards the space-filling property of the branching pattern, which induce the fractal-like network in xylems. This axiom is inspired from the observation that most distribution systems can be described by a branching network [156]. As a second axiom it is required the size-invariance of leaf and petiole, meaning that the capillary density in a cross sectional
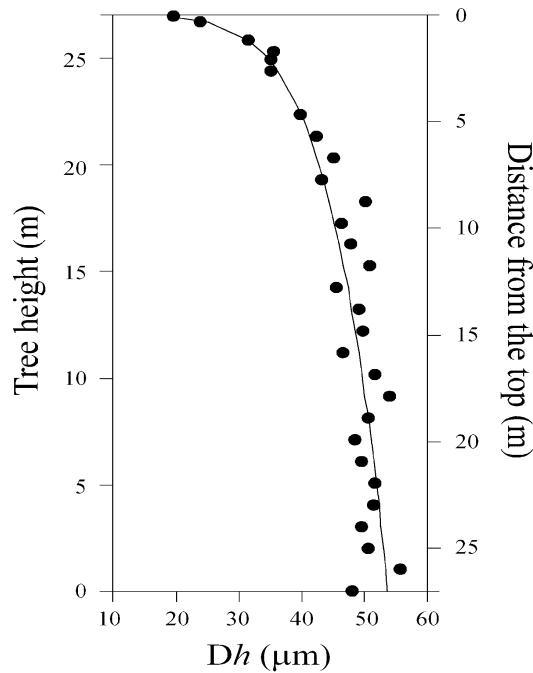
Fig. 5.1 Tapering in xylem conduits. Adapted from: "Convergent tapering of xylem conduits in different woody species", Anfodillo et al.

area remains constant across levels. The third axiom is related to the area-preserving branching condition, which is a bio-mechanical constraints that assures that at each level, branches split in smaller ones whose area sums to the original one. The fourth axiom requires the minimization of the total hydrodynamic resistance term [156, 157], representing the optimization criteria in which we are mainly focused.

Based on these assumptions it is possible to derive a plethora of scaling relations (refer to [157] for more details), and among others, the scaling exponent of $1/4$ of the tapering of xylem conduits (see Appendix C). The WBE-model has turned out to be the reference model for analyzing xylems tapering, however, several criticisms have arisen by showing the inadequacy of some biological assumptions and of theoretical derivations, and demand for an improved biophysical model, which is capable to overcome these fundamental issues ([117], [164], [162], [78], [79], [104]).

5.2 The model

In this section we explicit the mathematical framework of the theory, which is based on the variational formalism. Johann Bernoulli in 1696 was the first to apply the variational principle and find the optimal solution which minimizes the total time for a sliding object to descend from an

higher to a lower point. The curve, called brachistochrone (from Ancient Greek, meaning "shortest time"), is found by using tools from the calculus of variations [118].

In our model, xylems are thin pipes that continuously taper from roots to petioles and individually feed single leaves. Similarly to the WBE model, we still make the basic assumption that the xylem profiles are optimized by evolution to minimize the hydraulic resistance. However, contrary to the WBE assumption that pipes have a constant cross sectional radius within a given branching level, we highlight that tapering of xylem conduits is a continuous effect running through the whole path from roots to leaves.

By following the variational formalism, we start by defining the main functional of the model $\mathcal{F}[\sigma(h), \dot{\sigma}(h), h]$, which accounts for the Hagen-Poiseuille resistance term of the sap solute inside xylems, integrated from the stump of trees to leaves. The variational approach requires the minimization of the functional and the derivation of the Euler-Lagrange (EL) equations . The optimal function that minimizes the integrated hydraulic term is the solution of the EL equations. As we will show below, we are enforced to introduce a Lagrange multiplier term in the functional to limit the optimal solution to the biologically feasible space.

Hagen-Poiseuille term

The Hagen-Poiseuille law has been experimentally derived by solving the laminar flow dynamics of an incompressible and Newtonian fluid, inside a thin cylindrical pipe with a constant circular cross section [140]. The volumetric flow rate Q of a laminar fluid in a cylindrical pipe with circular cross sections is proportional to the applied pressure gradient ΔP, according to the Hagen-Poiseuille law [140, 154, 164]:

$$Q = \frac{|\Delta P| \pi R^4}{8 \mu L}, \tag{5.1}$$

where L is the length and R is the radius of the pipe, μ is the fluid viscosity, and ΔP is the pressure gradient between the tip and the base. The total resistance Ω of the pipe is defined as the inverse of the volumetric flow rate, thus $\Omega \sim L/R^4$, meaning that the flow resistance increases with the length of the pipe and decreases with the fourth power of the radius of the pipe.

For non circular cross section it is common to introduce a constant factor in the numerator of 5.1 (for more details see [42] and Appendix B). In this model we absorbed this constant in the parameter of the Lagrange

multiplier. The Hagen-Poiseuille law governs the dynamics of the flow of sap solute inside xylems in the limit of infinitesimal parts of the xylem, where the radius stands constant. The total hydraulic resistance in xylems is obtained after integrating the infinitesimal resistances $d\Omega(h)$ across the whole xylem path length. In terms of the cross-sectional area, defined as $\sigma(h) = \pi \left(\frac{d(h)}{2}\right)^2$, we get:

$$d\Omega(h) = \Omega(h)dh \sim \frac{c}{\sigma^2(h)}dh \tag{5.2}$$

where c is a positive real constant. Equation (5.2) is valid also for other general shapes, such as rectangular, triangular and ellipsoidal conduits as far as the lengths of the cross-sectional areas are of the same order of magnitude. Thus, the explicit expression of the starting functional $\mathcal{F}[\sigma(h), \dot{\sigma}(h), h]$, becomes the following:

$$\mathcal{F}[\sigma(h), \dot{\sigma}(h), h] = \int dh \left[\frac{1}{\sigma^2(h)}\right] \tag{5.3}$$

Lagrange multiplier term

The EL equations of the functional 5.3 are solved by the non biological xylem profiles $\sigma(h) = \sigma_{max} = \sigma_{trunk}$, which are contrary to any experimental evidence. In order to avoid this unfeasible solutions, we added a series of Lagrange multipliers, to put a cost in the tapering of xylems, which, especially in the proximity of leaves where the xylems enter steeply into the leaves. Thus the Lagrange multipliers are introduced with the following series:

$$a_1\dot{\sigma}(h) + a_2\dot{\sigma}^2(h) + a_3\dot{\sigma}^3(h) + \ldots + a_n\dot{\sigma}^n(h). \tag{5.4}$$

and the cost becomes infinite if $\dot{\sigma}(h) \to \infty$. The first term $a_1\dot{\sigma}(h)$ is trivial because its integral depends only on the boundary values of σ, thus the next simplest term is $a_2\dot{\sigma}^2(h)$, while higher order terms need the further parameters $a_3, ...a_n$ to be introduced. Thus, we start to consider for simplicity only the second term:

$$\mathcal{K}(h) = \frac{\alpha}{2}\,\dot{\sigma}^2(h). \tag{5.5}$$

where $\frac{\alpha}{2} \neq 0$ is the parameter of the Lagrange multiplier. Then, the new constrained action $\mathcal{F}_\alpha$ becomes:

$$\mathcal{F}_\alpha[\sigma(h), \dot\sigma(h), h, \alpha] = \int dh \left[\frac{1}{\sigma^2(h)} + \frac{\alpha\dot\sigma^2(h)}{2} \right] \qquad (5.6)$$

From the classical physical perspective, eq. (5.6) can be considered as the action of a classical particle with a kinetic term $K(h) = \frac{\alpha\dot\sigma^2(h)}{2}$ and a mass α in the effective potential $V(h) = \frac{1}{\sigma^2(h)}$.

Instead of considering the Euler-Lagrange equations of the above functional we focus on the *energy* of the system, defined as follows:

$$E = \frac{\partial L}{\partial \dot\sigma(h)} - L = \frac{\alpha\dot\sigma^2(h)}{2} - \frac{1}{\sigma^2(h)}. \qquad (5.7)$$

which is a nonlinear differential equation for $\sigma(h)$, and independent from h.

5.2.1 Optimal solution

We solve Eq. (5.7) analytically by putting on one side $\dot\sigma(h)$:

$$\dot\sigma(h) = \sqrt{\frac{2}{\alpha}}\sqrt{E + \frac{1}{\sigma^2(h)}} \qquad (5.8)$$

where $\alpha > 0$ and we fixed the boundary condition to be $\sigma(0) = 0$, in order to avoid a multitude of parameters. This condition is an approximation, because xylems do not completely close near the leaf; they indeed maintain a width proportional to the size of a cell. Thus, equation (5.8) can be analytically solved:

$$\sigma(h) = h^{1/2} \left(\frac{8}{\alpha}\right)^{1/4} \sqrt{1 + \frac{E}{\sqrt{2\alpha}}h} \qquad (5.9)$$

where E and α are free parameters of the theory. Since $\alpha > 0$, real solutions exist if $E < 0$. The stationarity point of 5.9 coincides to putting to 0 Equation 5.8. Thus, we have that:

$$\sigma_F = \sigma(h_{\max}) = \frac{1}{\sqrt{|E|}}. \qquad (5.10)$$

We can define $h_{\max}$ as follows:

$$h_{\max} = \sqrt{\frac{\alpha}{2}} \frac{1}{|E|}. \tag{5.11}$$

Thus, by substituting (Eqs. 5.10 and 5.11) in (Eq. 5.9) we have that:

$$\sigma(h) = \left(\frac{8}{\alpha}\right)^{1/4} \sqrt{h} \left(1 - \frac{h}{2h_{\max}}\right)^{1/2}. \tag{5.12}$$

The above xylem profile results in a tapered function, with the parameters h_{max} and α that depend on the specific tree. In the limit of $h \ll h_{\max}$, we have that $\sigma(t) \sim \sqrt{t}$, which is equivalent to the well-known power-law expression of $d(h) \sim h^{1/4}$, as predicted by the WBE model.

5.2.2 Data fitting

We validated our theory by performing a data fitting analysis of the predicted xylem profile 5.12, as derived from the variational approach, to experimental xylem profiles in a comprehensive dataset of 72 angiosperms. We considered the parametrized optimal solution 5.12 and transformed it to the more intuitive form:

$$\sigma(h) = \sigma_F \sqrt{\frac{h}{h_{max}} \cdot \left(2 - \frac{h}{h_{max}}\right)} \tag{5.13}$$

and fitted it to data points of angiosperms, by holding fixed h_{max} to correspond to the experimental heights.

We performed data fitting with the *lsqcurvefit* algorithm as implemented in Matlab. It is a nonlinear least-squares solver, useful in solving nonlinear data-fitting problems. Mathematically, it is equivalent to solving the following minimization problem (see $https://uk.mathworks.com/help/optim/ug/lsqcurvefit$

$$\min_x \sum_i (F(x, x_i) - y_i)^2. \tag{5.14}$$

We can map the above variables in terms of our problem as follows: $x_i \equiv h_i$ are the experimental height measures, $y_i \equiv \sigma_i(h)$ are the cross-sectional measures, $x \equiv \sigma_F$ is the free parameter to be fitted, and F is the nonlinear curve Eq. (5.13). We chosen $\sigma(0) = 1$ as initial value. To check the robustness of the initial point selection we used also the values $10^{-6}, 10^0, 10^{+6}$. To crosscheck, we performed the fitting analysis

with another algorithm called *fminsearch* and implemented in Matlab $https://uk.mathworks.com/help/matlab/ref/fminsearch.html$.

The parametrized optimal solution 5.13 carries a particularly interesting form because with a simple scaling $\Sigma(t) = \sigma(h)/\sigma_F$ and $t = h/h_{\max}$, we get:

$$\Sigma(t) = \sqrt{t(2-t)}, \tag{5.15}$$

which is generally valid for all tree species and independent of the height. $\Sigma(t)$ and $0 < t < 1$ are dimensionless variables. For visual purposes, we made the following transformation of the height axis in the log space, in order to plot the collapse:

$$x = \left(\log\left(1 + \frac{x_0}{x}\right) - \log(x_0)\right) \cdot \left(1 + \left(\frac{x_0}{x}\right)^b\right); \tag{5.16}$$

$$x_0 = 10^{-5}; \tag{5.17}$$

$$b = \frac{1}{2}. \tag{5.18}$$

where x_0 and b are two parameters chosen to equally distribute data points in the whole range $0 < t < 1$. This transformation does not affect the relationship among the variables σ and h. It serves only as an auxiliary tool for showing results.

In Figure 5.2 is shown the outcome of the individual-based data fitting, in the particular cases of the *Caudatum* and *Starfoot* species, for both the WBE and our model. Although the theoretical curves are almost indistinguishable near the apex, it could be highlighted the deviation from the power-law behavior in the proximity of the stump of the trees, with our theory that accurately fits data points.

5.2.3   WBE formulation

The WBE model predicts the scaling law for the tapering of xylem conduits with the exponent of $1/4$ in the diameter, which becomes $1/2$ for cross sectional areas. Thus, xylem conduits are predicted to taper as follows:

$$\sigma(h) = A \cdot \sqrt{\frac{h}{h_{\max}}} \tag{5.19}$$

where A is the free parameter with the dimensions of a cross sectional area, and $h_{\max}$ corresponds to the experimental heights. By rescaling $\sigma(h)$ and
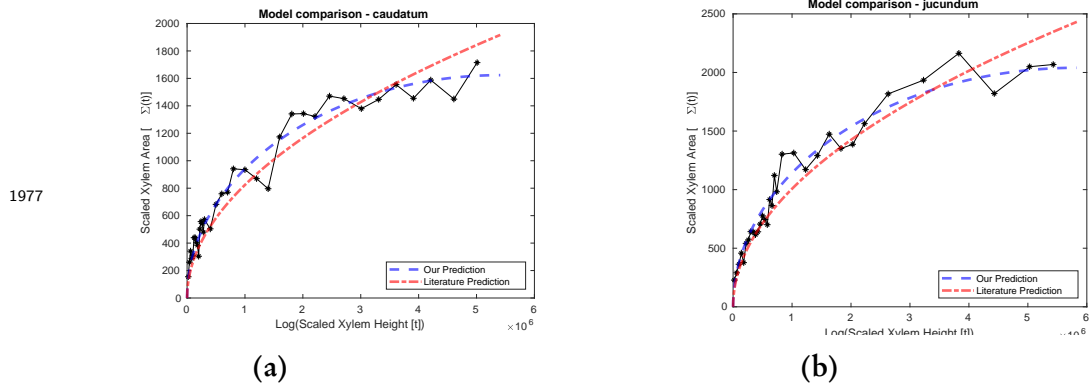
Fig. 5.2 Xylem tapering: A model comparison in individual trees. In panel (a) and (b) we show two examples of angiosperms, the Caudatum and Starfoot trees. Data points are depicted with the black dots, while in dashed-blue and red curves we plot the theoretical curves obtained after fitting with data points respectively the function (Eq 5.12) of our model and the function (Eq. 5.19) of the WBE model.

h as follows:

$$\Sigma(h) = \sigma(h)/A, \tag{5.20}$$

$$t = h/h_{\max}, \tag{5.21}$$

we can rewrite 5.19 in a universal way, independent from any parameter:

$$\Sigma(t) = t^{1/2} \tag{5.22}$$

which becomes a straight line in the space of the t-axis after the log transformation Eq. (5.16).

5.3   Collapse on the universal curve

We hypothesized that xylems, like many other biological and complex systems are characterized by having the key feature of scale invariance ([54], [156],[152]). To test the validity of this hypothesis, we showed that in the space of Σ and t variables, all trees collapse in the single universal curve 5.15. In Figure 5.3 data points of each tree are scaled as follows: $\Sigma(t) = \sigma/\sigma_F$ and $t = h/h_{\max}$. Then, we binned the t-axis in bins with distinct widths, each containing approximately the same number of data points. Figure 5.3 shows that the universal curve of our model 5.15 (the red-dotted line) matches almost perfectly the averages of the cross sectional areas (black bold dots) and their standard deviations (bars), performed for all bins, meaning that the majority of xylem profiles approximately collapse in the average in the universal curve 5.15. We repeated this same analysis for the WBE model, however, Figure 5.3 shows that the universal curve predicted by the WBE model, does not match data points with the same accuracy of our model.

We plotted in log space the two parameters h_{max} and σ_F and studied their relationship (see Figure 5.4).

(a)   (b)
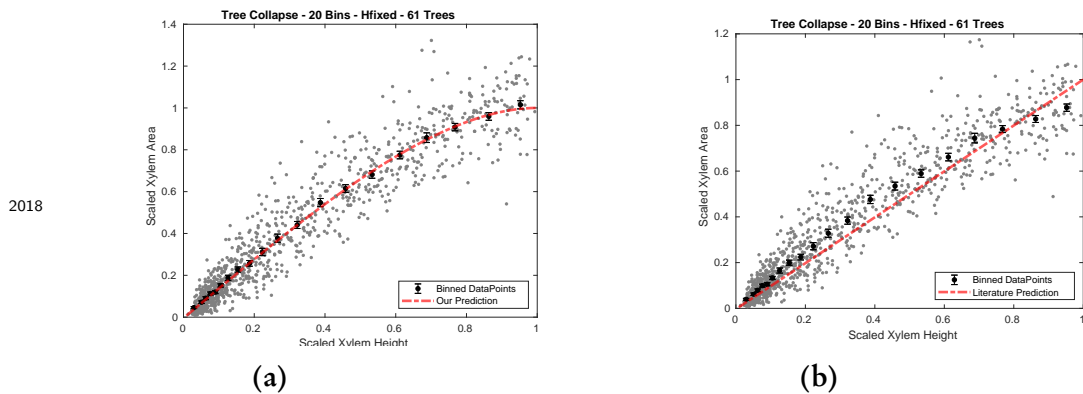
Fig. 5.3 The collapse of 61 trees. In panel (a), we show all the empirical 61 trees in scaled coordinates $\Sigma(t) = \sigma(h)/\sigma_F$ and $t = h/h_{\max}$. For visual purposes, we stretched the height axis with a suitable logarithmic transformation (see Eq. (5.16)), in order to uniformly distribute data points in the whole range $0 < t < 1$. Empirical data are shown in gray dots. We binned the $t-$axis in 20 intervals, each interval containing the same number of data points. Bold dots represent averages, while bars are the standard deviations of the mean. The red dotted line represents the universal curve in the scaled coordinates, as derived theoretically in (Eq. 5.15). In panel (b), we repeated the same analysis described in (a), but employing a polynomial function Eq. (5.19) as derived from the WBE model.
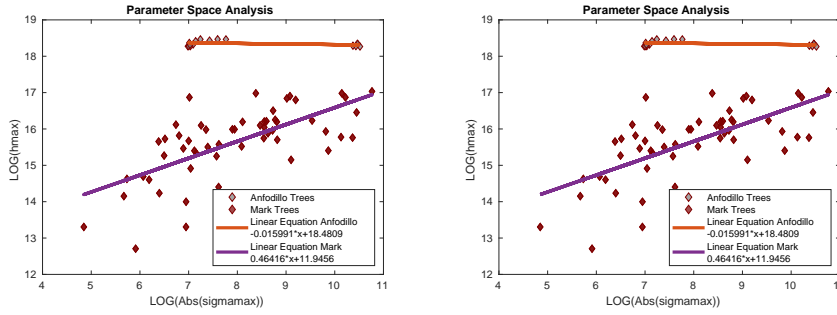
Fig. 5.4 Parameter Space in log space. We studied the parameter distribution in the log space. Linear regression fitting has been done on both distributions resulting in the colored lines.

We rearranged equation 5.12 as follows:

$$\sigma(h) = \sigma_{\text{stump}} \cdot \sqrt{\left(\frac{h}{h_F} \cdot \left(2 - \frac{h}{h_F} \right) \right)} \tag{5.23}$$

and fitted it to data points in order to derive the free parameters h_F which coincides with the total heights. We checked for scale invariance by performing the following transformations: $\Sigma(t) = \frac{\sigma(h)}{\sigma_{stump}}$ and $t = \frac{h}{h_F}$. In the space of $\Sigma(t)$ and t, they fall in the universal curve (in panel **a)** of Figure 5.5). In panel **b)** and **c)** we plotted in log-space the predicted versus empirical heights as derived from our model and WBE. We fitted the distributions with a linear regression before plotting in log space and found that the linear curve has a slope of 0.97 for our model and 0.67 for the WBE model.

Finally, we present an interesting relation in the parameters space of $log(\sigma_F)$ vs $log(h_{max})$ and $log(E)$ vs $log(\alpha)$ (Figure 5.6a, Figure 5.6b). We fitted the distributions with a linear regression.

(a)　　　　　　　　　(b)　　　　　　　　　(c)
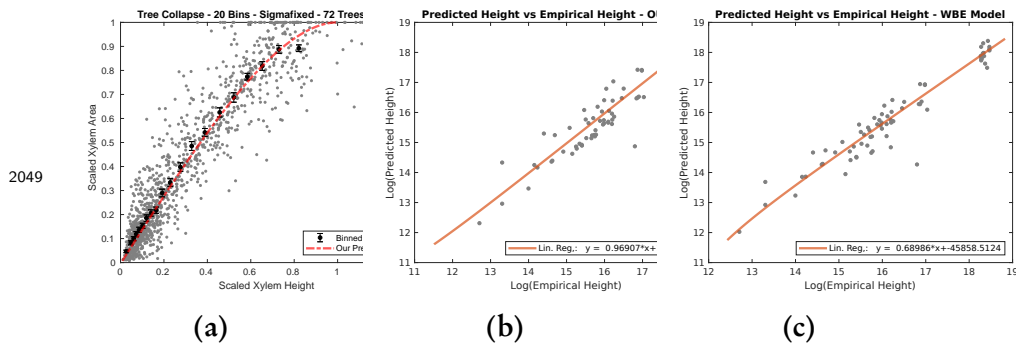
Fig. 5.5 The collapse of 61 trees with fixed $\sigma_{\max}$. We plot the empirical 61 trees in the space of the scaled cross sections $\Sigma(t) = \sigma(h)/\sigma_{max}$ and of the scaled heights $t = h/h_F$. As done in (Fig. 5.4), we stretched the height axis (Eq. 5.16), in order to uniformly distribute data points in the whole range $0 < t < h/h_F$. Empirical measures of the 61 trees are the gray dots in the plot. We binned data points in 20 bins, with each bin containing the same number of data points. Bold dots are averages of $\Sigma(t)$ at each bin, while bars are the standard deviations of the mean. The red dotted line represents the universal curve in the scaled coordinates, as derived theoretically. We have not plot 11 trees, since they don't strictly follow the universal curve. In panels **b)** and **c)** we showed in scatterplot the predicted vs empirical heights for each of the 61 trees, in the log space. Before plotting in log space we found the linear regression as derived from our theory and the WBE models.
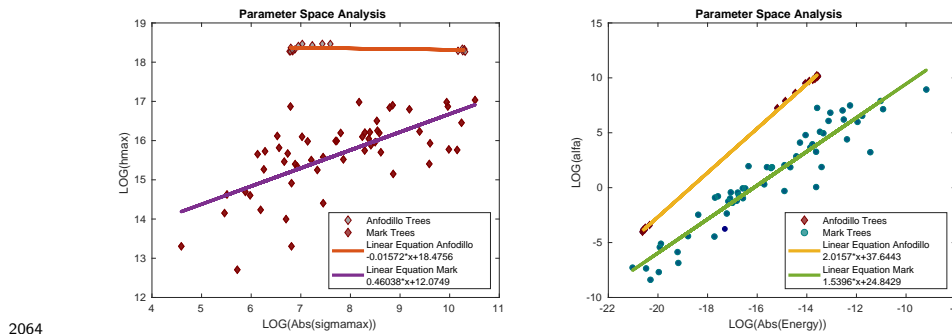
Fig. 5.6 Parameter Space of the WBE parameters in log space.

5.3.1 Statistical Robustness

We performed the statistical robustness of the collapse of our model by increasing the number of bins from 30 to 100 (Figure 5.7). For 100 bins, fluctuations become relevant.
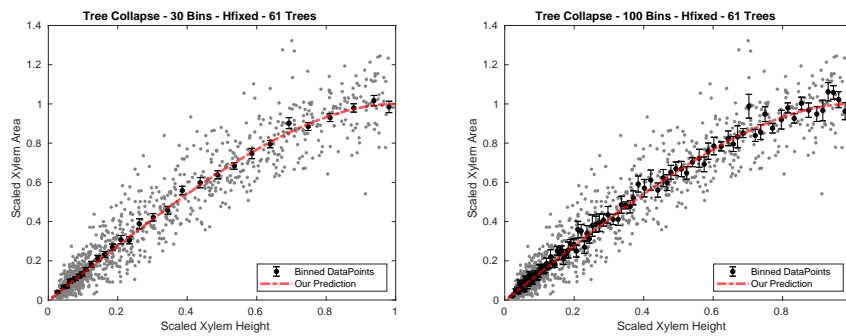
Fig. 5.7 Robustness of the collapse. Here we show the trees' collapse when the number of bins range from 30 to 100.)

5.4   Conclusion and Discussion

In this chapter we addressed the open problem of the tapering of xylem profiles in vascular plants and studied their physical properties within a variational formulation. We modeled this phenomenon with a Lagrangian made up of a Poiseuille resistance term (Eq. 5.2), constrained by a Lagrange multiplier (Eq. 5.5). The Euler-Lagrange solutions lead to a tapered shape for xylems, as observed in several experimental studies. The main result of this investigation rests on the emergence of the scale invariance symmetry of xylems profiles, which greatly simplifies the complexity of the theory with a single universal curve.

As compared to the WBE, our model is able to extract from a very general principle of optimization, the analytical expression of xylems profiles. In addition, data points match more accurately our predictions than the WBE model especially at the stump of the trees, where large deviations between theoretical predictions and empirical data points have been reported. Our model considers only the principle of minimizing the cost of the constrained Poiseuille term, instead of considering additional and unnecessary biological principles.

Based on the Lagrangian formulation, we were able to obtain the tapered structural shape of each individual xylems in the angiosperm dataset. The resulting tapered curves steeply widen near twigs, where it is concentrated the majority of resistance (up to 93% [117]), and then smoothly widen until the basis of the tree (Eq. 5.15) (see Figure 5.3). In the proximity of the twigs, the xylems closely follow the 1/4 polynomial functions, as predicted by the WBE model [157]. Another significant result of our model is related to the correlation of the heights of trees to the cross-sectional areas at the stump of trees. This might extend our possibility to predict tree heights during growth of the trees. Theoretically,

we considered the simplest model by adding only one multiplier term in the Lagrangian, however we could in principle have introduced higher order terms in order to increase the accuracy of our description of the xylem profiles. This analysis is left for further studies. We think that this study paves the way for more biological models that will be able to predict the maximum heights in angiosperms, based on the fact that we have derived the analytical profiles of xylems.

We performed the same statistical robustness for the WBE model by considering the power-law function $\sigma(h) = \sigma_F \sqrt{\frac{2h}{h_{max}}}$. We show in Figure 5.8 the collapse of the fitted trees, as predicted by the WBE model:
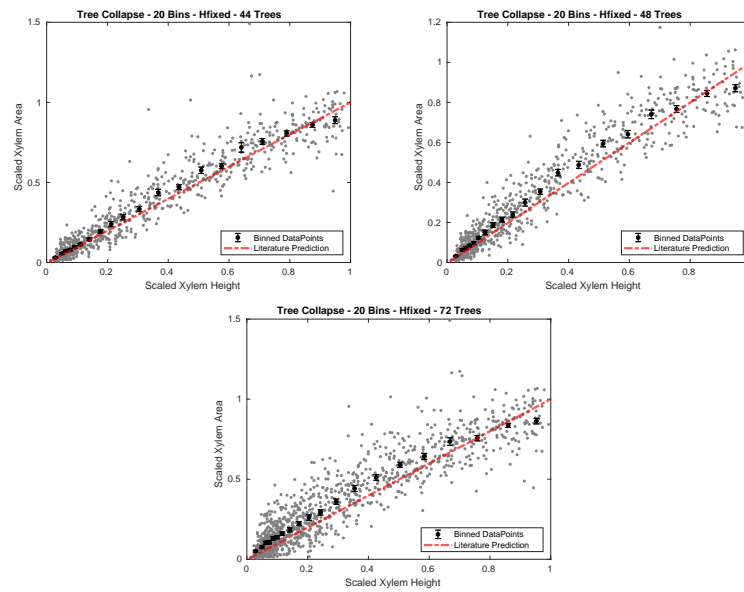
Fig. 5.8 Tree collapse with a power-law. We plot experimental data points in gray and rescale them in the following way: $\Sigma(h) = \sigma(h)/\sigma_F$, $t = h/h_{max}$. We binned data points as before. The red-dotted line represents the WBE polynomial curve in the rescaled space, thus we have that $\Sigma(t) = \sqrt{t}$. In the left figure we plotted only the first 44 trees that best-fit with the power-law 1/4, while in the right figure we chosen to plot the first 48 trees. A deviation from the universal curve becomes clear in this latter case. In the bottom plot we show the collapse of all trees, which present a large deviation from the theoretical prediction, invalidating the power-law 1/4 prediction.

Chapter 6

Conclusions and Perspectives

In this thesis we have quantitatively addressed by means of statistical tools the role of evolutionary selection in shaping the physical traits in biological systems to best adapt their niche.

In the first part of this thesis, we have employed a recently implemented algorithm for studying biological systems, based on the concept of Pareto optimality in competing objective functions. By investigating for signatures of Pareto optimization in the Escherichia coli proteome, we found a triangular-shaped Pareto optimal front by projecting each protein in the space of solubility and hydrophobicity, whose vertices correspond to archetypal proteins specialized in distinct tasks, such as regulatory processes, membrane transport, outer-membrane pore formation, catalysis, and binding. Furthermore, they occupy different subcellular compartments, namely, cytoplasmic, inner membrane, outer membrane, and outer membrane bounded periplasmic space.

In chapter 3, we analyzed the Human Connectome Project (HCP) dataset of cognitive and behavioral scores in 1206 humans through Pareto optimality. When projected in the morphospace of time preferences for reward, which is evaluated with the Delay Discounting Task (DDT), we found a Pareto triangular distribution in which each of the three vertices included individuals who used a particular strategy to discount reward. These archetypes accounted for variability on many cognitive, personality, and socioeconomic status variables, as well as differences in brain structure and functional connectivity, with only a weak influence of genetics. In summary, time preference for reward reflects a core variable that biases human phenotypes via natural and cultural selection. To date, the degree to which biological systems are optimized remains an outstanding problem. Based on these findings and recent literature, it is evident that the Pareto optimality approach is a powerful method to investigate the signatures

of natural selection, and as a prospect, we could adopt this method to unravel further Pareto optimal fronts in biological systems. Possible generalizations regard the application of Pareto optimality in multi-dimensional morphospaces.

Finally, chapter 4 was dedicated to disclose the theoretical mechanism of the well-known tapering phenomenon in the xylem structure in angiosperms. We presented a framework based on the variational formulation with the postulate of minimizing the hydrodynamic resistance cost. The main result of this investigation rests on the emergence of the scale invariance symmetry of xylems profiles, which greatly simplifies the complexity of the theory with a single universal curve. Based on the Lagrangian formulation, we were able to obtain the tapered structural shape of each individual xylem in the angiosperm dataset. The resulting tapered curves steeply widen near twigs, where the majority of resistance is concentrated and then smoothly widen until the basis of the trees. In prospect, this model could be generalized in order to predict the maximum heights in vascular plants.

In summary, we disclosed two Pareto fronts in two very different biological systems which are signatures of multi-objective evolutionary optimization with tradeoffs. Furthermore, we provided a model of the tapering in xylems, which aims to find the best profile which minimized the total energy of the fluid flow.

References

[1] AINSLIE, G. W. Impulse control in pigeons 1. *Journal of the experimental analysis of behavior 21*, 3 (1974), 485–489.

[2] ANOKHIN, A. P., GOLOSHEYKIN, S., GRANT, J. D., AND HEATH, A. C. Heritability of delay discounting in adolescence: a longitudinal twin study. *Behavior genetics 41*, 2 (2011), 175–183.

[3] ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., ET AL. Gene ontology: tool for the unification of biology. *Nature genetics 25*, 1 (2000), 25.

[4] AUSTIN, E. J., DEARY, I. J., WHITEMAN, M. C., FOWKES, F. G. R., PEDERSEN, N. L., RABBITT, P., BENT, N., AND MCINNES, L. Relationships between ability and personality: Does intelligence contribute positively to personal and social adjustment? *Personality and Individual Differences 32*, 8 (2002), 1391–1411.

[5] BALL, P. Water as an active constituent in cell biology. *Chemical reviews 108*, 1 (2008), 74–108.

[6] BATCHELOR, G. K. *An introduction to fluid dynamics*. Cambridge university press, 2000.

[7] BECHARA, A. Decision making, impulse control and loss of willpower to resist drugs: a neurocognitive perspective. *Nature neuroscience 8*, 11 (2005), 1458.

[8] BENJAMINI, Y., AND HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* (1995), 289–300.

[9] BICKEL, W. K., MILLER, M. L., YI, R., KOWAL, B. P., LINDQUIST, D. M., AND PITCOCK, J. A. Behavioral and neuroeconomics of drug

addiction: competing neural systems and temporal discounting processes. *Drug and alcohol dependence 90* (2007), S85–S91.

[10] BICKEL, W. K., AND MUELLER, E. T. Toward the study of trans-disease processes: A novel approach with special reference to the study of co-morbidity. *Journal of dual diagnosis 5*, 2 (2009), 131–138.

[11] BICKEL, W. K., YI, R., LANDES, R. D., HILL, P. F., AND BAXTER, C. Remember the future: working memory training decreases delay discounting among stimulant addicts. *Biological psychiatry 69*, 3 (2011), 260–265.

[12] BIOUCAS-DIAS, J. M. A variable splitting augmented lagrangian approach to linear spectral unmixing. In *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, 2009. WHISPERS'09. First Workshop on* (2009), IEEE, pp. 1–4.

[13] BIOUCAS-DIAS, J. M. A variable splitting augmented lagrangian approach to linear spectral unmixing. In *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, 2009. WHISPERS'09. First Workshop on* (2009), IEEE, pp. 1–4.

[14] BOETTIGER, C. A., MITCHELL, J. M., TAVARES, V. C., ROBERTSON, M., JOSLYN, G., D'ESPOSITO, M., AND FIELDS, H. L. Immediate reward bias in humans: fronto-parietal networks and a role for the catechol-o-methyltransferase 158val/val genotype. *Journal of Neuroscience 27*, 52 (2007), 14383–14391.

[15] BOUSSINESQ, J. Mémoire sur l'influence des frottements dans les mouvements réguliers des fluids. *J. Math. Pures Appl 13*, 377-424 (1868), 21.

[16] BOYD, S., AND VANDENBERGHE, L. *Convex optimization*. Cambridge university press, 2004.

[17] BRANKE, J., BRANKE, J., DEB, K., MIETTINEN, K., AND SLOWIŃSKI, R. *Multiobjective optimization: Interactive and evolutionary approaches*, vol. 5252. Springer Science & Business Media, 2008.

[18] BUTLER, G. S., AND OVERALL, C. M. Proteomic identification of multitasking proteins in unexpected locations complicates drug targeting. *Nature reviews Drug discovery 8*, 12 (2009), 935.

[19] CAMCHONG, J., MACDONALD III, A. W., NELSON, B., BELL, C., MUELLER, B. A., SPECKER, S., AND LIM, K. O. Frontal hyperconnectivity related to discounting and reversal learning in cocaine subjects. *Biological psychiatry 69*, 11 (2011), 1117–1123.

[20] CAMILLONI, C., BONETTI, D., MORRONE, A., GIRI, R., DOBSON, C. M., BRUNORI, M., GIANNI, S., AND VENDRUSCOLO, M. Towards a structural biology of the hydrophobic effect in protein folding. *Scientific reports 6* (2016), 28285.

[21] CARROLL, J. B. *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge University Press, 1993.

[22] CHAN, W. S. Delay discounting and response disinhibition moderate associations between actigraphically measured sleep parameters and body mass index. *Journal of sleep research 26*, 1 (2017), 21–29.

[23] CHANDLER, D. Interfaces and the driving force of hydrophobic assembly. *Nature 437*, 7059 (2005), 640.

[24] CHAO, L.-W., SZREK, H., PEREIRA, N. S., AND PAULY, M. V. Time preference and its relationship with age, health, and survival probability. *Judgment and Decision Making 4*, 1 (2009), 1.

[25] CHESSON, H. W., LEICHLITER, J. S., ZIMET, G. D., ROSENTHAL, S. L., BERNSTEIN, D. I., AND FIFE, K. H. Discount rates and risky sexual behaviors among teenagers and young adults. *Journal of Risk and uncertainty 32*, 3 (2006), 217–230.

[26] CIRYAM, P., TARTAGLIA, G. G., MORIMOTO, R. I., DOBSON, C. M., AND VENDRUSCOLO, M. Widespread aggregation and neurodegenerative diseases are associated with supersaturated proteins. *Cell reports 5*, 3 (2013), 781–790.

[27] COELLO, C. C. Evolutionary multi-objective optimization: a historical view of the field. *IEEE computational intelligence magazine 1*, 1 (2006), 28–36.

[28] COLLETTE, Y., AND SIARRY, P. *Multiobjective optimization: principles and case studies.* Springer Science & Business Media, 2013.

[29] COLOM, R., JUNG, R. E., AND HAIER, R. J. Distributed brain sites for the g-factor of intelligence. *Neuroimage 31*, 3 (2006), 1359–1365.

[30] Comstock, J. P., and Sperry, J. S. Tansley review no. 119 theoretical considerations of optimal conduit length for water transport in vascular plants. *The New Phytologist 148*, 2 (2000), 195–218.

[31] Cona, G., Koçillari, L., Palombit, A., Bertoldo, A., Maritan, A., and Corbetta, M. Archetypes of human cognition defined by time preference for reward and their brain correlates: an evolutionary trade-off approach. *Neuroimage* (2018).

[32] Cornejo, E., Abreu, N., and Komeili, A. Compartmentalization and organelle formation in bacteria. *Current opinion in cell biology 26* (2014), 132–138.

[33] Costafreda, S. G., Brammer, M. J., David, A. S., and Fu, C. H. Predictors of amygdala activation during the processing of emotional stimuli: a meta-analysis of 385 pet and fmri studies. *Brain research reviews 58*, 1 (2008), 57–70.

[34] Coverstone-Carroll, V., Hartmann, J., and Mason, W. Optimal multi-objective low-thrust spacecraft trajectories. *Computer methods in applied mechanics and engineering 186*, 2-4 (2000), 387–402.

[35] Cutello, V., Narzisi, G., and Nicosia, G. A multi-objective evolutionary approach to the protein structure prediction problem. *Journal of The Royal Society Interface 3*, 6 (2006), 139–151.

[36] Cutler, A., and Breiman, L. Archetypal analysis. *Technometrics 36*, 4 (1994), 338–347.

[37] Daly, M., and Wilson, M. Carpe diem: Adaptation and devaluing the future. *The quarterly review of Biology 80*, 1 (2005), 55–60.

[38] De Simone, A., Dhulesia, A., Soldi, G., Vendruscolo, M., Hsu, S.-T. D., Chiti, F., and Dobson, C. M. Experimental free energy surfaces reveal the mechanisms of maintenance of protein solubility. *Proceedings of the National Academy of Sciences 108*, 52 (2011), 21057–21062.

[39] Deary, I. J., Johnson, W., and Houlihan, L. M. Genetic foundations of human intelligence. *Human genetics 126*, 1 (2009), 215–232.

[40] Dekel, E., and Alon, U. Optimality and evolutionary tuning of the expression level of a protein. *Nature 436*, 7050 (2005), 588.

[41] DILL, K. A. Dominant forces in protein folding. *Biochemistry 29*, 31 (1990), 7133–7155.

[42] DRAZIN, P. G., AND RILEY, N. *The Navier-Stokes equations: a classification of flows and exact solutions*. No. 334. Cambridge University Press, 2006.

[43] EDGEWORTH, F. Y., AND PSYCHICS, M. London: C. *Keegan Paul & Co* (1881).

[44] EISENBERG, D. T., MACKILLOP, J., MODI, M., BEAUCHEMIN, J., DANG, D., LISMAN, S. A., LUM, J. K., AND WILSON, D. S. Examining impulsivity as an endophenotype using a behavioral approach: a drd2 taqi a and drd4 48-bp vntr association study. *Behavioral and Brain Functions 3*, 1 (2007), 2.

[45] FISCHL, B., SALAT, D. H., BUSA, E., ALBERT, M., DIETERICH, M., HASELGROVE, C., VAN DER KOUWE, A., KILLIANY, R., KENNEDY, D., KLAVENESS, S., ET AL. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron 33*, 3 (2002), 341–355.

[46] FLOYD, R. G., SHANDS, E. I., RAFAEL, F. A., BERGERON, R., AND MCGREW, K. S. The dependability of general-factor loadings: The effects of factor-extraction methods, test battery composition, test battery size, and their interactions. *Intelligence 37*, 5 (2009), 453–465.

[47] FONSECA, C. M., AND FLEMING, P. J. An overview of evolutionary algorithms in multiobjective optimization. *Evolutionary computation 3*, 1 (1995), 1–16.

[48] FOSSATI, A., BARRATT, E. S., CARRETTA, I., LEONARDI, B., GRAZIOLI, F., AND MAFFEI, C. Predicting borderline and antisocial personality disorder features in nonclinical subjects using measures of impulsivity and aggressiveness. *Psychiatry research 125*, 2 (2004), 161–170.

[49] FRANKENHUIS, W. E., AND DEL GIUDICE, M. When do adaptive developmental mechanisms yield maladaptive outcomes? *Developmental psychology 48*, 3 (2012), 628.

[50] GALLAGHER, T., BJORNESS, T., GREENE, R., YOU, Y.-J., AND AVERY, L. The geometry of locomotive behavioral states in c. elegans. *PloS one 8*, 3 (2013), e59865.

[51] GASS, S., AND SAATY, T. The computational algorithm for the parametric objective function. *Naval research logistics quarterly 2*, 1-2 (1955), 39–45.

[52] GEERLIGS, L., TSVETANOV, K. A., AND HENSON, R. N. Challenges in measuring individual differences in functional connectivity using fmri: the case of healthy aging. *Human brain mapping 38*, 8 (2017), 4125–4156.

[53] GHAHREMANI, D. G., LEE, B., ROBERTSON, C. L., TABIBNIA, G., MORGAN, A. T., DE SHETLER, N., BROWN, A. K., MONTEROSSO, J. R., ARON, A. R., MANDELKERN, M. A., ET AL. Striatal dopamine d2/d3 receptors mediate response inhibition and related activity in frontostriatal neural circuitry in humans. *Journal of Neuroscience 32*, 21 (2012), 7316–7324.

[54] GISIGER, T. Scale invariance in biology: coincidence or footprint of a universal mechanism? *Biological Reviews 76*, 2 (2001), 161–209.

[55] GLASSER, M. F., SOTIROPOULOS, S. N., WILSON, J. A., COALSON, T. S., FISCHL, B., ANDERSSON, J. L., XU, J., JBABDI, S., WEBSTER, M., POLIMENI, J. R., ET AL. The minimal preprocessing pipelines for the human connectome project. *Neuroimage 80* (2013), 105–124.

[56] GLEASON, S. M., BUTLER, D. W., ZIEMIŃSKA, K., WARYSZAK, P., AND WESTOBY, M. Stem xylem conductivity is key to plant water balance across australian angiosperm species. *Functional Ecology 26*, 2 (2012), 343–352.

[57] GOLDSTEIN, R. Z., AND VOLKOW, N. D. Dysfunction of the prefrontal cortex in addiction: neuroimaging findings and clinical implications. *Nature reviews neuroscience 12*, 11 (2011), 652.

[58] GORDON, E. M., LAUMANN, T. O., ADEYEMO, B., HUCKINS, J. F., KELLEY, W. M., AND PETERSEN, S. E. Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cerebral cortex 26*, 1 (2014), 288–303.

[59] GOVINDARAJAN, S., AND AMSTER-CHODER, O. Where are things inside a bacterial cell? *Current opinion in microbiology 33* (2016), 83–90.

[60] GOWDY, J., ROSSER JR, J. B., AND ROY, L. The evolution of hyperbolic discounting: Implications for truly social valuation of the future. *Journal of Economic Behavior & Organization 90* (2013), S94–S104.

[61] GRAHN, J. A., PARKINSON, J. A., AND OWEN, A. M. The cognitive functions of the caudate nucleus. *Progress in neurobiology 86*, 3 (2008), 141–155.

[62] GRANT, P., ABBOTT, I., SCHLUTER, D., CURRY, R., AND ABBOTT, L. Variation in the size and shape of darwin's finches. *Biological Journal of the Linnean Society 25*, 1 (1985), 1–39.

[63] GREEN, L., AND MYERSON, J. A discounting framework for choice with delayed and probabilistic rewards. *Psychological bulletin 130*, 5 (2004), 769.

[64] GREEN, L., MYERSON, J., AND CALVERT, A. L. Pigeons'discounting of probabilistic and delayed reinforcers. *Journal of the Experimental Analysis of Behavior 94*, 2 (2010), 113–123.

[65] GRISKEVICIUS, V., TYBUR, J. M., DELTON, A. W., AND ROBERTSON, T. E. The influence of mortality and socioeconomic status on risk and delayed rewards: a life history theory approach. *Journal of personality and social psychology 100*, 6 (2011), 1015.

[66] HACKE, U. G., AND SPERRY, J. S. Functional and ecological xylem anatomy. *Perspectives in plant ecology, evolution and systematics 4*, 2 (2001), 97–115.

[67] HACKE, U. G., SPERRY, J. S., WHEELER, J. K., AND CASTRO, L. Scaling of angiosperm xylem structure with safety and efficiency. *Tree physiology 26*, 6 (2006), 689–701.

[68] HAIMES, Y. On a bicriterion formulation of the problems of integrated system identification and system optimization. *IEEE transactions on systems, man, and cybernetics 1*, 3 (1971), 296–297.

[69] HART, Y., SHEFTEL, H., HAUSSER, J., SZEKELY, P., BEN-MOSHE, N. B., KOREM, Y., TENDLER, A., MAYO, A. E., AND ALON, U. Inferring biological tasks using pareto analysis of high-dimensional data. *Nature methods 12*, 3 (2015), 233.

[70] HLINKA, J., PALUŠ, M., VEJMELKA, M., MANTINI, D., AND CORBETTA, M. Functional connectivity in resting-state fmri: is linear correlation sufficient? *Neuroimage 54*, 3 (2011), 2218–2225.

[71] HONG, L. E., GU, H., YANG, Y., ROSS, T. J., SALMERON, B. J., BUCHHOLZ, B., THAKER, G. K., AND STEIN, E. A. Association of nicotine

addiction and nicotine's actions with separate cingulate cortex functional circuits. *Archives of general psychiatry 66*, 4 (2009), 431–441.

[72] HUBERT, L., AND ARABIE, P. Comparing partitions journal of classification 2 193–218. *Google Scholar* (1985).

[73] JIN, Y. A comprehensive survey of fitness approximation in evolutionary computation. *Soft computing 9*, 1 (2005), 3–12.

[74] KIRBY, K. N., AND MARAKOVIĆ, N. N. Delay-discounting probabilistic rewards: Rates decrease as amounts increase. *Psychonomic bulletin & review 3*, 1 (1996), 100–104.

[75] KIRBY, K. N., PETRY, N. M., AND BICKEL, W. K. Heroin addicts have higher discount rates for delayed rewards than non-drug-using controls. *Journal of Experimental psychology: general 128*, 1 (1999), 78.

[76] KOÇILLARI, L., FARISELLI, P., TROVATO, A., SENO, F., AND MARITAN, A. Signature of pareto optimization in the escherichia coli proteome. *Scientific reports 8*, 1 (2018), 9141.

[77] KOREM, Y., SZEKELY, P., HART, Y., SHEFTEL, H., HAUSSER, J., MAYO, A., ROTHENBERG, M. E., KALISKY, T., AND ALON, U. Geometry of the gene expression space of individual cells. *PLoS computational biology 11*, 7 (2015), e1004224.

[78] KOZŁOWSKI, J., AND KONARZEWSKI, M. Is west, brown and enquist's model of allometric scaling mathematically correct and biologically relevant? *Functional Ecology 18*, 2 (2004), 283–289.

[79] KOZŁOWSKI, J., AND KONARZEWSKI, M. West, brown and enquist's model of allometric scaling again: the same questions remain. *Functional Ecology 19*, 4 (2005), 739–743.

[80] KUHN, H. Aw tucker, nonlinear programming. In *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability* (1951), pp. 481–492.

[81] KYTE, J., AND DOOLITTLE, R. F. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology 157*, 1 (1982), 105–132.

[82] Lagorio, C. H., and Madden, G. J. Delay discounting of real and hypothetical rewards iii: Steady-state assessments, forced-choice trials, and all real rewards. *Behavioural processes 69*, 2 (2005), 173–187.

[83] Li, N., Ma, N., Liu, Y., He, X.-S., Sun, D.-L., Fu, X.-M., Zhang, X., Han, S., and Zhang, D.-R. Resting-state functional connectivity predicts impulsivity in economic decision-making. *Journal of Neuroscience 33*, 11 (2013), 4886–4895.

[84] Liu, X., Hairston, J., Schrier, M., and Fan, J. Common and distinct networks underlying reward valence and processing stages: a meta-analysis of functional neuroimaging studies. *Neuroscience & Biobehavioral Reviews 35*, 5 (2011), 1219–1236.

[85] MacLean, E. L., Hare, B., Nunn, C. L., Addessi, E., Amici, F., Anderson, R. C., Aureli, F., Baker, J. M., Bania, A. E., Barnard, A. M., et al. The evolution of self-control. *Proceedings of the National Academy of Sciences 111*, 20 (2014), E2140–E2148.

[86] Marcus, D., Harwell, J., Olsen, T., Hodge, M., Glasser, M., Prior, F., Jenkinson, M., Laumann, T., Curtiss, S., and Van Essen, D. Informatics and data mining tools and strategies for the human connectome project. *Frontiers in neuroinformatics 5* (2011), 4.

[87] Marler, R. T., and Arora, J. S. Survey of multi-objective optimization methods for engineering. *Structural and multidisciplinary optimization 26*, 6 (2004), 369–395.

[88] McGhee, G. R. *The geometry of evolution: adaptive landscapes and theoretical morphospaces*. Cambridge University Press, 2006.

[89] Miettinen, K. *Nonlinear multiobjective optimization*, vol. 12. Springer Science & Business Media, 2012.

[90] Miettinen, K. *Nonlinear multiobjective optimization*, vol. 12. Springer Science & Business Media, 2012.

[91] Mobini, S., Grant, A., Kass, A. E., and Yeomans, M. R. Relationships between functional and dysfunctional impulsivity, delay discounting and cognitive distortions. *Personality and Individual Differences 43*, 6 (2007), 1517–1528.

[92] Mørup, M., and Hansen, L. K. Archetypal analysis for machine learning and data mining. *Neurocomputing 80* (2012), 54–63.

[93] Mørup, M., and Hansen, L. K. Archetypal analysis for machine learning and data mining. *Neurocomputing 80* (2012), 54–63.

[94] Myerson, J., Green, L., and Warusawitharana, M. Area under the curve as a measure of discounting. *Journal of the experimental analysis of behavior 76*, 2 (2001), 235–243.

[95] Niwa, T., Ying, B.-W., Saito, K., Jin, W., Takada, S., Ueda, T., and Taguchi, H. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of escherichia coli proteins. *Proceedings of the National Academy of Sciences 106*, 11 (2009), 4201–4206.

[96] Nomi, J. S., and Uddin, L. Q. Developmental changes in large-scale network connectivity in autism. *NeuroImage: Clinical 7* (2015), 732–741.

[97] Norberg, U. M., and Rayner, J. M. Ecological morphology and flight in bats (mammalia; chiroptera): wing adaptations, flight performance, foraging strategy and echolocation. *Phil. Trans. R. Soc. Lond. B 316*, 1179 (1987), 335–427.

[98] Olson, M. E., Anfodillo, T., Rosell, J. A., Petit, G., Crivellaro, A., Isnard, S., León-Gómez, C., Alvarado-Cárdenas, L. O., and Castorena, M. Universal hydraulics of the flowering plants: vessel diameter scales with stem length across angiosperm lineages, habits and climates. *Ecology letters 17*, 8 (2014), 988–997.

[99] Orfanoudaki, G., and Economou, A. Proteome-wide sub-cellular topologies of e. coli polypeptides database (stepdb). *Molecular & Cellular Proteomics* (2014), mcp–O114.

[100] Pareto, V. *Manuale di economia politica*, vol. 13. Societa Editrice, 1906.

[101] Parker, G. A., and Smith, J. M. Optimality theory in evolutionary biology. *Nature 348*, 6296 (1990), 27.

[102] Pessoa, L. On the relationship between emotion and cognition. *Nature reviews neuroscience 9*, 2 (2008), 148.

[103] Peters, J., and Büchel, C. The neural mechanisms of intertemporal decision-making: understanding variability. *Trends in cognitive sciences 15*, 5 (2011), 227–239.

[104] Petit, G., and Anfodillo, T. Plant physiology in theory and practice: an analysis of the wbe model for vascular plants. *Journal of Theoretical Biology 259*, 1 (2009), 1–4.

[105] Petit, G., Anfodillo, T., and De Zan, C. Degree of tapering of xylem conduits in stems and roots of small pinus cembra and larix decidua trees. *Botany 87*, 5 (2009), 501–508.

[106] Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. Methods to detect, characterize, and remove motion artifact in resting state fmri. *Neuroimage 84* (2014), 320–341.

[107] Rachlin, H., and Green, L. Commitment, choice and self-control 1. *Journal of the experimental analysis of behavior 17*, 1 (1972), 15–22.

[108] Raup, D. M. Geometric analysis of shell coiling: general problems. *Journal of Paleontology* (1966), 1178–1190.

[109] Raup, D. M. Geometric analysis of shell coiling: coiling in ammonoids. *Journal of Paleontology* (1967), 43–65.

[110] Reynolds, B. A review of delay-discounting research with humans: relations to drug use and gambling. *Behavioural pharmacology 17*, 8 (2006), 651–667.

[111] Ritchie, S. J., Booth, T., Hernández, M. d. C. V., Corley, J., Maniega, S. M., Gow, A. J., Royle, N. A., Pattie, A., Karama, S., Starr, J. M., et al. Beyond a bigger brain: Multivariable structural brain imaging and intelligence. *Intelligence 51* (2015), 47–56.

[112] Robinson, E. C., Jbabdi, S., Glasser, M. F., Andersson, J., Burgess, G. C., Harms, M. P., Smith, S. M., Van Essen, D. C., and Jenkinson, M. Msm: a new flexible framework for multimodal surface matching. *Neuroimage 100* (2014), 414–426.

[113] Robson, A. J., and Samuelson, L. The evolutionary foundations of preferences. In *Handbook of social economics*, vol. 1. Elsevier, 2011, pp. 221–310.

[114] Rocha, R. P., Koçillari, L., Suweis, S., Corbetta, M., and Maritan, A. Homeostatic plasticity and emergence of functional networks in a whole-brain model at criticality. *Scientific Reports* (2018).

[115] ROGERS, A. R. Evolution of time preference by natural selection. *The American Economic Review* (1994), 460–481.

[116] ROITBERG, B. D., SIRCOM, J., ROITBERG, C. A., MANGEL, M., ET AL. Life expectancy and reproduction. *Nature 364*, 6433 (1993), 108–108.

[117] ROSELL, J. A., OLSON, M. E., AND ANFODILLO, T. Scaling of xylem vessel diameter with plant size: causes, predictions, and outstanding questions. *Current Forestry Reports 3*, 1 (2017), 46–59.

[118] ROSS, I. M. *A primer on Pontryagin's principle in optimal control.* Collegiate publishers, 2015.

[119] ROUNDS, J. S., BECK, J. G., AND GRANT, D. M. Is the delay discounting paradigm useful in understanding social anxiety? *Behaviour Research and Therapy 45*, 4 (2007), 729–735.

[120] RUDNER, D. Z., AND LOSICK, R. Protein subcellular localization in bacteria. *Cold Spring Harbor perspectives in biology* (2010), a000307.

[121] SALIMI-KHORSHIDI, G., DOUAUD, G., BECKMANN, C. F., GLASSER, M. F., GRIFFANTI, L., AND SMITH, S. M. Automatic denoising of functional mri data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage 90* (2014), 449–468.

[122] SANTORO, A., LATORA, V., NICOSIA, G., AND NICOSIA, V. Pareto optimality in multilayer network growth. *Physical Review Letters 121*, 12 (2018), 128302.

[123] SAVAGE, V. M., BENTLEY, L. P., ENQUIST, B. J., SPERRY, J. S., SMITH, D., REICH, P. B., AND VON ALLMEN, E. Hydraulic trade-offs and space filling enable better predictions of vascular structure and function in plants. *Proceedings of the National Academy of Sciences* (2010), 201012194.

[124] SCHAFFER, J. D. Multiple objective optimization with vector evaluated genetic algorithms. In *Proceedings of the First International Conference on Genetic Algorithms and Their Applications, 1985* (1985), Lawrence Erlbaum Associates. Inc., Publishers.

[125] SCHINDEL, D. E. Unoccupied morphospace and the coiled geometry of gastropods: architectural constraint or geometric covariance? *Causes of evolution* (1990), 270–304.

[126] SEOANE, L. F., AND SOLÉ, R. Phase transitions in pareto optimal complex networks. *Physical Review E 92*, 3 (2015), 032807.

[127] SEOANE, L. F., AND SOLÉ, R. The morphospace of language networks. *arXiv preprint arXiv:1803.01934* (2018).

[128] SHEFTEL, H., SHOVAL, O., MAYO, A., AND ALON, U. The geometry of the p areto front in biological phenotype space. *Ecology and evolution 3*, 6 (2013), 1471–1483.

[129] SHEFTEL, H., SZEKELY, P., MAYO, A., SELLA, G., AND ALON, U. Evolutionary trade-offs and the structure of polymorphisms. *Phil. Trans. R. Soc. B 373*, 1747 (2018), 20170105.

[130] SHINOZAKI, K., YODA, K., HOZUMI, K., AND KIRA, T. A quantitative analysis of plant form-the pipe model theory: I. basic analyses. *Japanese Journal of ecology 14*, 3 (1964), 97–105.

[131] SHINOZAKI, K., YODA, K., HOZUMI, K., AND KIRA, T. A quantitative analysis of plant form-the pipe model theory: Ii. further evidence of the theory and its application in forest ecology. *Japanese Journal of Ecology 14*, 4 (1964), 133–139.

[132] SHOVAL, O., SHEFTEL, H., SHINAR, G., HART, Y., RAMOTE, O., MAYO, A., DEKEL, E., KAVANAGH, K., AND ALON, U. Evolutionary trade-offs, pareto optimality, and the geometry of phenotype space. *Science* (2012), 1217405.

[133] SMITH, S. M., MILLER, K. L., SALIMI-KHORSHIDI, G., WEBSTER, M., BECKMANN, C. F., NICHOLS, T. E., RAMSEY, J. D., AND WOOLRICH, M. W. Network modelling methods for fmri. *Neuroimage 54*, 2 (2011), 875–891.

[134] SOHN, S. Y., KANG, J. I., NAMKOONG, K., AND KIM, S. J. Multidimensional measures of impulsivity in obsessive-compulsive disorder: cannot wait and stop. *PLoS One 9*, 11 (2014), e111739.

[135] SOLON, A. P., AND HOROWITZ, J. M. Phase transition in protocols minimizing work fluctuations. *Physical review letters 120*, 18 (2018), 180605.

[136] SPADONE, S., DELLA PENNA, S., SESTIERI, C., BETTI, V., TOSONI, A., PERRUCCI, M. G., ROMANI, G. L., AND CORBETTA, M. Dynamic reorganization of human resting-state networks during visuospatial

attention. *Proceedings of the National Academy of Sciences* (2015), 201415439.

[137] SPEARMAN, C. " general intelligence," objectively determined and measured. *The American Journal of Psychology 15*, 2 (1904), 201–292.

[138] SPERRY, J. S. Evolution of water transport and xylem structure. *International Journal of Plant Sciences 164*, S3 (2003), S115–S127.

[139] STADLER, W., AND DAUER, J. 'multicriteria optimization in engineering: A tutorial and survey'. *Structural optimization: Status and promise 150* (1993).

[140] SUTERA, S. P., AND SKALAK, R. The history of poiseuille's law. *Annual Review of Fluid Mechanics 25*, 1 (1993), 1–20.

[141] SUTHERLAND, M. T., MCHUGH, M. J., PARIYADATH, V., AND STEIN, E. A. Resting state functional connectivity in addiction: lessons learned and a road ahead. *Neuroimage 62*, 4 (2012), 2281–2295.

[142] SZEKELY, P., KOREM, Y., MORAN, U., MAYO, A., AND ALON, U. The mass-longevity triangle: Pareto optimality and the geometry of life-history trait space. *PLoS computational biology 11*, 10 (2015), e1004524.

[143] SZEKELY, P., KOREM, Y., MORAN, U., MAYO, A., AND ALON, U. The mass-longevity triangle: Pareto optimality and the geometry of life-history trait space. *PLoS computational biology 11*, 10 (2015), e1004524.

[144] TARTAGLIA, G. G., PECHMANN, S., DOBSON, C. M., AND VENDRUS-COLO, M. Life on the edge: a link between gene expression levels and aggregation rates of human proteins. *Trends in biochemical sciences 32*, 5 (2007), 204–206.

[145] TARTAGLIA, G. G., PECHMANN, S., DOBSON, C. M., AND VENDR-USCOLO, M. A relationship between mrna expression levels and protein solubility in e. coli. *Journal of molecular biology 388*, 2 (2009), 381–389.

[146] TENDLER, A., MAYO, A., AND ALON, U. Evolutionary tradeoffs, pareto optimality and the morphology of ammonite shells. *BMC systems biology 9*, 1 (2015), 12.

[147] TENDLER, A., MAYO, A., AND ALON, U. Evolutionary tradeoffs, pareto optimality and the morphology of ammonite shells. *BMC systems biology 9*, 1 (2015), 12.

[148] THØGERSEN, J. C., MØRUP, M., DAMKIÆR, S., MOLIN, S., AND JELSBAK, L. Archetypal analysis of diverse pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways. *BMC bioinformatics 14*, 1 (2013), 279.

[149] TOMPKINS, D. A., JOHNSON, P. S., SMITH, M. T., STRAIN, E. C., EDWARDS, R. R., AND JOHNSON, M. W. Temporal preference in individuals reporting chronic pain: discounting of delayed pain-related and monetary outcomes. *Pain 157*, 8 (2016), 1724.

[150] VAN ESSEN, D. C., SMITH, S. M., BARCH, D. M., BEHRENS, T. E., YACOUB, E., UGURBIL, K., CONSORTIUM, W.-M. H., ET AL. The wu-minn human connectome project: an overview. *Neuroimage 80* (2013), 62–79.

[151] VENDRUSCOLO, M., AND DOBSON, C. M. Chemical biology: More charges against aggregation. *Nature 449*, 7162 (2007), 555.

[152] VISWANATHAN, G. M., AFANASYEV, V., BULDYREV, S., MURPHY, E., PRINCE, P., AND STANLEY, H. E. Lévy flight search patterns of wandering albatrosses. *Nature 381*, 6581 (1996), 413.

[153] WARD JR, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association 58*, 301 (1963), 236–244.

[154] WASHBURN, E. W. The dynamics of capillary flow. *Physical review 17*, 3 (1921), 273.

[155] WESLEY, M. J., AND BICKEL, W. K. Remember the future ii: meta-analyses and functional overlap of working memory and delay discounting. *Biological psychiatry 75*, 6 (2014), 435–448.

[156] WEST, G. B., BROWN, J. H., AND ENQUIST, B. J. A general model for the origin of allometric scaling laws in biology. *Science 276*, 5309 (1997), 122–126.

[157] WEST, G. B., BROWN, J. H., AND ENQUIST, B. J. A general model for the structure and allometry of plant vascular systems. *Nature 400*, 6745 (1999), 664.

[158] WILSON, E. O. Caste and division of labor in leaf-cutter ants (hymenoptera: Formicidae: Atta). *Behavioral ecology and sociobiology 7*, 2 (1980), 157–165.

[159] XU, L.-W., YANG, F.-Q., QIN, S., ET AL. A parametric bootstrap approach for two-way anova in presence of possible interactions with unequal variances. *Journal of Multivariate Analysis 115* (2013), 172–180.

[160] YANES, J. A., RIEDEL, M. C., RAY, K. L., KIRKLAND, A. E., BIRD, R. T., BOEVING, E. R., REID, M. A., GONZALEZ, R., ROBINSON, J. L., LAIRD, A. R., ET AL. Neuroimaging meta-analysis of cannabis use studies reveals convergent functional alterations in brain regions supporting cognitive control and reward processing. *Journal of Psychopharmacology 32*, 3 (2018), 283–295.

[161] ZADEH, L. Optimality and non-scalar-valued performance criteria. *IEEE transactions on Automatic Control 8*, 1 (1963), 59–60.

[162] ZAEHLE, S. Effect of height on tree hydraulic conductance incompletely compensated by xylem tapering. *Functional Ecology 19*, 2 (2005), 359–364.

[163] ZASLAVER, A., KAPLAN, S., BREN, A., JINICH, A., MAYO, A., DEKEL, E., ALON, U., AND ITZKOVITZ, S. Invariant distribution of promoter activities in escherichia coli. *PLoS computational biology 5*, 10 (2009), e1000545.

[164] ZIMMERMANN, M. *Xylem structure and the ascent of sap*. Springer series in wood science. Springer-Verlag, 1983.

[165] ZITZLER, E. *Evolutionary algorithms for multiobjective optimization: Methods and applications*, vol. 63. Citeseer, 1999.

[166] ZITZLER, E., AND THIELE, L. Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. *IEEE transactions on Evolutionary Computation 3*, 4 (1999), 257–271.

Appendix A

Pareto fronts identified as convex hulls in the morphospace

In this appendix we provide the full proof of the theorem given in ([132], see Supplementary Materials), that identifies the optimal Pareto front solutions as convex hull of archetypes in the morphospace. This theorem is valid for finite dimensional vector space and denoted by $V : \mathcal{R}^k$, which is endowed with a norm topology such as the locally convex Hausdorff space, and an inner-product norm $||x|| = \sqrt{x \cdot x} \; \forall x \in V$.

Definition 1. Pareto optimal solutions

The Pareto front of a finite subset of V, called X, is the set of points P(X) which are Pareto optimal, namely for each $y \in V, \notin X$ there exists $x_i \in X$ such as $||y - x_i|| > ||x - x_i||$.

Definition 2. Convex hull of X

Convex hull of X are defined as follows:

$$CH(X) = \{x \in V : x = \sum_{n=1}^{M} \alpha_n x_n \geq 0 \; (n = 1, ..., M), \sum_{n=1}^{M} \alpha_n = 1\}$$

(A.1)

Theorem 1. (Hahn-Banach)

Let V be a Hausdorff locally convex topological vector space, and let A and B be two non-empty disjoint closed convex subsets of V with B compact. Then, there exists a continuous linear function $h : V \to \mathcal{R}$ and a number $\gamma \in \mathcal{R}$ such that $h(a) < \gamma, \forall a \in A$ and $h(b) > \gamma, \forall b \in B$.

Theorem 2. (P(X)=CH(X))

Suppose that $x \in CH(X)$ and $x \notin P(X)$. Then, there exists $y \in V, y \neq x$ such that:

$$(y - x_n) \cdot (y - x_n) \leq (x - x_n) \cdot (x - x_n) \ (n = 1, ..., M) \qquad \text{(A.2)}$$

We can rewrite it in the following way:

$$y \cdot y - 2(y - x) \cdot x_n - x \cdot x \leq 0 \ (n = 1, ..., M) \qquad \text{(A.3)}$$

By definition we have that $\alpha_1, ..., \alpha_M, \alpha_n \geq 0 \ (n = 1, ..., M), \sum_{n=1}^{M} \alpha_n = 1$, such that $x = \sum_{n=1}^{M} \alpha_n x_n$. We can straightforwardly obtain what follows:

$$(y - x) \cdot (y - x) \leq 0 \qquad \text{(A.4)}$$

but the inner product of a vector is by definition nonnegative, thus $(y - x) \cdot (y - x) = 0$ and $y = x$, in contradiction with $y \neq x$, and thus $CH(X) \subset P(X)$.

To prove the opposite, suppose that $x \in P(X)$ and $x \notin CH(X)$, meaning that CH(X) and $\{x\}$ are nonempty, disjoint, closed and convex subsets of V. According to the Hahn-Banach theorem there exist $v \in V$ and $\gamma \in R$ such that:

$$v \cdot x > \gamma \qquad \text{(A.5)}$$

and

$$v \cdot x_n < \gamma \ (n = 1, ..., M) \qquad \text{(A.6)}$$

and thus:

$$v \cdot (x - x_n) > 0 \ (n = 1, ..., M) \qquad \text{(A.7)}$$

We make the following definition:

$$t = min_{x_n \in X} \frac{2v \cdot (x - x_n)}{v \cdot v} > 0 \qquad \text{(A.8)}$$

Then there exists $x_i \in X$ such that:

$$(x - tv - x_i) \cdot (x - tv - x_i) > (x - x_i) \cdot (x - x_i) \qquad \text{(A.9)}$$

and obtain a contradiction:

$$t > \frac{2v \cdot (x - x_i)}{v \cdot v} \geq t \qquad \text{(A.10)}$$

2728 implying that $x \notin CH(X)$ is false. Thus we have that $P(X) \subset CH(X)$.

Appendix B

Non-circular pipe flow constant

In this appendix we show some case of non-circular pipes that has been investigated in [42]. In case when the xylem cross section is elliptic $x^2/a^2 + y^2/b^2 = 1$, it can be shown that the volumetric flow rate has the following expression ([42], [15], [6]):

$$Q = \frac{\pi P a^3 b^3}{4\mu(a^2 + b^2)} \tag{B.1}$$

which becomes a Poiseuille volumetric flow if $a = b$.

Further examples have been studied in ([42], [15]) such as pipes with an equilateral triangular cross section, with a flow rate given by:

$$Q = \frac{Gh^4}{60\sqrt{3}\mu} \tag{B.2}$$

where $2h/\sqrt{3}$ is the side length of the cross section.

In case of a rectangular channel of height h and width l we have the volumetric flow rate given as ([42],[15]):

$$Q = \frac{Gh^3 l}{12\mu} - \frac{16Gh^4}{\pi^5\mu} \sum_{n=1}^{\infty} \frac{1}{(2n-1)^5} \tag{B.3}$$

Other different cross sectional shapes have been considered, however all these generalized formulas show that the volumetric flow rate is proportional to the fourth power of the cross sectional radius $Q \propto R^4$, where the constant depends on the shape of the cross section.

Appendix C

WBE model (West et al. 1999)

The WBE model is based on the four axioms:

1. the branching pattern follows a space-filling mechanism [156], which ensures biologically that all leaves are serviced by capillaries,

2. terminal elements are size-invariant, meaning that the capillary density in a cross sectional area remains constant across levels,

3. the total hydrodynamic resistance is minimized,

4. the bio-mechanical constraints are uniform, which assures that at each level branches split in smaller ones whose area sums to the original one.

Based on these axioms, it can be derived a continuously branching network for xylems, going from roots to leaves, which is structured with k successive levels of branching, with a bundle of parallel and identical cylindrical pipes at each level.

C.0.1 Notation

We define the *branch radii* as:

$$\beta_k \equiv \frac{r_{k+1}}{r_k} \equiv n^{-a/2} \tag{C.1}$$

the *tube radii* as:

$$\bar{\beta}_k = \frac{a_{k+1}}{a_k} = n^{-\bar{a}/2} \tag{C.2}$$

and the *branch lengths* as:

$$\gamma_k \equiv \frac{l_{k+1}}{l_k} \tag{C.3}$$

where n is the branching ratio, defined also as the number of daughter branches as derived from a parent branch.

In [157] authors consider that the total number of tubes is preserved at each branching, thus n is taken independent of k. This condition has been generalized in Savage et al [123].

C.0.2 Derivation of the $1/4$ tapering exponent

The first axiom of volume-filling states that [156]:

$$\gamma_k = n^{-1/3} \tag{C.4}$$

The total number of tubes is preserved at each branching so $n = \frac{n_{k+1}}{n_k}$ is independent of the k-level and we also have that $n_k = n_N n^{N-k}$, where N is the total number of branching generations from roots to leaves.

Authors assumed that xylem tapering is constant across levels, meaning that $\bar{a}$ is independent of k and thus the tube radius scales as:

$$\frac{a_k}{a_N} = \left(\frac{r_k}{r_N}\right)^{\bar{a}/a} \tag{C.5}$$

and the branch lengths as follows:

$$\frac{l_k}{l_N} = \left(\frac{r_k}{r_N}\right)^{2/3a} \tag{C.6}$$

From the area-preserving as derived from the bio-mechanical axiom, we have that $a = 1$.

The resistance R_k^i of a single xylem i within the branch segment k is given by the Hagen-Poiseuille law:

$$R_k^i = \frac{8\mu l_k}{a_k^4} \tag{C.7}$$

and the total resistance of a given xylem along the whole path is the sum of all k contributing branch segments. By substituting our notation for l_k and a_k we have the following relation: where $l_T = \sum_{k=0}^{N} l_k = l_0/(1 - n^{-1/3})$. When $l_T \gg l_N$, R_i will depend mostly on the degree of tapering, that is for $\bar{a} < 1/6$ the resistance increases with path length l_T, while for $\bar{a} > 1/6$ the resistance reaches a minimum, and is a constant, independent from the total xylem height l_T. By choosing $\bar{a} = 1/6$, which minimizes xylem

tapering, we have that the tube radius scales as:

$$\frac{a_k}{a_N} = \left(\frac{r_k}{r_N}\right)^{1/6} \tag{C.8}$$

and by combining it to the branch lengths

$$\frac{l_k}{l_N} = \left(\frac{r_k}{r_N}\right)^{2/3} \tag{C.9}$$

we derive the classic $a_k/a_N \propto (l_k/l_N)^{1/4}$ relation.