On the molecular organization of a succinyl-CoA-producing cell-free system: A cryo-EM and computational approach.

Dissertation

for the degree of
Doctor of Natural Sciences (Dr. rer. Nat.)

Submitted to the
Faculty of Natural Sciences I – Biosciences –
of Martin Luther University,
Halle-Wittenberg

Presented by:
Ioannis Skalidis
Born on 17.09.1991 in Athens (Greece)

Defended on 16.06.2023

Assessors:
Jun. Prof. Dr. Panagiotis L. Kastritis
Prof. Dr. Milton Stubbs
Prof. Dr. Robert Tampé

To my mother, my sister
and to Linda.

Table of contents

Table of figures

List of tables

List of abbreviations

ABC	Ammonium Bicarbonate
Acetyl-CoA	Acetyl-coenzyme A
ACN	Acetonitrile
ACP	Acyl Carrier Protein
ACS	Acetyl-CoA synthase
ADP	Adenosine Diphosphate
AGC	Automatic Gain Control
AI	Artificial Intelligence
AKG	α-ketoglutarate
AKGDD	α-ketoglutarate-dependent Dioxygenase
AKT	Protein kinase B
ANOVA	Analysis of Variance
APK	AMP-activated Protein Kinase
ATP	Adenosine Triphosphate
BCKDHc	Branched-chain Ketoacid Dehydrogenase Complex
BSA	Bovine Serum Albumin
BSA	Buried Surface Area
C-ter	C-terminal
CaMKII	Ca^{2+}/Calmodulin-dependent Protein Kinase II
CAPRI	Critical Assessment of Predicted Interactions
CASP	Critical Assessment of Methods for Protein Prediction
CC	Cross-Correlation
CCD	Charge Coupled Device
CCM	Complete Culture Media
CFS	Cell-free System
CLSM	Confocal Laser Scanning Microscopy
CNN	Convolutional Neural Network
CoA	Coenzyme A
cryo-EM	Cryogenic Electron Microscopy
cryo-TEM	cryo-Transmission Electron Microscope

C_s	Spherical Aberration
CTF	Contrast Transfer Function
D2	Dihedral 2
Da	Dalton
DED	Direct Electron Detector
DH	Dehydratase
DL	Deep Learning
DNA	Deoxyribonucleic Acid
DNL	*de novo* Lipogenesis
DQE	Detective Quantum Efficiency
DS	Desolvation Energy
DTT	Dithiothreitol
E1o	2-oxoglutarate Dehydrogenase
E1p	Pyruvate Dehydrogenase (lipoamide)
E2b	Dihydrolipoamide Acyltransferase
E2o	Dihydrolipoyl Succinyltransferase
E2p	Dihydrolipoyl Acetyltransferase
E3	Dihydrolipoyl Dehydrogenase
E3BP	E3 Binding Protein
E3BPo	E3 Binding Protein of oxoglutarate dehydrogenase complex
EM	Electron Microscopy
EMDB	Electron Microscopy Data Bank
ER	NADPH-dependent β-enoyl Reductase
ERK	extracellular signal-regulated kinase
ES	Electrostatics
FAS	Fatty Acid Synthasae
FDR	False-discovery Rate
FEG	Field Emission Gun
FFNN	Feed-Forward Neural Network
FPLC	Fast Performance Liquid Chromatography
FSC	Fourier Shell Correlation
GO	Gene Ontology

GPCR	G Protein-coupled Receptor
GS	Gold-Standard
HCCS	Holocytochrome C Synthase
HDX-MS	Hydrogen Exchange Mass Spectrometry
HIF-1α	Hypoxia-inducible Factor 1 alpha
HIF-1β	Hypoxia-inducible Factor 1 beta
HRE	Hypoxia-related Element
I	Icosahedral
IAA	2-iodoacetamide
IDP	Intrinsically Disordered Protein
IDR	Intrinsically Disordered Region
IRS-1	Insulin Receptor Substrate-1
JNK	c-Jun N-terminal kinase
KAT	Lysine Acetyl-transferase
KDAC	Lysine Deacetylase
KR	β-ketoacyl Reductase
KS	β-ketoacyl Synthase
LC-MS/MS	Liquid Chromatography - Tandem Mass Spectrometry
LD	Lipoyl-binding Domain
Malonyl-CoA	Malonyl-coenzyme A
MAPK	Mitogen-activated protein kinase
MAT	malonyl-/acetyl-CoA-ACP Transacylase
MDa	Megadalton
ML	Machine Learning
MS	Mass Spectrometry
MTF	Modular Transfer Function
mTOR	mechanistic Target of Rapamycin
mTORC1	mechanistic Target of Rapamycin Complex 1
MW	Molecular Weight
N-ter	N-terminal
NCS	Non-Crystallographic Symmetry

NF-κB	Nuclear factor kappa-light-chain-enhancer of activated B cells
NMR	Nuclear Magnetic Resonance
NS-EM	Negative Staining Electron Microscopy
O	Octahedral
OGDHc	2-oxoglutarate Dehydrogenase Complex
PA	Palmitic acid
PAE	Predicted Align Error
PAGE	Polyacrylamide Gel Electrophoresis
PBS	Phosphate Buffer Saline
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PDHc	Pyruvate Dehydrogenase Complex
PHD	Prolyl-hydroxylase
PI3K	Phosphoinositide 3-kinase
PKC	Protein Kinase C
pLDDT	Predicted Local Distance Difference Test
PPI	Protein-Protein Interaction
PPT	phosphopantetheinyl transferase
PSBD	Peripheral Subunit Binding Domain
PTEN	Phosphatase and tensin homolog
PTM	Post-Translational Protein Modification
RNA	Ribonucleic acid
RNN	Recurrent Neural Network
ROS	Reactive Oxygen Species
RT	Room Temperature
SCP	Sodium Cacodylate
SDS	Sodium Dodecyl Sulfate
SEC	Size-exclusion Chromatography
SNR	Signal to Noise Ratio
SPA	Single Particle Analysis
STAGE	Stop And Go Extraction

Succinyl-CoA	Succinyl-coenzyme A
TCA	Tri-carboxylic Acid
TE	Thioesterase
TEM	Transmission Electron Microscopy
ThDP	Thiamine Diphosphate
TOR	Target of Rapamycin
TSC2	Tuberous Sclerosis Complex 2
ULK1	Unc-51 Like Autophagy Activating Kinase 1
vdW	Van der Waals
VHL	Von Hippel-Lindau Factor
WB	Western Blotting
XL-MS	Crosslinking Mass Spectrometry
XRD	X-ray Diffraction

1 Introduction

1.1 The overlooked signaling functions of metabolites

One of the main goals of modern biological research has been the discovery of disease drivers and more specifically, the molecular mechanisms that underly pathological phenomena such as malignancies. A plethora of techniques have been employed towards this goal, starting from genome sequencing and gene knockouts and recently moving towards -omics approaches, which the advent of network biology has bolstered. These techniques have revealed the basic molecular pathway framework that, when disrupted by numerous internal or external factors, can result in pathological conditions. As a result, a general paradigm has been established: the cell will employ either intracellular or extracellular sensor proteins to quantify environmental stimuli. Thus, a signaling cascade will begin, primarily through protein-protein interactions (PPI), until it reaches the nucleus[1]. The cascading signal will alter gene expression to respond and adapt to the triggering stimulus. Any disruption of this process will most likely lead to a non-canonical state for the cell, resulting in pathological phenotypes of varying severity.

Recently, however, research has been broadening its scope to include a facet of cell signaling that had gone unheeded. Apart from the classic protein-protein interaction networks that propagate cellular signaling, the products of the cell's metabolism can also participate in signaling networks. Observations on the effect of metabolites in a cell go as far back as the 1950s, when it was shown that the cellular concentration equilibrium between glucose and lactose could alter gene expression[2] or how post-translational modifications of proteins can regulate their function[3]. Currently, cellular signaling and metabolism are viewed as highly intertwined[4,5] with metabolic networks displaying a degree of complexity comparable to their PPI signaling network counterparts. Metabolites are not only used by the cell for energy production and storage or as building blocks of critical cellular components (*e.g.,* proteins, nucleic acids, lipids, and sugars) but the variety of substrates, intermediates, and products that participate in the metabolic networks can also affect substantial

change to cellular fate and behavior. Below, three different examples of metabolites will be examined, namely acetyl-coenzyme A (Acetyl-CoA), α-ketoglutarate (AKG), and palmitic acid (PA), along with the corresponding enzymatic complexes that regulate their availability.

1.2 Acetyl-CoA, α-ketoglutarate and palmitic acid and their role in metabolic signaling

1.2.1 Acetyl-CoA

Acetyl-coenzyme A can genuinely be described as a cornerstone of a cell's metabolic pathways (**Figure 1**). It is a central component of a plethora of metabolic pathways, is involved in multiple reactions, and participates in many cellular non-metabolic processes, with its involvement as a substrate for protein acetylation being a prime example[6,7]. Pyruvate is first catalyzed into acetyl-CoA to enter the tri-carboxylic acid (TCA) cycle. Acetyl-CoA is also one of the lipid synthesis pathway's precursors with malonyl-coenzyme A (Malonyl-CoA)[8]. Two main chemical groups comprise Acetyl-CoA: (a) an acetyl- group and (b) a coenzyme A are connected via a thioester bond. In turn, coenzyme A is the result of a 3'-5' adenosine diphosphate (ADP) group connected to β-mercaptoethylamine and pantothenic acid. The largest intracellular source of Acetyl-CoA is through the decarboxylation of pyruvate by the pyruvate dehydrogenase complex (PDHc). PDHc is the enzymatic complex that transforms the pyruvate coming from glycolysis into Acetyl-CoA and then feeds it into the TCA cycle for energy production through the formation of adenosine triphosphate (ATP).

It is essential to highlight that Acetyl-CoA levels in the cell are under tight control through various homeostatic mechanisms, as Acetyl-CoA availability substantially impacts cell fate via multiple mechanisms. Significant reductions of Acetyl-CoA levels in the cytosol have been shown to activate autophagy pathways[9]. When Acetyl-CoA is depleted, the AMP-activated protein kinase (APK) is activated, leading to further activation of the Unc-51 Like Autophagy Activating Kinase 1 (ULK1). ULK1 acts as a

phosphatase, with many of its downstream targets being active participants in the formation of the autophagosome[10]. Apart from ULK1 activation, when cytosolic Acetyl-CoA concentrations lower, the mechanistic Target of Rapamycin Complex 1 (mTORC1) is additionally inhibited. During typical Acetyl-CoA concentration conditions, mTORC1 inhibits URK1, thus inhibiting mTOR signaling and preventing the activation of autophagic pathways[10]. The cytosolic concentration balance between Acetyl-CoA and coA is another parameter in the regulation of cell death. When the ratio between the two remains high, an Acetyl-CoA-dependent signaling cascade begins, activating Ca^{2+}/calmodulin-dependent protein kinase II (CaMKII). CaMKII is the main phosphatase that activates Caspase-2, a lynchpin protein of many anti-apoptotic pathways in the cell[11].

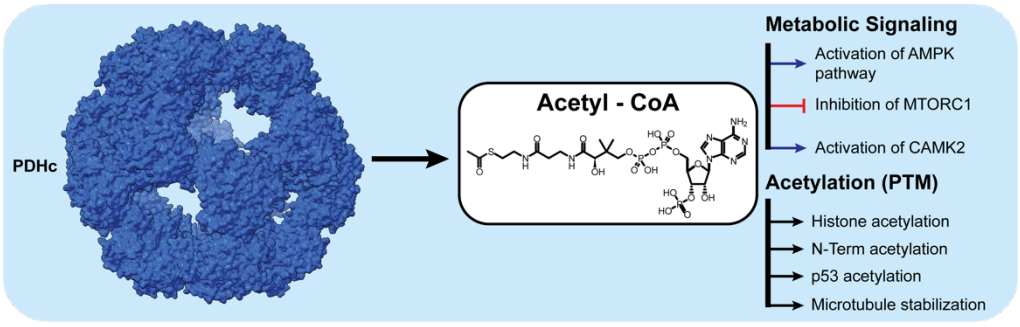


Figure 1: Acetyl-coenzyme A is a central metabolite of the cell's bioenergetic pathways.
Acetyl-CoA is produced by the pyruvate dehydrogenase complex (PDHc (EMD-7610)) and affects multiple signaling pathways through different mechanisms. Here, the icosahedral core of PDHc is shown for simplicity. Figure reproduced from[12].

Protein acetylation, either during translation (*N-ter* acetylation) or purely as a post-translational protein modification (PTM), relies on the usage of acetyl- groups donated by Acetyl-CoA. When a protein is acetylated at its *N-ter* during translation, an acetyl- group is covalently attached to the first *N-ter* residue's amino- group (usually a serine, threonine, valine, cysteine or alanine) after the cleavage of the methionine. Acetylation of the *N-ter* has significant implications for a protein's function, localization, and stability[13]. During PTM, the main residues that are targeted for acetylation are the amino- groups of lysine residues. Lysine acetylation has been shown to have multiple implications on a protein's fate, and its effects on protein catalytic activity, localization, stability, and PPIs have been widely studied[14].

Protein acetylation is broken down into two main processes: (a) the addition of the acetyl- group to a target Lys by a class of enzymes known as lysine acetyl-transferases (KATs) and (b) the removal of the acetyl- group from an already acetylated lysine by the class of enzymes known as lysine deacetylases (KDACs). Histone modification, a process where different chromatin areas become accessible or inaccessible to the transcription apparatus, is mediated by KATs/HDACs[15]. This process leads to the regulation and altering of a cell's epigenetic profile, which, in turn, defines gene expression as a response to environmental changes[16]. Tubulin subunits, the building blocks of the cell's microtubule skeleton, represent another intriguing acetylation target. In general, all microtubule PTM sites are located on their outer surface, with the only exception being the α-tubulin's Lys40 acetylation which happens in the microtubule lumen[17]. Researchers have been aware of this specific modification for quite some time, but its importance has only recently been further investigated, showing its implications in autophagy, cell migration[17], and intracellular protein trafficking[18]. Finally, protein acetylation is directly implicated in anti-oncogenic processes through its effect on p53, the protein otherwise known as the "guardian of the genome". Histone modification (as described above) regulates the transcriptional activity of the p53 gene, and the p53 protein itself is further activated by being acetylated at lysine residues of its *N-ter* domain, thus inducing transcription of other downstream genes with anti-oncogenic effect[19].

1.2.2 α-ketoglutarate

Formerly known as 2-oxoglutaric acid, α-ketoglutaric acid (α-ketoglutarate/AKG) is a critical intermediate product of the TCA cycle. Its availability is controlled by the 2-oxoglutarate dehydrogenase complex (OGDHc), as it is OGDHc that converts it through decarboxylation of AKG into succinyl-coenzyme A (Succinyl-CoA). The conversion of AKG to Succinyl-CoA is a rate-determining step of the TCA cycle, as glutamate can act as a substrate for anaplerotic reactions that will upregulate the intracellular concentration of AKG[20]. Additionally, apart from its role in energy production, AKG plays a large part in biosynthetic pathways by acting as a substrate for the production of proline, leucine, glutamine, and glutamate, with the latter two,

other than their role as protein building blocks, also displaying a significant role in multiple metabolic and signaling processes[21,22].

AKG displays metabolic signaling activity through two main signaling pathways (**Figure 2**). The first is related to tumorigenesis through the cell's hypoxia sensing and response pathway. AKG is the primary substrate of a class of α-ketoglutarate-dependent dioxygenases (AKGDD), the prolyl-hydroxylases (PHD). PHDs hydroxylate the hypoxia-inducible factor 1 alpha (HIF-1α), using AKG and oxygen as substrates, also producing succinate during the process[23]. HIF-1α and hypoxia-inducible factor 1 beta (HIF-1β) are the two subunits of the HIF-1 transcription factor, which controls the expression of oxygen-sensing response related genes[24]. When oxygen levels are low, a condition called hypoxia, HIF-1 promotes glycolysis-related gene transcription, which ensures that ATP continues to be generated even in the absence of oxygen when oxidative phosphorylation is ineffective. Cancer cells, which undergo rapid propagation, often have to survive in hypoxic environments and heavily utilize this pathway to accommodate their need for vast quantities of ATP[25]. In non-hypoxic conditions, high levels of AKG in the cytosol will activate the prolyl hydroxylase 1-3, leading to the hydroxylation of HIF-1α. This modification will allow the Von Hippel-Lindau factor (VHL) to recognize and bind to HIF-1α, designating the protein as a target for proteasome-mediated degradation[26]. A cancer cell will leverage this mechanism through the reduction of cytosolic AKG concentrations. Reduced AKG levels will inhibit AKGDDs, leading to the stabilization of HIF-1α which in turn will be transported to the nucleus and, through its dimerization with HIF-1β, will lead to the promotion of hypoxia-related elements (HRE) transcription[25], ensuring cell survival and proliferation.

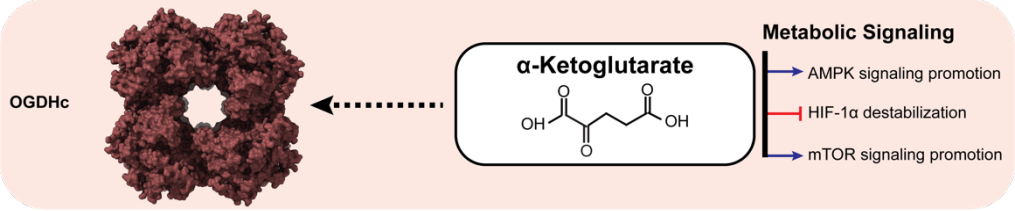


Figure 2: **α-ketoglutarate is directly implicated in metabolic pathways.**
AKG levels are regulated by 2-oxoglutarate dehydrogenase complex (OGDHc, Electron Microscopy Data Bank ID, EMD-0108). The cubic core of OGDHc is shown for simplicity. Figure reproduced from[12].

The second pathway through which AKG displays its metabolic signaling capabilities is the Target of Rapamycin (TOR) signaling pathway[27] where its anti-aging effect is observed. In *C. elegans* models, AKG can delay phenotype appearance related to aging, thus increasing their lifespan. This effect is mediated through the ATP synthase, a membrane-embedded protein complex of significant importance to the cells due to its involvement in energy metabolism. ATP synthase has been identified as a novel AKG binding target, with an inhibitory effect. Inhibition of ATP synthase reduces the production of ATP and consequently oxygen consumption, which in turn activates autophagy[27]. The reduction in produced ATP also affects the ATP/ADP ratio in the cell, triggering the phosphorylation of TOR suppressor Tuberous Sclerosis Complex 2 (TSC2) by the APK, and further strengthens autophagy signaling in the cell[28].

1.2.3 Palmitic acid

Palmitic acid (PA) is the human body's most common saturated fatty acid (**Figure 3**). It is synthesized endogenously via the *de novo* lipogenesis (DNL) metabolic pathway. In DNL, during a continuous synthesis spiral, Malonyl-CoA is elongated to a final length of 16 (16:0) carbons. The main carbon source for the synthesis is Acetyl-CoA, and the process is controlled by the key enzyme fatty acid synthase (FAS)[29]. Through tight homeostatic mechanisms, PA synthesis via DNL is kept in balance with extracellular intake, suggesting that PA is critical for a plethora of cellular processes[30]. Deregulation of intracellular PA levels often leads to pathological phenotypes, with the most striking example being its connection to what has been described as the Warburg effect[31], a critical characteristic of multiple cancer cell phenotypes. The Warburg effect describes a cancer cell's switch to anaerobic glycolysis pathways and leverage of DNL before there is a need for such an adaptation due to reduced oxygenation through the lack of local tissue hematosis.

The switch to anaerobic glycolysis pathway utilization by the cancer cells creates an intracellular PA excess and leads to increased diacylglycerol levels, activating the protein kinase C (PKC). PKC, in turn, phosphorylates the insulin receptor substrate-1 (IRS-1) and reduces its activation, which leads to Phosphoinositide 3-kinase/ Protein

kinase B (PI3K/AKT) signaling pathway inhibition[32]. The PI3K/AKT pathway is one of the main pathways involved in cell cycle regulation, and its de-activation can affect cell proliferation, quiescence, and the onset of oncogenic phenotypes. The AKT signaling pathway is further influenced by intracellular PA levels via a p38-mediated activation of the phosphatase and tensin homolog (PTEN) tumor suppressor[32]. Additionally, increased intracellular PA concentration in adipose tissue facilitates inflammation responses and the onset of related diseases such as insulin resistance-related obesity.

Multiple key pathways related to inflammation responses are activated by PA, with prime examples being the nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB), mitogen-activated protein kinase (MAPK), and PKC-mediated pathways. Activation of these pathways leads to increased cytokine production, including tumor necrosis factor (TNF) and interleukin-10, resulting in a perpetual inflammatory state[33]. The inflammatory response is further strengthened by the induced phosphorylation of c-Jun N-terminal/extracellular signal-regulated kinases (JNK/ERK), key participants in the MAPK signaling pathway, and can lead to severe pathological conditions related to circulation, due to the onset of metabolic syndrome[34]. Intriguingly, it has been observed that in hepatocellular carcinoma, increased concentrations of PA display anti-tumorigenic effects, by downregulating the mechanistic target of rapamycin (mTOR) signaling pathway, thus reducing cancer invasiveness and proliferation[35].

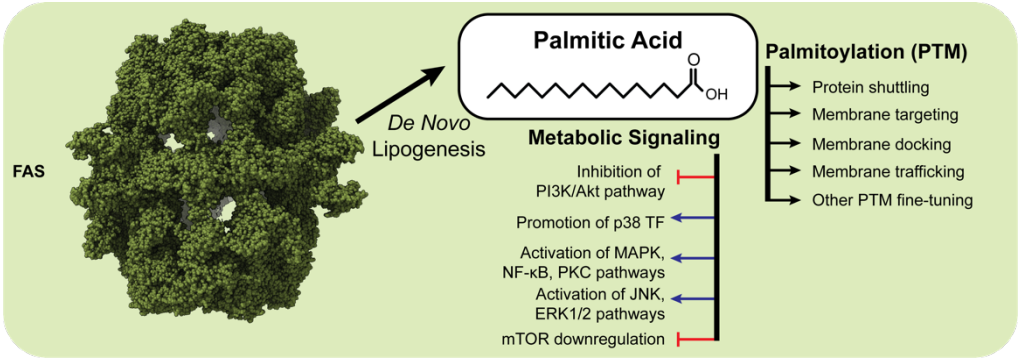


Figure 3: **Palmitic acid is a main product of *de novo* lipogenesis.**
DNL is carried out by the fatty acid synthase (FAS, (EMD-4577)) and PA can influence different signaling targets, directly or through post-translational modifications. The structure of FAS is shown for simplicity. Figure reproduced from[12].

Apart from its participation in cellular signaling pathways, PA is the main substrate of palmitoylation, another form of protein PTM. When a protein is palmitoylated, a palmitate group is connected to a cysteine with a thioester bond (S-palmitoylation) or, even more rarely, to a serine or threonine (O-palmitoylation). Palmitoylation is a dynamic process and can be reversed[36]. The dynamic character of this PTM is probably the reason for its widespread functionality, including, among others, the modulation of protein membrane trafficking and docking[37]. A palmitate group, when added to a soluble protein, will provide it with additional hydrophobic properties and act as an anchor that will allow it to more easily dock on a membrane. This can work for non-soluble, transmembrane proteins as well, for example, G protein-coupled receptors (GPCR), as the increased hydrophobicity through the addition of the palmitate group will provide increased stability, prevent their aggregation until their placement in the membrane, even provide the means to increase binding specificity to a specific membrane and its components[38]. As is the case with other PTMs, palmitoylation also can modulate protein trafficking amongst cellular compartments, such as the transfer of Ras proteins from the Golgi to post-Golgi membrane compartments[39]. Finally, palmitoylation can "fine-tune" other existing PTMs on the same protein, especially when myristoylation and prenylation occur in proximal sites[40].

1.3 Acetyl-CoA, α-ketoglutarate, and palmitic acid availability are regulated by large enzymatic complexes.

1.3.1 The pyruvate dehydrogenase complex controls the availability of acetyl-CoA

Production of Acetyl-CoA is heavily reliant on available carbon sources. The carbon required can be produced either inside the mitochondria, or have cytosolic origins. When glucose levels are low, CoA is acetylated by the acetyl-CoA synthase (ACS), with the alcohol dehydrogenase providing the necessary acetyl- moiety, by hydrolyzing ATP and ethanol as the carbon source[7]. Acetyl-CoA can also be produced through the branched-chain keto-amino-acid catabolism pathway. In this, branched-

chain keto-amino-acids, such as leucine, isoleucine, and valine, are transaminated to α-ketoacids and then decarboxylated to either isovaleryl-coenzyme A, alpha-methylbutyryl-coenzyme A, or isobutyryl-coenzyme A, respectively, by the branched-chain ketoacid dehydrogenase complex (BCKDHc), located in the mitochondrial matrix. After these -coenzyme A intermediates have been formed; they are then subjected to a long process of dehydrogenation, then carboxylation, and finally are hydrated to form one final -coenzyme A intermediate that will then be broken down to Acetyl-CoA and acetoacetate or Succinyl-CoA depending on the starting amino-acid.

Another metabolic pathway that is used for Acetyl-CoA production, when glucose levels are low, and takes place in the mitochondria, is the β-oxidation of fatty acids. It's an intensive process where fatty acids are first converted to acyl-CoA, which is then converted, after a 4-step reaction, to an acyl-CoA with a main chain shorter by two carbons and Acetyl-CoA. The first part of the reaction is catalyzed by the acyl-CoA dehydrogenase and the other three by the so-called tri-functional mitochondrial protein, a protein that displays three different functionalities of 2-enoyl coenzyme A hydratase, long-chain 3-hydroxy acyl-coenzyme A dehydrogenase and long-chain 3-ketoacyl CoA thiolase.

When glucose levels are sufficiently high, glycolysis produces large amounts of citrate through the TCA cycle, which is then transported from the mitochondria to the cytosol and other organelles, where it is broken down to Acetyl-CoA and oxaloacetate by the activity of ATP citrate lyase. Another source of Acetyl-CoA is through the conversion of pyruvate. One example of this conversion is by the activity of pyruvate formate lyase, which results in the production of Acetyl-CoA and formic acid.

Even when considering all the above-mentioned Acetyl-CoA production methods, the main pathway that produces the majority of Acetyl-CoA is through glycolysis for the subsequent oxidative decarboxylation of pyruvate, also known as the pyruvate dehydrogenase reaction. This reaction uses pyruvate as a substrate to produce Acetyl-CoA, NADH, and CO_2 and is catalyzed by the pyruvate dehydrogenase complex (PDHc).

The pyruvate dehydrogenase complex (schematic shown in **Figure 4**) is one of the largest soluble macromolecular assemblies in the cell, with a molecular weight of approximately ten megadalton (MDa). It is comprised of multiple copies of three main

different enzymatic subunits, the pyruvate dehydrogenase (lipoamide) or E1p, the dihydrolipoyl acetyltransferase or E2p, and the dihydrolipoyl dehydrogenase or E3. The prokaryotic PDHc core is comprised of 24 E2p monomers assembled into a cube. In contrast, the eukaryotic PDHc core is arranged into a dodecahedron containing 60 copies of the E2o enzyme and displays icosahedral (I) symmetry in eukaryotes. Each vertex of the dodecahedron contains an E2p trimer. The E2p displays a quite distinct domain-linker structure composed of four structured domains, each one connected to the other through a flexible linker. All four ordered domains are of paramount importance to the overall complex's function and include two lipoyl-binding domains (LD), a peripheral subunit binding domain (PSBD) responsible for binding the E1p peripheral subunits, and finally the catalytic core-forming domain. The other two components of the complex, the E1p and the E3, are also found in multiple copies and located at the core structure's periphery, bound non-covalently but quite tightly to the outer region at their corresponding binding sites. Individual structures of the ordered domains of the PDHc exist and have been resolved through multiple techniques, such as X-ray crystallography[41-43], cryo-electron microscopy (cryo-EM)[44], and nuclear magnetic resonance (NMR) spectroscopy[45].

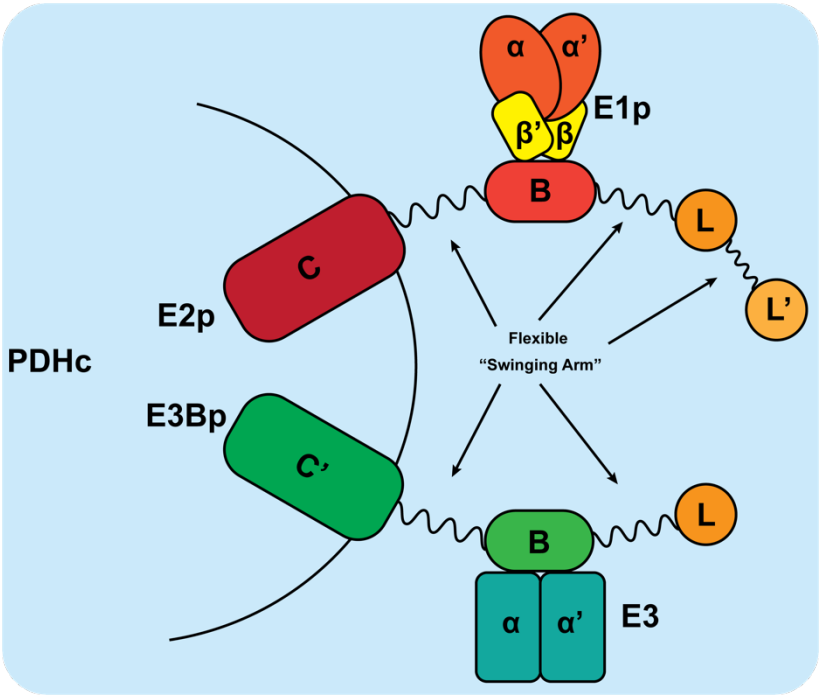


Figure 4: Organization of the different proteins comprising the pyruvate dehydrogenase complex (PDHc).
The flexible linkers connecting the lipoyl domains allow for easier transfer of substrates among the E1p, E2p, and E3 subunits needed for the creation of Acetyl-CoA. E2p: dihydrolipoyl acetyltransferase, E3BP: E3 Binding Protein, C: core domain, PSBD: peripheral subunit binding domain, E1p: pyruvate dehydrogenase (lipoamide), E3: dihydrolipoyl dehydrogenase, L: lipoyl domain. Figure reproduced from[12].

For the PDHc reaction to take place, reaction intermediates must be transferred from one of the complex's subunits to the next. First, the pyruvate is bound by the E1p, then decarboxylated, and the lipoyl-moiety is attached to the LD domain of the E2p. The LD transfers the acyl- group to the catalytic part of the E2p, where it is transferred to a bound coenzyme A. After the reaction is finished, Acetyl-CoA is released, and the LD is regenerated by the E3 through re-oxidation in order to prepare it for another reaction cycle. This movement of the E2p's LD domain, where it cycles through each of the complex' subunits, can be described as a "swinging arm" mechanism, where the LD, through its movements, transports the reaction intermediates from one active site to the next to complete the reaction. This kind of flexibility of the E2p arm is mostly attributed to the unresolved linker regions that connect the ordered E2p domains[46]. These linker regions have usually a length of 25 to 30 amino-acids and, thanks to their flexibility, which is mostly attributed to their predicted disorder properties, can facilitate the movement of the ordered domains, thus coupling the active sites to each other to

complete the reaction mechanism[47]. If the flexible linkers are fully extended, and assuming a Cα-Cα distance of 3.8 Å, a maximum distance of up to 122 Å can be calculated between the E2p core domain and the PSBD, 194 Å between the PSBD and the first LD, and finally another 129 Å between the first and second LD domains. Of course, while these distances seem quite large, they are the theoretical maximum extended conformations of the linker domains. Based on experimental data, the complete PDHc seems to have an overall diameter of ~ 500 Å, with the external densities being at least after the ~ 300 Å region of the complex, with an empty region between the E2p core and the external subunits[48]. Again, the flexible regions are mostly observed as extended[48], and immunological studies have confirmed very low immunological cross-reactivity[49], meaning that the flexible linkers of the E2p do not share significant antigenic epitopes with other protein regions. This observation hints at their uniqueness as a class of domain linkers that can assume varying conformations in order to facilitate the PDHc reaction. As each E2p subunit that takes part in the complex' core assembly extends a lipoyl- arm in the periphery, this means that the core's peripheral space is quite crowded. 60 different lipoyl- arms must assume specific conformations to avoid unfavorable van der Waals (vdW) interactions due to the spatial proximity with the remaining *N-ter* domains.

Another interesting observation concerning the lipoyl- arms, strengthening the case made for their unique characteristics, is that even though the linkers have been shown to be disordered, they do not assume random coil conformations[50] (**Figure 5**). This hints at the functional role of the predicted-as-disordered linker regions, as they must assume transient but suitable conformations that will facilitate the interactions between the ordered domains for the PDHc reaction to take place. The importance of the linker regions has been further verified by mutagenesis studies when it was shown that mutating the linker regions or varying their length negatively affects growth rates in prokaryotic organisms[47]. When the total number of disordered linkers present in the PDHc is taken into account (48-60 E2p subunits per core), it is evident that any change to the linker sequence will have a greatly amplified effect on the complex' stability and functionality.

Nevertheless, the cumulative effect of the disordered regions on the total conformational landscape of the lipoyl- domains or the causal relationship between the disordered regions, the ordered domains, and the overall molecular redundancy is

still poorly understood, even though there are recent molecular dynamics observations on how the neighboring E1p and E3 subunits can impose flexibility changes on the E2o disordered linkers[51]. As mentioned above, the PSBD of the E2p is tethering the E1p to the periphery of the complex' core. The same function, but for the E3, is performed by the E3 binding protein (E3BP), which is also part of the PDHc core structure. It is present in ~12 copies, contributing another level of complexity to the structural interpretation of the mechanism that underlies the PDHc reaction. Structurally, there are a lot of similarities between the E3BP and the E2p, but they display functional differences. First, E3BP displays no catalytic function in regards to coA acetylation; second, its PSBD is dedicated to tethering solely the E3 in place; third, in contrast to the E2p that contains two LDs in its *N-ter* sequence, the E3BP has only a single LD. As is the case with the E2p, the E3BP also contains flexible linkers, predicted to be disordered, connecting its LD and PSBD to the core domain[52,53].

Taking into account the multitude of proteins that comprise the PDHc, an astounding architecture is revealed: approximately, its core is formed by 48-60 E2p and 6-12 E3BP proteins, and in the periphery, up to 48-60 E1p and 12 E3 proteins can be found. Over 100 LDs have to navigate through the protein subunits and transfer the reaction intermediates to complete the reaction, meaning that the observed disordered linkers connecting the LDs and the PSBDs to the core have a large role in defining their conformational freedom. Additionally, in this crowded space, the on/off rates of the LDs and their interactions with the surrounding polypeptide chains will affect the generation of Acetyl-CoA.

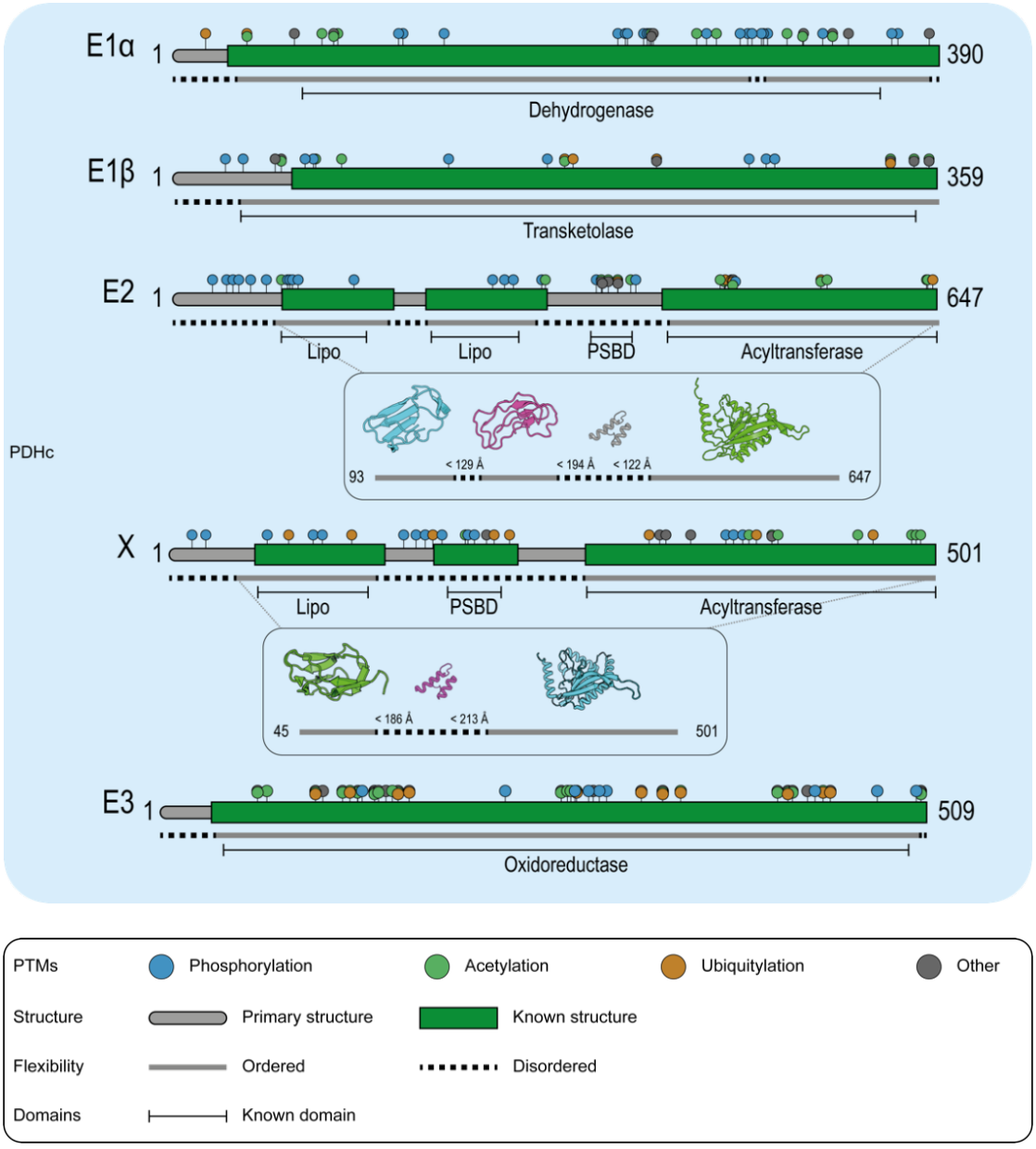


Figure 5: **Sequence analysis and characterization of human PDHc proteins.**
Protein sequences were analyzed regarding PTMs, known structure, domains, and flexibility. It is shown that, beyond regulation imposed by the flexible linkers that are predicted to be disordered, additional regulation is conferred by multiple PTMs as well as structural redundancy (see text). Data about PTMs were obtained from PhosphoSitePlus (https://www.phosphosite.org/), known structures were retrieved from the Protein Data Bank (https://www.rcsb.org/) and manually validated. Homologous structures were identified using HHpred (https://toolkit.tuebingen.mpg.de/tools/hhpred). Domain annotations were retrieved from the pfam-database (https://pfam.xfam.org/). Disorder prediction was performed using SPOT-disorder2 (https://sparks-lab.org/server/spot-disorder2/). Structures shown for PDHc E2: the structures of the first (protein data bank accession code, PDB ID: 1FYC) and second (2DNE) lipoyl domain, a placeholder (from E3BP; low homology) PSBD structure (1ZY8) and the catalytic core (6H60) and the lipoyl domain of E3BP (2DNC), the PSBD (1ZY8) and core structure (6H60). The maximum distances of the disordered regions are calculated using a Cα-Cα distance of 3.8 Å. Illustrations of 3D protein models were generated using PyMOL (https://pymol.org). Figure reproduced from[12].

The importance of PDHc is also highlighted by the existence of dedicated phosphatases and kinases, that, through the post-translational modifications that affect the PDHc, add another regulatory layer to its function[54]. Their importance has been specifically observed through studying the diminished mitochondrial function in various cancer types[55]. Through phosphorylation of distinct serine residues in the E1p sequence (Ser264, Ser271, Ser203) that are located in two of the structure's flexible loops (loop A: 259-282 and loop B: 195-205), Acetyl-CoA production can be diminished or even completely inactivated, based on the specificity of the phosphorylation of each loop[54]. More specifically, in the case of loop A phosphorylation, thiamine diphosphate (ThDP) anchoring to the active site of the E1p is hindered, while loop B phosphorylation hinders the chelation of a coordinated Mg^{2+} ion by the ThDP group[56]. The missing density for both loops in experimentally resolved crystallographic structures of E1p strongly hints towards their structural disorder[56,57]. When the loops are phosphorylated, the bulky phosphoryl groups introduce steric clashes and dismantle the hydrogen bond network that would otherwise maintain the loops' ordered conformations[56], leading to E1p catalytic efficiency loss, a phenomenon that is pathologically manifested as PDHc deficiency[57]. Another important PTM site of E1p with pathological implications is the phosphorylation of Tyr301[58], as in various cancer studies, it has been implicated with resistance to therapy and promotion of the Warburg effect. Apart from the previously mentioned PTMs, there's a large number of other PTM sites that have been identified for all the proteins that comprise the complete PDHc, with the majority of them located in the proteins' ordered regions. It is still unclear how PTMs affect the disordered linkers that govern the reaction intermediates' transfer or how these modifications translate to a physiological effect.

1.3.2 α-Ketoglutaric acid availability is regulated by the 2- oxoglutarate dehydrogenase complex

AKG is a key intermediate in the TCA cycle. The isocitrate dehydrogenase produces AKG by decarboxylation of the isocitrate, which is then used as a substrate for the 2-oxoglutarate dehydrogenase complex to perform the next step in the reaction and convert it to succinyl-CoA. AKG is located at a critical junction of the TCA cycle,

and its production is a critical anaplerotic step, as AKG can be also generated and fed into the TCA cycle *e.g.,* through the oxidative deamination of glutamate, performed by the glutamate dehydrogenase. Nevertheless, it is OGDHC that controls the available AKG concentrations, functioning as a rate-limiting step of the complete TCA cycle reaction pathway.

OGDHc is a large enzymatic complex, displaying a strikingly similar organization to the PDHc (**Figure 6**). It is a 4 MDa molecular machine, comprised of multiple copies of three basic enzymes: the 2-oxoglutarate dehydrogenase (E1o), the dihydrolipoyl succinyltransferase (E2o), and, the same as in PDHc, the dihydrolipoyl dehydrogenase (E3). The E2o forms a cubic, 24-meric core structure with octahedral symmetry, whereas the multiple copies of E1o and E3 are located at the core's periphery. It is quite interesting, that another oxo-acid dehydrogenase complex located in the inner mitochondrial matrix, the BCKDHc, shares exactly the same cubic architecture of its similarly 24-meric core structure[59]. In contrast to the PDHc, it has been shown to lack dedicated phosphatases or kinases, and there has not been observed an analog to the E3BP of PDHc to be part of its core structure[60].

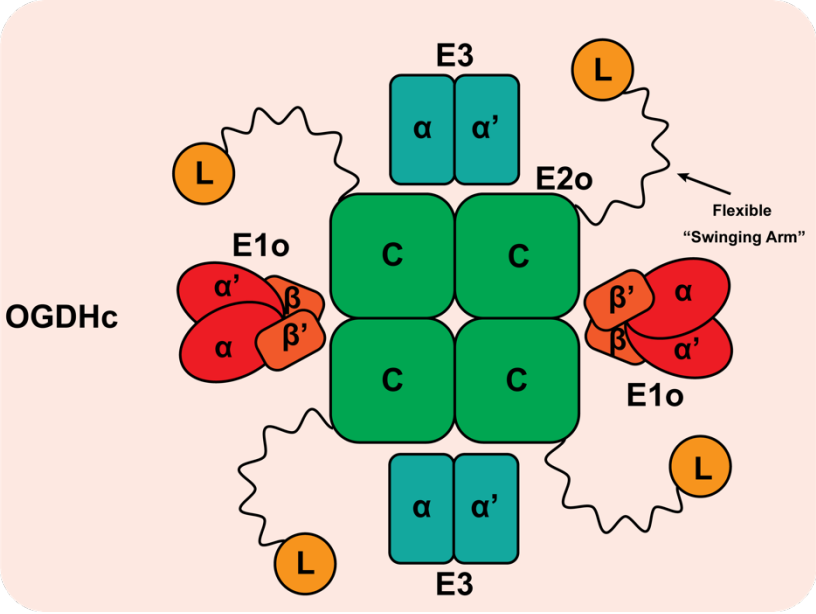


Figure 6: Organization of the 2-oxoglutarate dehydrogenase complex (OGDHc). The disordered flexible arms holding the lipoyl domains facilitate the substrate channeling among the core E2o proteins and the peripheral E1o and E3 subunits. E1o: 2-oxoglutarate decarboxylase, E2o: dihydrolipoyl succinyltransferase, E3: dihydrolipoyl dehydrogenase, L: lipoyl domain. Figure reproduced from[12].

The mechanism for the recruitment and tethering of the complex's peripheral subunits, combined with the so-far observed lack of an E3BP analog, is of particular research interest. The *N-ter* of the OGDHc E2o subunit contains an LD with a similar function as the LDs of the PDHc E2p, tethered to the core domain by a 73-residue-long flexible linker, but, in contrast to the E2p, lacks a dedicated PSBD to bind the E1o and keep it in proximity to the core. A hypothesis can be made that, since there is a high degree of similarity between the E2p and E2o subunits, a similar mechanism for the peripheral subunit binding should exist. If so, an unobserved thus far disorder-to-order transition should take place at the flexible linker region of the E2o, allowing for the docking of the peripheral subunits that are critical for the completion of the reaction and making up for the lack of a dedicated subunit that would function similarly to the E3BP. It is also probable that an analog to the PDHc's E3BP also exists, in the form of another protein that will play a direct role in the regulation of the peripheral subunit's proximity to the core. Recently, Heublein *et al.* discovered a novel interacting subunit of the human OGDHc, called KGD4, that interacts both with the OGDHc and the mitochondrial ribosome and seems to display a mitochondrial moonlighting function[61]. The observed interactions showed that Kgd4 binds to the E1o and E2o with its *N-ter*, whereas its *C-ter* domain interaction with the E3 helps to keep the subunit tethered to the vicinity of the OGDHc core[61].

Structurally, only the core acyltransferase domain of the OGDHc E2o is resolved, with the lipoyl- domain and its flexible linker that tethers it to the core remaining unresolved, mainly due to its inherent flexibility, with the linker predicted as disordered. It is though, once again, this same flexibility that allows the lipoyl- arm to assume various conformations and regulate the transfer of reaction intermediates, as well as the interactions between the complex's subunits[62]. The disordered linker region, when fully extended, can cover a maximum distance of 289 Å, similar to the disordered linker of the PDHc E2p. This reach could be the defining factor for the transfer of the reaction's intermediate products between subunits and, consequently, the key regulatory factor of the complete OGDHc reaction (**Figure 7**).

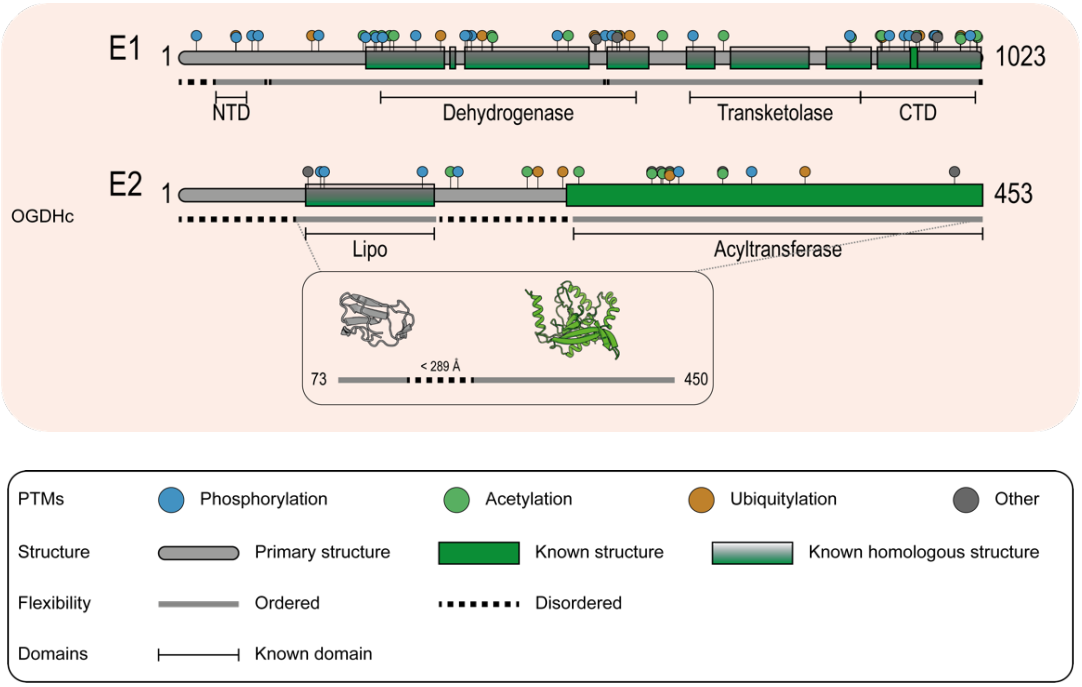


Figure 7: Sequence analysis and characterization of human OGDHc proteins. Protein sequences were analyzed regarding PTMs, known structure, domains, and flexibility. Data about PTMs were obtained from PhosphoSitePlus (https://www.phosphosite.org/), known structures were retrieved from the Protein Data Bank (https://www.rcsb.org/) and manually validated. Homologous structures were identified using HHpred (https://toolkit.tuebingen.mpg.de/tools/hhpred). Domain annotations were retrieved from the pfam-database (https://pfam.xfam.org/). Disorder prediction was performed using SPOT-disorder2 (https://sparks-lab.org/server/spot-disorder2/). Structures shown are a homologous lipoyl structure from PDHc E2 (1FYC) and the core of OGHDc (6H05). The maximum distances of the disordered regions are calculated using a Cα-Cα distance of 3.8 Å. Illustrations of 3D protein models were generated using PyMOL (https://pymol.org). Figure reproduced from[12].

It is of note that, as is the case with PDHc, there are 24 E2o lipoyl arms that coordinate the transfer of the reaction's intermediate products, hinting that, once again, non-covalent interactions between them will play a significant role in defining and limiting the available conformational space.

The E1o was very recently characterized, and its dimeric structure was resolved at high resolution[63]. In contrast to the PDHc E1p, the E1o is almost two times larger, and in its subunits, the dehydrogenase and transketolase activities are fused. Even though there are no dedicated kinases and phosphatases, the E1o displays allosteric interactions that control its activity and, as a result, the function of the complete OGDHc[64,65]. There are PTM sites identified for both the E1o and E2o subunits. In the case of E2o, some of the PTMs appear to be located at the disordered

region that connects the core domain to the lipoyl- domain, although little is known about how these may affect the conformational variability and function of the flexible linker. It is also of note that due to its Succinyl-CoA production capabilities, the OGDHc can post-translationally modify other proteins participating in the TCA cycle, such as fumarase and PDHc, and has been shown to regulate the abundance of other metabolites in specific cell types, *e.g.,* neurons[66].

1.3.3 Palmitic acid is produced by the fatty acid synthase, a modular enzymatic complex

PA represents one of the main body components of both humans and animals. In humans specifically, up to 30% of their depot fat is mostly composed of PA and is one of the main lipids that are contained in breast milk. During fatty acid synthesis, PA is the first fatty acid that carbohydrates are converted into. Consequently, PA participates in a negative feedback loop and regulates the activity of acetyl-CoA carboxylase, the enzyme responsible for the conversion of acetyl-CoA to malonyl-CoA. In turn, malonyl-CoA is one of the main sources of carbon for fatty acid chain elongation, meaning that with diminished malonyl-CoA production, PA generation will also halt. Fatty acid synthase, another large enzymatic complex, is the main controller of the *de novo* generation of PA. Its function is to convert Acetyl-CoA and Malonyl-CoA into long-chain saturated fatty acids, accompanied by NADPH oxidation[67].

The main enzymatic components of fatty acid synthesis organize utilizing architecturally distinct evolutionary mechanisms that are reflected across the kingdoms of life. Fatty acid synthesis in prokaryotic organisms, plants, and mitochondria is carried out by seven distinct single-functionality enzymes and the acyl carrier protein (**Figure 8**). In contrast, in animals and fungi, the different functionality enzymes have fused into large polypeptide chains. More specifically, in fungi, FAS is a complex with an A_6B_6 stoichiometry, whereas in humans, a single A_2 complex exists, with each chain containing ~2500 residues. The structure of the complete human FAS has yet to be elucidated, as only fragments of the complete structure have so far been structurally characterized. This is in line with other structural investigations that deal

with large enzymatic assemblies, as is the case with PDHc and OGDHc, even if FAS has an overall smaller size of around 0.5 MDa.

In order for the fatty acid synthesis reaction to begin, an acyl- moiety, derived by the acetyl-CoA, is loaded onto the acyl carrier protein (ACP) by the malonyl-/acetyl-CoA-ACP-transacylase (MAT). The ACP has been previously activated by post-translational modification of a conserved serine residue, through the covalent addition of a phosphopantetheinyl moiety of coenzyme A (CoA), performed by the phosphopantetheinyl transferase (PPT). MAT's second functionality is to then transacylate the malonyl- moiety of Malonyl-CoA to the ACP. In the next step, β-ketoacyl synthase (KS) condenses the acyl- intermediate and the malonyl-ACP to β-ketoacyl-ACP (during the first cycle of elongation, this results in the creation of acetoacetyl-ACP). The β-carbon of the β-ketoacyl-ACP is then reduced by the β-ketoacyl reductase (KR) accompanied by NADPH oxidation, and the resulting β-hydroxyacyl-ACP is dehydrated to a β-enoyl intermediate with the help of a dehydratase (DH). The last two steps of the complete reaction are the reduction of the previously formed β-enoyl intermediate by the NADPH-dependent β-enoyl reductase (ER), leading to the formation of a C4 acyl- substrate that is elongated each time by two more carbons derived from Malonyl-CoA, to a final length of 16 carbons (in the case of palmitic acid) up to 18 atoms (stearic acid) and its release from the ACP by a thioesterase (TE)[68].

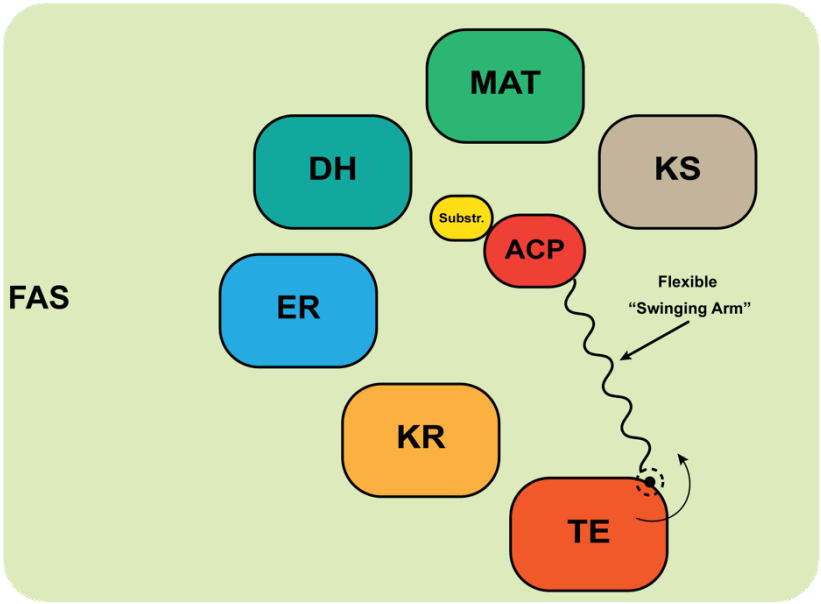


Figure 8: The fatty acid synthase (FAS) subunit organization.
The acyl-carrier protein (ACP) is linked to the thioesterase (TE) domain through a flexible disordered arm that allows access to the covalently connected substrate to multiple subunits with different activities needed for *de novo* lipogenesis. TE: thioesterase, KR: β-ketoacyl reductase, ER: β-enoyl reductase, DH: dehydratase, MAT: malonyl−/acetyl-CoA-ACP-transacylase, KS: β-ketoacyl synthase, ACP: Acyl- Carrier Protein, Substr.: Substrate. Figure reproduced from[12]

According to previously published molecular dynamics simulations[69], the binding of the ACP to the different enzymatic sites of the FAS chain was thought to be completely stochastic and not regulated by the FAS complex. In contrast, a study by Singh *et al.* revealed the presence of a γ subunit in the *Saccharomyces cerevisiae* FAS that was proven to regulate its enzymatic activity according to NADP abundance, by affecting the FAS higher-order structure and limiting the available conformational space of the ACP, impacting reaction intermediate transfer[70]. The ACP subunit (spanning residues 2125 to 2192 in the yeast FAS), follows the paradigm of the LD in PDHc and OGDHc and is flexibly tethered at both its termini by flexible linkers (~60 residues long and ~120 residues long at the *N-ter* and *C-ter* respectively), which have yet to be structurally characterized in high resolution by either X-ray crystallography or cryo-EM. Structural studies have nevertheless captured the ACP[70-72], despite the difficulties imposed by the flexible linkers at both sides, which enable its large conformational variability. The inability of structural characterization, along with computational predictions, hint at the role of structural disorder of the flexible linkers flanking the ACP, and they have been so far observed only in low-resolution cryo-EM

reconstructions[73]. As to the existence of PTMs in the FAS structure, even though their exact effect on its function is still unknown, many different PTMs are present along its polypeptide chain, located on both disordered and ordered structural regions (**Figure 9**).

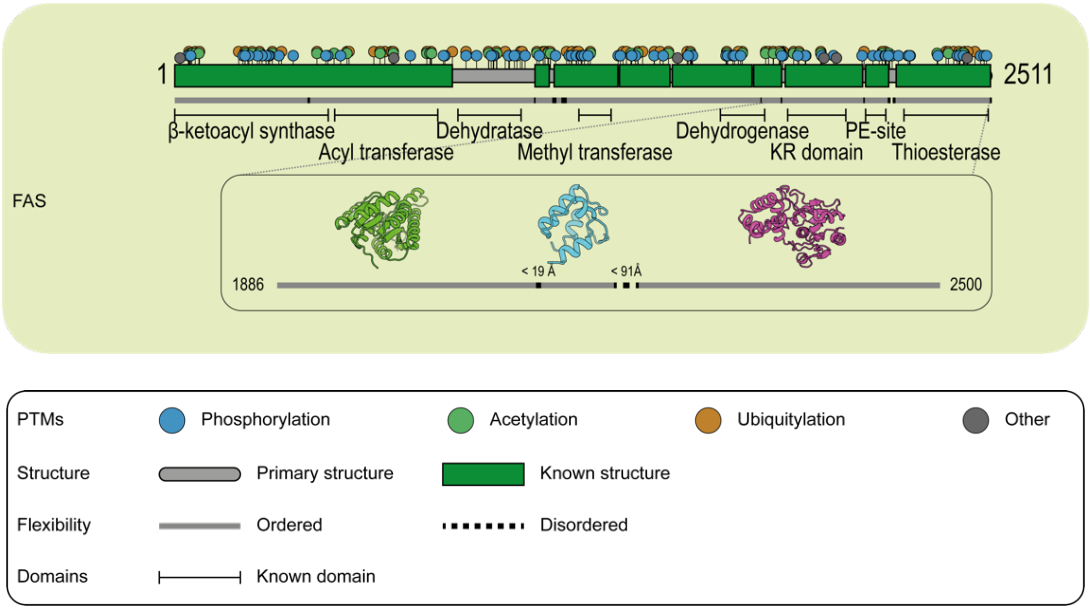


Figure 9: Sequence analysis and characterization of human FAS protein.
Protein sequence was analyzed regarding PTMs, known structure, domains and flexibility. Data about PTMs were obtained from PhosphoSitePlus (https://www.phosphosite.org/), known structures were retrieved from the Protein Data Bank (https://www.rcsb.org/) and manually validated. Homologous structures were identified using HHpred (https://toolkit.tuebingen.mpg.de/tools/hhpred). Domain annotations were retrieved from the pfam-database (https://pfam.xfam.org/). Disorder prediction was performed using SPOT-disorder2 (https://sparks-lab.org/server/spot-disorder2/). Structures shown are the KR-domain (5C37), the acyl-carrier (PE-site; 2CG5) and the thioestarese (4Z49). The maximum distances of the disordered regions are calculated using a Cα-Cα distance of 3.8 Å. Illustrations of 3D protein models were generated using PyMOL (https://pymol.org). Figure reproduced from[12].

As mentioned earlier, in regards to the shuttling of reaction intermediates across the multiple enzymatic domains of FAS by the ACP, the enzyme employs an analogous mechanistic paradigm as in the cases of PDHc and OGDHc: a lipoyl- or acyl- carrier domain that is tethered to the main structure via a flexible linker is responsible for this transfer, with the linker's manifested disorder allowing for the exploration of the wide conformational space necessary for the responsible domain to perform this type of functionality. In both the fungal and human FAS, there are multiple

copies of the flexible lipoyl- arms, with each containing six and two, respectively. The lipoyl- arms are spatially confined within the complex's overall structure, but there should still be available conformational space for their intercommunication[74] a fact that, along with the regulatory effect on the overall reaction by the γ subunit[70], should provide clues to their regulation. Despite that, the higher-order regulation of the multiple carrier proteins remains undiscovered. In the case of the fungal FAS, its cage-like overall structure[72,75] could limit unwanted protein interactions of the ACP through its spatial confinement in the FAS dome. Lacking a complete structural characterization, especially in a native context, structure-function insights and further understanding of the regulatory mechanism of FAS remain as open questions.

1.4 Large enzymatic complexes and their inclusion in "protein communities"

In the life cycle of a cell, the coordination and implementation of multiple processes are of paramount importance for its overall survival and propagation. For this reason, multiple proteins that participate in the same cellular process are often organized into larger assemblies called "protein communities"[72,76,77]. By employing this type of spatial organization, the proteins participating in a specific process, *e.g.,* a metabolic or signaling pathway, will be brought together, and their intermediate products will be linked. A protein community can act as a type of compartmentalization, especially when a complete reaction or signaling pathway cannot be spatially isolated in a dedicated membrane scaffold or organelle. In the case of proteins that belong specifically to the same metabolic pathway, the term "metabolon" can also describe this type of multi-protein organization[78]. The protein complexes discussed in section 1.3 fall in this exact category. Indeed, a protein that is taking part in the formation of a community can acquire different conformations that will facilitate its functionality[79] in this context. The conformational changes in a community-embedded protein can vastly differ, both in structure and function, when compared to the same protein studied in isolation[80].

Until recently, structural biology has focused on studying overexpressed proteins, as they can offer other advantages: an overexpressed, purified sample will be homogeneous, and the protein under study can be expressed in concentrations

that are adaptable to any type of experimental methodology and especially to traditional structural biology methods, such as X-ray diffraction (XRD), nuclear magnetic resonance (NMR) spectroscopy and single-particle cryogenic electron microscopy (cryo-EM). Nevertheless, overexpressed protein samples lack information about the native context. Ideally, a modern structural biology approach for the investigation of a large enzymatic complex would be to supplement previous *in vitro* protein structural knowledge with the study of the same protein as part of its community, in order to more fully understand its native structure and function in the cellular context.

1.5 Employing cell extracts to retain information of a protein complex's native state

Cellular extracts may be the most suitable sample that would allow a researcher to retain the native context during the investigation of a protein's structure and function, while biochemically exploiting its cell-free character. This was demonstrated by recent structural studies, showing the cell extract's feasibility for this type of study[72,81], and revealing structural insights that retain the protein's near-native context. The study of cell extracts is by no means a new concept, as, combined with electron microscopy, differential centrifugation, and gradient fractionation[82] in order to identify enzymatic complexes[83] and explore cellular ultrastructures and organelles[84], it was already well-established and recognized for its substantial contribution in cell biology. The groundbreaking discoveries of de Duve[85], Palade[86], and Claude[87] were recognized for their importance with the Nobel prize in Chemistry of 1974 and paved the way for another, more native way of studying the cell and its components.

It is to be anticipated that the complexity of a sample derived from cell extracts will pose some barriers to its analysis. A multitude of protein communities and other structural elements of the cell reside in the sample, requiring novel approaches that will allow for their deconvolution and characterization. The abundance of each component of a cell extract sample will also affect their proper analysis. Modern studies of cell extracts can leverage the recent technological advances in sample preparation, data collection, and analysis to ensure that a minimally perturbed cell

extract sample will be produced and investigated at the highest possible resolution, therefore, allowing insights into its architecture.

1.6 Cryogenic electron microscopy as a modern tool for the visualization and investigation of complex protein mixtures

Cryogenic electron microscopy (cryo-EM) has proven itself in recent years as the go-to method for the visualization of complex samples. Ever since 2014, the so-called "resolution revolution" of the field has given rise to unparalleled structural discoveries[88,89]. Large protein complex structures, previously unattainable by other structural determination methods, *e.g.,* X-ray single crystal diffraction and solution nuclear magnetic resonance spectroscopy, are finally within reach, with examples ranging from important molecular machinery and complexes, such as the mammalian brain V-ATPase and the ribosome-Sec61-OST complex, to pathological proteins, as is the case with the tau filaments, the culprits behind tauopathies which include Alzheimer's disease[90-92]. Every field that surrounds and is connected to cryo-EM has rapidly advanced, from microscope components such as electron sources, electromagnetic lenses, and direct electron detectors[93,94], EM sample preparation[95], to image analysis and computational methods[96-99], leading to the structural elucidation of increasingly complex macromolecular assemblies[100].

One of the main characteristics that increased the suitability of cryo-EM for structural investigations of large protein assemblies is its sample preparation method. For a protein sample to be inserted into the cryo-transmission electron microscope (cryo-TEM), it needs to be embedded in a thin layer of amorphous (or vitreous) ice. The process is called vitrification, and proteins that are vitrified are maintained in a frozen-hydrated state, assuming random orientations inside the vitreous ice. For the vitrification process to be successful, protein-related alterations are not necessary, as is the case with proteins that are designated for crystallization prior to X-ray crystallography and may require prior modifications, such as mutations or truncations. Additionally, due to the absence of any canonical organization in vitreous ice, proteins are free to assume various structural conformations that can be captured, distinguished in the primary data as distinct "structural signatures", and analyzed[101].

Recent examples in the literature include capturing the pulsing motion of the active fatty acid synthase[102] or the various ribosomal conformational changes throughout elongation of the polypeptide chain during protein synthesis[103]. Finally, cryo-EM can greatly contribute in understanding the always elusive disordered regions of a protein's structure, especially in the context of their functions in metabolons[12] and with the constant surge of new developments in image processing of cryo-EM data, their visualization may be within reach[104].

1.6.1 Computational advances in cryo-EM image analysis and the advent of artificial intelligence

Advances in cryo-EM instrumentation have led to a rapidly expanding amount of collected data that needs to be processed with specialized image analysis software. Even though sample complexity itself does not hinder high-resolution structure determination from cryo-EM data[105], the sheer amount of data required for such complex calculations, when aiming at the characterization of a heterogeneous sample, will act as a bottleneck that software has to overtake to achieve high-resolution reconstructions. To this end, the field of algorithm development for cryo-EM image analysis is constantly integrating the latest advances in computation and creating analysis pipelines that bridge traditional computational methods with machine learning to analyze structurally heterogeneous samples[96-99].

A wide array of computational tools is now accessible to a structural biologist working on cryo-EM data analysis, tools that can be applied not only to purified samples but have been successfully used to analyze complex, heterogeneous protein samples as well[79]. It is of note that, recently, advances in protein structure prediction, performed by artificial intelligence (AI)[106,107], enable a structural biologist to access robust and (often) reliable structural models that can facilitate the reconstruction of low-abundant proteins in complex samples, *e.g.,* native cell extracts, removing the barrier of high resolution necessary for accurate structure determination.

1.7 Native cell extracts can be leveraged for biotechnological applications

Several compounds of biotechnological interest are produced through in-cell metabolic engineering[108]. The methodology has been implemented in a plethora of cases[109,110] to enhance the desired product metabolic pathway, thus creating a newly engineered strain-of-choice[111]. The advantages of this specific methodology are clear, but recently, multiple drawbacks, such as toxic by-product accumulation[112], prohibiting cell growth times[113], and general concerns over the environmental impact of genetically-modified organisms[114], have incentivized scientists to search for alternatives.

This search has led researchers towards the use of cell-free systems (CFS)[115] in an attempt to ameliorate the process of biotechnological production of substances by dissociating, to a degree, the process from a cultivated species. Even though CFS come with their own disadvantages, *e.g.,* a lack of internal homeostatic mechanisms[116], research has focused on two main branches in their development. The first branch concerns the *in vitro* synthesis of the sought-after metabolic pathway by separately overexpressing and purifying all the components, then mixing them to recreate the pathway[117] artificially. For this approach, many parameters need to be addressed and barriers to overcome[118]. Still, its success can be summarized by the widespread use of the PURE system for protein synthesis[119] and, of course, the polymerase chain reaction (PCR)[120].

The second branch of CFS development leans towards the use of crude cellular lysates[121]. During this approach, a quantity of the desired microorganism will be lysed, the crude cellular lysate will be isolated, and then the necessary substrates of a targeted metabolic pathway will be added to the lysate for product generation to begin. This approach attempts to rectify some of the issues mentioned above but again requires *a priori* metabolic engineering of the desired microorganism strain to maximize the in-cell concentration of the primary reaction components[122]. Without this step, the production rate cannot be easily optimized due to the complexity of the lysate itself.

The use of fractionated, native cell lysates[81] could pose an attractive methodology that could greatly improve the second approach, apart from employing

genetic engineering strategies. A native lysate can be simplified by size-exclusion chromatography into what can then be used as an “adaptive protein toolbox”. After fractionation, each of the separate fractions will include a specific subset of the cell’s total protein content. These fractions can be characterized by multiple structural and/or biochemical techniques, including, *e.g.,* classic biochemical assays or immunoblotting and are amenable to further structural characterization with cryo-EM and mass spectrometry[79,123], while maintaining accessibility for direct biochemical manipulation of all biosynthetic pathways that can be identified in-extract.

1.7.1 Leveraging CFS for the production of succinyl-CoA-derived compounds of biotechnological interest

In particular, succinic acid and L-lysine represent two essential biotechnological products with a wide range of applications[124,125] and a constantly growing market demand, making their production a desirable target for the application of cell-free methodologies. Both products can have as a starting substrate succinyl-CoA through distinct biosynthetic processes[126,127]. Optimization of their biotechnological yield has so far only focused on the metabolic engineering of the respective organisms[128] and not of the pathway *ex vivo*. This is perhaps due to deep interconnectivity with other primary metabolic pathways, such as succinyl-CoA production that is performed and firmly regulated by the oxoglutarate dehydrogenase complex (OGDHc)[129]. OGDHc is a complex of great importance as it is one of the central regulators of metabolic flux in the TCA cycle, meaning that cells maintain its availability under tight control[130,131]. Furthermore, OGDHc presents high sensitivity to reactive oxygen species (ROS)[132], meaning that its possible inhibition due to oxidative stress could prove detrimental to a cell’s overall metabolism.

1.8 Aims of the study

The use of cell-free systems (CFS) derived from fractionated cellular lysates holds tremendous potential not only for understanding fundamental biochemical and signal transduction pathways but also for biotechnological applications. Nevertheless, the protein communities contained within have not yet been studied in detail, and their members have up-to-date not been resolved at near-atomic resolution. To fully take advantage of their application, a researcher should not treat them as a biochemical "black box", but should be able to understand the structure of the main components and how these interact with each other in the context of the native lysate fraction. For this purpose, the aims of this study are to:

- Combine all the latest advancements in the field of cryo-electron microscopy, along with the most recent developments in AI-guided atomic model prediction, into a robust workflow that facilitates the investigation of protein community members in the context of a native lysate fraction. This workflow will allow for the simultaneous identification, characterization, and reconstruction completely *de novo* of various captured structural signatures that belong to protein community members displaying both metabolic and signaling functions and inform for the suitability of the native lysate fraction for further application as a cell-free system.
- Characterize a native cell extract fraction with succinyl-CoA-producing capabilities across scales, and focus on elucidating the reaction's main component, the 2-oxoglutarate dehydrogenase complex. The combination of multiple techniques can provide unprecedented insights into its active, endogenous structure, information that can then be correlated to its biochemical function, thus revealing the governing principles behind the role and interactions of one of the main components of the eukaryotic cell's respiratory pathway.

2 Materials and methods

2.1 Materials

2.1.1 Chemicals and enzymes

Table 1: Key resources table with chemicals and enzymes used in the current study.

Chemicals and enzymes	**Source**	**Identifier**
1,4-Dithiothreit, min. 99 %, p.a.	Carl Roth	6908.4
2-Iodoacetamide	Sigma-Aldrich	8047440100
Acetonitrile	Sigma-Aldrich	900667-100ML
Acrylamide/Bis solution, 37.5:1	Serva	10688.01
Agar-Agar, bacteriological highly pure	Carl Roth	2266.3
Ammonimum bicarbonate	Sigma-Aldrich	A6141-500G
Ammonium acetate, ≥97 %, p.a., ACS	Carl Roth	7869.2
Ammonium persulfate	Serva	13376.02
Aprotinin from bovine lung	Sigma-Aldrich	A1153-1MG
Bestatin, 10 mg	Sigma-Aldrich	10874515001
Bovine Serum Albumin	Sigma-Aldrich	A2153-10G
Bradford Solution	Sigma-Aldrich	B6916-500ML
Bromophenol Blue	Sigma-Aldrich	318744-500ML
BS_3	ThermoFisher Scientific	21580
Cell counting kit 8	Sigma-Aldrich	96992-500TESTS-F

Chloroacetamide	Sigma-Aldrich	C0267-500G
Clarity Western ECL substrate	BIO-RAD	170-5060
Coenzyme A	Sigma-Aldrich	C3144-25MG
D (+)-Glucose p. a., ACS, anhydrous	Carl Roth	X997.2
D-Sucrose, ≥99,5 %, p.a.	Carl Roth	4621.1
Dextrin for microbiology (from potato starch)	Carl Roth	3488.1
di-Potassium hydrogen phosphate trihydrate	Carl Roth	6878.1
di-Potassium hydrogen phosphate, ≥99 %, p.a., anhydrous	Carl Roth	P749.1
DNAse I	Sigma-Aldrich	10104159001
DTT		
E-64	Sigma-Aldrich	E3132-1MG
ECL fluorescent mixture	BIO-RAD	1705062
EDTA disodium salt dihydrate, min. 99 %, p.a., ACS	Carl Roth	8043.2
Ethanol	Carl Roth	7301.1
FM 4-64	ThermoFisher Scientific	T13320
Formic acid	Carl Roth	4724.1
Glucose	Carl Roth	X997.1
Glutaraldehyde	Sigma-Aldrich	G5882-100ML
Glycerol	Carl Roth	6962.1
Glycine	Serva	23391.02
HEPES PUFFERAN®, min. 99.5 %, p.-1 kg	Carl Roth	9105.3
Iron (III) sulphate hydrate, 80 %, pure	Carl Roth	0492.1

Isopropanol	Carl Roth	CP41.1
Leupeptin	Sigma-Aldrich	L2884-1MG
LysC	Sigma-Aldrich	LYSC9001-20UG
Magnesium chloride hexahydrate, min. 99 %, p.a., ACS	Carl Roth	2189.1
Magnesium sulphate heptahydrate, ≥99 %, p.a., ACS	Carl Roth	P027.1
Methanol	Carl Roth	4627.6
Milk powder	Carl Roth	T145.3
MitoTracker Orange	ThermoFisher Scientific	M7510
NAD^+	Sigma-Aldrich	10127965001
Osmiumtetroxide	Sigma-Aldrich	201030-100MG
Pefabloc	Sigma-Aldrich	11585916001
Pepstatin A	Sigma-Aldrich	77170-5MG
Peptone ex casein	Carl Roth	8986.1
Phosphate buffered saline tablets (PBS)	Sigma-Aldrich	P4417
Potassium chloride min. 99.5 %, -1 kg	Carl Roth	6781.1
Potassium dihydrogen phosphate, ≥99 %, p.a., ACS	Carl Roth	3904.2
Precision plus protein all blue standards (marker)	BIO-RAD	161-0373
Roti®-Quant 5X	Carl Roth	K015.1
Sodium cacodylate	Carl Roth	5169.1
Sodium chloride 99,5 %, p.a., ACS, ISO	Carl Roth	3957.2
Sodium dodecyl sulfate (SDS)	Carl Roth	0183.2

Sodium nitrate, ≥99 %, p.a., ACS, ISO	Carl Roth	A136.1
TEMED	Carl Roth	2367.3
Thiamine Diphosphate		
Tris	Carl Roth	AE15.2
Tris hydrochloride	Carl Roth	9090.2
Trypsin	Sigma-Aldrich	T6567-5X20UG
Tryptone	Sigma-Aldrich	T7293
Tween 20	Carl Roth	9127.1
Urea	Carl Roth	7638.1
Yeast extract, micro-granulated	Carl Roth	2904.3
α-ketoglutaric acid	Sigma-Aldrich	75890-25G
β-mercaptoethanol	Sigma-Aldrich	444203-250ML

2.1.2 Equipment and instruments

Table 2: Main equipment and instruments used in the current study.

Instrument	**Type**	**Company**
Incubator	Heracell 150i	Thermo Fisher Scientific
Tabletop Centrifuge	Heraeus Megafuge 40 R	Thermo Fisher Scientific
Bead beater	FastPrep-24™ 5G	MP Biomedicals™
Ultracentrifuge	OPTIMA™ MAX-XP (TLA110)	Beckman Coulter

FPLC System	ÄKTA pure 25 M	Cytiva (GE Healthcare)
Plate reader	Epoch 2 Microplate Spectrophotometer	Agilent (BioTek)
Gel Imaging System	ChemiDoc™ MP Imaging Systems	Bio-Rad
Thermomixer	ThermoMixer C	Eppendorf
Glow Discharge Cleaning System	PELCO easiGlow™	Ted Pella, Inc.
Vitrification instrument	Vitrobot Mark IV System	Thermo Fisher Scientific
Microscope (300 kV)	JEOL JEM-3200FS	JEOL
Camera	Gatan K2 Summit Direct Detection Camera	GATAN
Microscope (200 kV)	Thermo Fisher Scientific Glacios Cryo Transmission Electron Microscope (Cryo-TEM)	Thermo Fisher Scientific
Camera	Falcon 3EC Direct Electron Detector	Thermo Fisher Scientific
Confocal Laser Scanning Microscope	Zeiss LSM880	Carl Zeiss
Ultramicrotome	Ultracut S	Leica
Microscope (80 kV)	Zeiss EM 900 TEM	Carl Zeiss

2.1.3 Model organism

Chaetomium thermophilum var. thermophilum La Touche 1950 (*Thermochaetoides thermophila*[133]) was acquired from DSMZ (Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures, Germany), and the preserved in freeze-dried ampoule spores were cultivated as directed by company guidelines (DSMZ Media list Medium 188, temperature: 45 °C).

2.1.4 Antibody generation

Specific antibodies were commissioned from GenScript (GenScript USA Inc., NJ) against PDHc E2p-His-Tag (aa 29-459), OGDHc E1o-His-Tag (aa 611-818), OGDHc E2o-His-Tag (aa 39-420) and E3-His-Tag (aa 35-504) (**Table 3**). Briefly, each sequence, after being codon-optimized, was cloned into a pET-30a(+) vector, in the same open reading frame with a His-Tag. Proteins were expressed in 1 L of TB culture medium, followed by a two-step purification of the cell lysate supernatant (Ni-NTA purification, followed by a Superdex 200 size exclusion chromatography). After purification, the proteins produced were stored in PBS, 10 % glycerol, 0.2 % SDS, pH 7.4, and PBS, 10 % glycerol, pH 7.4 for the OGDHc and PDHc derived proteins, respectively. After immunization of New Zealand rabbits, affinity-purified antibodies were isolated from their serum and then shipped to the *Kastritis* laboratory. Secondary antibodies against all primary antibodies mentioned above were Goat Anti-Rabbit IgG H&L (HRP), which were acquired from Abcam and used according to the included instructions.

Table 3: Antibodies used in the current study.

Antibody	**Source**	**Identifier**
Rabbit polyclonal antibody a-E2p against *C. thermophilum* E2p-His-Tag (29-459)	Custom-made by Genscript	RRID: AB_2888985

Rabbit polyclonal antibody a-E1o against *C. thermophilum* E1o-His-Tag (611-818)	Custom-made by Genscript	RRID: AB_2924900
Rabbit polyclonal antibody a-E3 against *C. thermophilum* E3-His-Tag (35-504)	Custom-made by Genscript	RRID: AB_2893235
Rabbit polyclonal antibody a-E2o against *C. thermophilum* E2o-His-Tag (39-420)	Custom-made by Genscript	RRID: AB_2924899
Goat Anti-Rabbit IgG H&L (HRP)	Abcam	ab205718

2.1.5 Kits

Table 4: Kits used in the current study.

Kit name	**Source**	**Identifier**
Pyruvate dehydrogenase activity assay kit	Sigma-Aldrich®	MAK183
α-ketoglutarate dehydrogenase activity assay kit	Sigma-Aldrich®	MAK189

2.1.6 Software and algorithms

Table 5: Software and algorithms used in the current study.

Software	**Source**	**Identifier**
AlphaFold-Multimer	[134]	https://github.com/deepmind/alphafold
AlphaFold2	[106]	https://github.com/deepmind/alphafold
ARP/wARP	[135]	https://www.embl-hamburg.de/ARP/
BoxPlotR	[136]	shiny.chemgrid.org/boxplotr/

ColabFold	[137]	https://github.com/sokrypton/ColabFold
COOT	[138]	https://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot/
cryoSPARC	Structura Biotechnology	https://cryosparc.com/
Cytoscape	[139]	https://cytoscape.org/
EPU	Thermo Fisher Scientific	https://www.thermofisher.com/de/de/home/electron-microscopy/products/software-em-3d-vis/epu-software.html
Fiji	[140]	https://imagej.net/Fiji
findMySequence	[141]	https://gitlab.com/gchojnowski/findmysequence/-/tree/master/
Gen5™	BioTek Instruments	https://www.biotek.com/products/software-robotics-software/gen5-microplate-reader-and-imager-software/
ggpubr	[142]	https://rpkgs.datanovia.com/ggpubr/index.html
HADDOCK	[143]	http://haddock.science.uu.nl/services/HADDOCK2.2
Image Lab Software 6.1	BIO-RAD	https://www.bio-rad.com/de-de/product/image-lab-software
ISOLDE	[144]	https://isolde.cimr.cam.ac.uk/
Jalview	[145]	https://www.jalview.org/
MaxQuant 1.6.1	[146]	https://www.maxquant.org/
MS Excel	Microsoft Corporation	https://www.microsoft.com/en-ww/microsoft-365/excel
Omokage	[147]	https://pdbj.org/emnavi/omo-search.php
Pandas	[148]	https://pandas.pydata.org/
Phenix	[149]	https://www.phenix-online.org

PyMOL	Schrödinger, inc	https://pymol.org/
Relion 3.0	[150]	https://github.com/3dem/relion
SEGGER	[151]	https://www.cgl.ucsf.edu/chimerax/docs/user/tools/segment.html
UCSF ChimeraX	[152]	https://www.rbvi.ucsf.edu/chimerax/
UNICORN 7	GE Healthcare Europe GmbH	https://www.gelifesciences.com/en/us/shop/chromatography/software/unicorn-7-p-05649
xiVIEW	[153]	https://xiview.org/xiNET_website/index.php
ZEN Black image analysis software	Carl Zeiss GmbH	https://www.zeiss.com/microscopy/en/products/software/zeiss-zen.html

2.2 Methods

A general theoretical approach to this thesis' main methods, namely cryogenic electron microscopy, cryo-EM image analysis, and protein 3D modeling, along with several sources for more detailed information, can be found in Appendix chapter **7.1**. Below are the technical descriptions of the methods that were used in this thesis.

2.2.1 Model organism culture

After initial cultivation, the mycelium was propagated in liquid Complete Culture Media (CCM), containing per 1,000 mL of ddH_2O: 5.00 g tryptone, 1.00 g peptone, 1.00 g yeast extract, 15.00 g dextrin, 3.00 g sucrose, 0.50 g $MgSO_4$ x 7 H_2O, 0.50 g NaCl, 0.65 g K_2HPO_4 x 3 H_2O and 0.01 g $Fe_2(SO_4)^3$ x H_2O. CCM was adjusted to pH 7.1. Final cultures were performed as follows: Solid media plate cultures: liquid CCM

was supplemented with 15.00 g of Agar/1000 mL ddH_2O and the plates were then inoculated with mycelium and grown at 52°C. Liquid media cultures: 2,000 mL Erlenmeyer flasks were filled to 40% of total volume (800 mL) with liquid CCM media, small pieces of freshly grown mycelium from Agar plates were added and then incubated under shaking at 110 rpm and 10% CO_2 for 20 hours.

2.2.2 Cell imaging

Confocal laser scanning microscopy (CLSM) imaging was performed as follows: A Zeiss LSM880 (Carl Zeiss, Germany) with a 40X objective lens without immersion (plan-Apochromat 40X/0.95 N.A.) was used to image the samples and images were acquired with the ZEN Black image analysis software (Carl Zeiss, Germany). For membrane staining with FM 4-64 (ThermoFisher Scientific, USA), the dye was diluted in DMSO to a stock concentration of 10 mM; 1 µL of stock solution was then added to 1 ml of *C. thermophilum* liquid cell culture, reaching a working concentration of 10 µM. Samples were imaged after 2 to 3 min of incubation time. For mitochondria staining with MitoTracker orange (ThermoFisher Scientific, USA), the dye was diluted in DMSO to a stock concentration of 10 µM; 5 µL of stock solution was then added to 1 ml of *C. thermophilum* liquid cell culture, reaching a working concentration of 50 nM. Samples were imaged after 10 m of incubation time. Transmission electron microscopy (TEM) imaging was performed as follows: freshly propagated liquid culture *C. thermophilum* filaments were fixed with 3% glutaraldehyde (Sigma, Taufkirchen, Germany) in 0.1 M sodium cacodylate buffer (SCP; pH 7.2) for five hours at room temperature. After fixation, the samples were rinsed in SCP and postfixed with 1% osmium tetroxide (Roth, Karlsruhe, Germany) in SCP for one hour at room temperature. Subsequently, the samples were rinsed with water, dehydrated in a graded ethanol series, infiltrated with epoxy resin according to Spurr[154], polymerized at 70 °C for 24 hours, and then cut to 70 nm ultra-thin sections with an Ultracut S ultramicrotome (Leica, Germany). After cutting, the sections were applied on copper grids with formvar coating, and uranyl acetate and lead citrate were added for post-staining in a specialized EM-Stain device (Leica, Germany). For

imaging, a Zeiss EM 900 TEM (Carl Zeiss, Germany) operating at 80 keV was used. All image processing was performed using the Fiji software[140].

2.2.3 Cell-free system preparation

To prepare the cell-free system, carefully grown mycelium[155] was isolated with the use of a 180 µm-pore size sieve, then washed 3 times with PBS at 3,000 g, 4 mins, and 4°C. After removing any residual moisture, the pellet was freeze-ground with a liquid N_2 pre-chilled mortar and stored until usage at -80 °C. Approximately 8 g of the freeze-ground material was lysed in 20 mL of Lysis Buffer (100 mM HEPES pH 7.4, 5 mM KCl, 95 mM NaCl, 1 mM $MgCl_2$, 1 mM DTT, 5% Glycerol, 0.5 mM EDTA, Pefabloc 2.5 mM, Bestatin 130 µM, 10 µg·mL^{-1} DNAse, E-64 40 µM, Aprotinin 0.5 µM, Pepstatin A 60 µM, Leupeptin 1 µM), with 3 repeats of 6.5 mps shaking speed for 25 s, 4 °C in a Fastprep cell homogenizer and 3 mins of rest in ice in between. A 4,000 g centrifugation step was used to pellet the larger cell debris, and a 100,000 g high-speed centrifugation was then performed. The resulting supernatant was filtered through a 100 KDa cutoff centrifugal filter and concentrated to 30 mg·mL^{-1}.

The filtered, concentrated to 30 mg·mL^{-1} supernatant, was applied to a Biosep SEC-S4000 size exclusion chromatography column that is mounted to an ÄKTA Pure 25M FPLC (Cytiva, USA) system via a 500 µL loop. The column was equilibrated with a filtered, degassed buffer containing 200 mM of $CH_3COO^-NH4^+$ at pH 7.4 prior to sample application. The fraction volume was set to 250 µL and the flow rate to 0.15 mL·min^{-1}. Based on acquired MS data (see details below), fraction 6 was selected to be tested for suitability as a succinyl-CoA-producing cell-free system.

In order to produce an equivalent yeast sample for enzymatic activity comparison, the *Saccharomyces cerevisiae* strain from ATCC (American Type Culture Collection PO Box 1549 Manassas, VA 20108 USA; ATCC® 24657TM) was cultivated in YPDG medium (yeast extract 10.0 mg·mL^{-1}, peptone 10.0 mg·mL^{-1}, glucose 70.0 mg·mL^{-1}, dd H_20 1,000 mL) at 30°C for 5 h, to an OD_{595} of 2.5 (early exponential phase), then harvested at 3000 g for 5 min at 4 °C and washed with distilled water

(resulting pellet ~7 g). The subsequent protocol until the final sample preparation is identical to the *C. thermophilum* preparation described above.

2.2.4 Protein concentration determination

For each fraction produced after size exclusion chromatography, its total protein concentration was determined with the Bradford assay. In brief, 4 μL of each fraction were first introduced to a 96-well microplate, and then 240 μL of 1X Bradford solution (from a 5X stock, Roti®-Quant, Carl Roth) were added to each well. A standard curve of known concentrations of bovine serum albumin (BSA) was also used as a guide to measuring protein concentration by measuring absorbance at 595 nm with a BioTek Epoch2 microplate spectrophotometer.

2.2.5 Activity assays

Initial screening of the CFS for PDHc and OGDHc activity was performed with the use of the pyruvate dehydrogenase and α-ketoglutarate dehydrogenase kinetic activity assay kits that are listed in **Table 4**, following the vendor's instructions. Shortly, the activity is measured by a coupled assay reaction, which uses a colorimetric product measured at 450 nm, which is directly related to the in-sample enzymatic activity. For each well, 2 μL of the sample was added, and for the standard curve, sequential concentrations (0, 2, 4, 6, 8, 10 μL) of the kit NADH standard solution were added, in technical duplicates, in a 96-well microplate. This resulted in final NADH concentrations of 0, 2.5, 5, 7.5, 10, and 12.5 nM per well standards. The low molecular weight fraction 22 was used as a negative control, and the included in the kit standard was used as a positive control in each case. Activity calculations were performed as per the included protocol's instructions, for biological triplicates and technical duplicates of each measurement. The values obtained within the NADH standard curve linear range were used for activity calculations, with the highest average values used to plot the NADH standard curve. From all measurements (standard and sample), the final $A_{450\text{-final}}$ measurements were subtracted in order to correct for

background. For each sample, the absorbance difference from start time T_{start} to end time T_{end} was calculated as a difference in total absorbance $\Delta A_{450} = A_{450\text{-end}} - A_{450\text{-start}}$. The produced NADH (in $nM \cdot min^{-1} \cdot mL^{-1}$) was calculated with the corrected measurements after the use of the following equation:

$$nM \cdot min^{-1} \cdot mL^{-1}(milliunits \cdot mL^{-1}) = \frac{S_a}{(T_{end} - T_{start})SV}$$

with S_a representing the difference in generated NADH as calculated by the equation between the start and end time points per sample well. The sample volume, in mL, is represented by SV.

For detailed substrate kinetics, an OGDH activity assay was adapted from[56]. The enzyme assay was prepared in a reaction volume of 100 µL at 4°C, containing 100 mM NaCl, 30 mM K_2HPO_4 (pH 7.5), 2 mM $MgCl_2$, 2 mM ThDP, 4 mM α-ketoglutarate, 3 mM NAD^+, 0.4 mM CoA, 4 µL of Cell Counting Kit 8 reagent and 2 µL cell lysate containing OGDH. For K_M calculations, α-ketoglutarate was titrated from 50 to 5,000 µM, NAD^+ from 25 to 5,000 µM, and CoA from 5 to 1,000 µM. The reaction mixture (without α-ketoglutarate) was pre-incubated for 5 minutes at 37 °C for the substrate K_M calculations and at 25 °C to 65 °C with 5 °C intervals for temperature-dependent kinetic characterization, and the reaction started by the addition of α-ketoglutarate. Formazan product formation by WST-8 of the Cell counting kit 8 was monitored every minute for 1 hour at 460 nm, and concentration was calculated with the Lambert–Beer-Equation:

$$A_{460} = \varepsilon \cdot l \cdot c$$

ε represents the molar absorption coefficient of WST-8[156] ($\varepsilon = 3.07 \times 104\ M^{-1} \cdot cm^{-1}$), l the optical path length in cm, and c the sample concentration. Due to substrate excess inhibition, reaction rates were plotted against substrate concentrations using the double reciprocal Lineweaver-Burk plot[157], and K_M-values were determined at the abscissa intersection point of the asymptotic linear regression. All values reported in the OGDHc enzyme kinetics plots are reported in **Supplementary Table 1**.

2.2.6 Immunoblotting experiments

For Western Blotting (WB) experiments, in-house casted, freshly prepared, 1 mm thickness gels were used with the following composition: stacking phase is listed in **Table 6** and separating phase is listed in **Table 7**.

Table 6: Stacking phase gel ingredients.

Reagent	**Final concentration**	**Amount**
Acrylamide/Bis solution, 37.5:1	10%	3.34 mL (Stock 30% $^{w}/_{v}$)
Tris-HCl p*H* 8.8	0.37 M	2.46 mL (Stock 1.5 M)
Sodium dodecyl sulfate (SDS)	0.1%	50 µL (Stock 20%)
APS	0.04%	40 µL (Stock 10%)
TEMED	13.4 mM	20 µL (Stock 6.71 M)
dd H_2O	n/a	4.1 mL
Total	**n/a**	**10.01 mL**

Table 7: Separating phase gel ingredients.

Reagent	**Final concentration**	**Amount**
Acrylamide/Bis solution, 37.5:1	5.08%	0.85 mL (Stock 30% $^{w}/_{v}$)
Tris-HCl p*H* 6.8	0.5 M	1.25 mL (Stock 0.5 M)
Sodium dodecyl sulfate (SDS)	0.1%	25 µL (Stock 20%)
APS	0.04%	20 µL (Stock 10%)

TEMED	13.4 mM	10 µL (Stock 6.71 M)
dd H_2O	n/a	2.86 mL
Total	**n/a**	**5.02 mL**

Samples were previously mixed with a 4x loading dye (250 mM Tris-HCl pH 6.8, 40% v/v glycerol, 20% v/v β-mercaptoethanol, 0.2% w/v bromophenol blue, 8% w/v SDS) and then incubated at 100 °C for 5 min. For each native sample, around 400 ng of protein was loaded in each lane and 5 µL of Precision Plus Protein™ All Blue Prestained Protein Standards (Biorad, USA) was loaded in each gel as a marker. Concentration of recombinant control samples was around 8 ng. After loading, gel electrophoresis followed in a 1X electrophoresis buffer, freshly diluted from a 10X stock (144 g Glycine, 30.3 g Tris-base in ddH_2O) with an applied electrical field of 100 V for 1,5 h. A Trans-Blot® Turbo™ Transfer System (Biorad, USA) was used to transfer the gel contents to a nitrocellulose membrane for 20 min, with a pre-set, 25 V (1 A) applied field. Blocking was performed for 1 h, under constant stirring in 5% w/v TBST/milk solution and then the membranes were incubated at 4°C for 16 h with the primary antibody (0.2 $\mu g \cdot ml^{-1}$, 2% w/v TBST/milk). The primary antibody was removed with three washing steps of 2% w/v TBST/milk and then the membrane was incubated with the secondary antibody (0.1 $\mu g \cdot ml^{-1}$, 2% w/v TBST/milk) for 1 h. Three more washing steps of 2% w/v TBST/milk were applied and finally the membranes were screened with a ChemiDoc MP Imaging system (Biorad, USA) and freshly mixed ECL fluorescent mixture and optimal exposure times. Antibodies used can be found in **Table 3** and their generation is described in “Antibody generation”.

2.2.7 Mass spectrometry and cross-linking mass spectrometry sample preparation, data collection and analysis

All mass spectrometry related measurements of samples were performed by the Rappsilber Lab of TU Berlin, Department of Bioanalytics. Fractions 3 to 9 from the *C. thermophilum* native lysate fractionation described above were combined into two

pools (3-6 and 7-9). From each pool, 40 µL of sample was digested in-solution with trypsin as described previously[158,159]. To avoid protein precipitation during sample reduction and alkylation, 2 µL of 20% SDS were added. 1 µL of 200 mM DTT in 200 mM HEPES/NaOH pH 8.5 was added to reduce the protein samples, which were then incubated at 56 °C for 30 min. After reduction, alkylation followed with the addition of 2 µL of 400 mM chloroacetamide in 200 mM HEPES/NaOH, pH 8.5 and another incubation at 25°C for 30 min. All excess chloroacetamide was quenched by adding 2 µL of 200 mM DTT in HEPES/NaOH, pH 8.5. After reduction and alkylation, the samples were used for single-pot solid-phase-enhanced sample preparation[158,159]. For this preparation, 5 µL of 10% v/v formic acid and 2 µL of Sera-Mag Beads were added, along with enough acetonitrile (ACN) to achieve a final ACN percentage of 50% v/v, then followed by an 8 min incubation and bead capture on a magnetic rack. The beads were then washed two times with the addition of 200 µL 70% ethanol and one more time with 200 µL of ACN. After resuspension in 10 µL of 0.8 µg of sequencing grade modified trypsin in 10 µL 100 mM HEPES/NaOH, pH 8.5, the beads were incubated overnight at 37°C. The incubation was followed by a reverse phase cleanup step and then analyzed by liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) with a Q Exactive™ Plus Hybrid Quadrupole-Orbitrap™ Mass Spectrometer (ThermoFisher Scientific, USA). More specifically, an UltiMate™ 3000 RSLCnano System (ThermoFisher Scientific, USA) equipped with a trapping cartridge and an analytical column was used for peptide separation. For solvent A, 0.1% v/v formic acid in LC-MS grade water, and for solvent B, 0.1% v/v formic acid in LC-MS grade ACN were used. All peptides were loaded onto the trapping cartridge with a solvent A set flow of 30 $mL \cdot min^{-1}$ for 3 min and eluted with a 0.3 $mL \cdot min^{-1}$ for 90 min of analysis time, starting with a 2-28% solvent B elution, then increased to 40% B, another 80% B washing step and finally re-equilibration to starting conditions. The LC system was directly coupled to the mass spectrometer using a Nanospray-Flex ion source and a Pico-Tip Emitter 360 µm OD x 20 µm ID; 10 µm tip. The mass spectrometer was operated in positive ion mode with a spray voltage of 2.3 kV and a capillary temperature of 275 °C. Full scan MS spectra with a mass range of 350–1,400 m/z were acquired in profile mode using a resolution of 70,000 [maximum fill time of 100 ms or a maximum of 3E6 ions (automatic gain control, AGC)]. Fragmentation was triggered for the top 20 peaks with charge 2 to 4 on the MS scan (data-dependent

acquisition) with a 20 s dynamic exclusion window (normalized collision energy was 26 eV). Precursors were isolated with 1.7 m/z and MS/MS spectra were acquired in profile mode with a resolution of 17,500 (maximum fill time of 50 ms or an AGC target of 1E5 ions). For the data analysis, the MS raw data were analyzed by MaxQuant 1.6.1[160]. *C. thermophilum* proteome sequences were downloaded from Uniprot with Proteome ID UP000008066. The MS data were searched against *C. thermophilum* proteome sequences plus common contaminants sequence provided by MaxQuant. The default setting of MaxQuant was used with modification oxidation and acetyl (protein N-term). A false-discovery rate (FDR) cutoff of 1% was used for protein identification, and iBAQ intensity was used for label-free protein quantitation. When calculating iBAQ intensity, the maximum detector peak intensities of the peptide elution profile were used as the peptide intensity. Then, all identified peptide intensities were added and normalized by the total number of identified peptides.

For the preparation of the crosslinking mass spectrometry samples, a titration to identify the optimal crosslinker concentration was first performed (**Supplementary Figure 1**). Fractions 3 to 9 from the *C. thermophilum* native lysate fractionation described above (**Supplementary Figure 1A, B**) were pooled to a total volume of 1.4 mL and protein concentration of 0.48 $mg \cdot mL^{-1}$, then split into 7 parts of equal volume in each Eppendorf tube. In the first tube, no crosslinking agent was added to be kept as control, while to the rest, 5 µL of 0.16, 0.32, 0.63, 1.25, 2.5, and 5 mM of the crosslinking agent BS_3 was added, respectively. The samples were incubated for 2 h on ice, and the crosslinking reaction was then deactivated with the addition of 50 mM NH_4HCO_3, and incubated again for 30 min on ice. The samples were then transferred in an acetone-compatible tube, and 4 times the sample volume of cold (-20 °C) acetone was added; the tubes were vortexed, again incubated for 60 min at -20 °C and centrifuged for 10 min at 15,000 g. The supernatant was properly removed, and then the tubes were left open at RT in order for the acetone to evaporate (**Supplementary Figure 1C**). To visualize the titration results, each of the 7 tubes' pellets were resuspended in 1X SDS-PAGE sample loading buffer (diluted from a 4X stock of 250 mM Tris–HCl pH 6.8, 8% w/v sodium dodecyl sulfate (SDS), 0.2% w/v bromophenol blue, 40% v/v glycerol, 20% v/v β-mercaptoethanol) to a final protein concentration of 10 and 20 $\mu g \cdot mL^{-1}$. The samples were then boiled for 5 min at 90°C, loaded onto Mini-PROTEAN® Precast Gels (BioRad, USA) and electrophorized for

60 min at 150 V. After the electrophoresis, gels were washed twice in water, stained with Coomassie staining solution and then destained until the background of the gel was fully destained. After visual inspection of gels, an optimal concentration of 1 mM BS_3 was selected to proceed with sample crosslinking (**Supplementary Figure 1D**). With optimal crosslinker concentration determined, fractions 3 to 9 from the *C. thermophilum* native lysate fractionation described above were pooled again, but this time in two pools, 3-6 and 7-9. In a Protein LoBind Eppendorf tube for each pool, 100 µL of 8M urea was added and then centrifuged overnight in a centrifugal vacuum evaporator at RT. 100 µL from each pool was then transferred to each of the tubes containing urea powder and 100 mM of ammonium bicarbonate (ABC) was added. DTT was then also added to a final concentration of 2.5 mM (from a 100 mM DDT stock dissolved in 100 mM ABC) and incubated for 30 min at RT. After the incubation was over, 2-Iodoacetamide (IAA) was added to a final concentration of 5 mM followed by incubation for 30 min, at RT in the dark. The reaction was quenched by the addition of 2.5 mM DTT to prevent modification of serine in the trypsin active sites. LysC (1µg·µL^{-1} stock, in 1:100 ratio, w/w) was subsequently added and samples were again incubated at RT for 4.5 h. An addition of 50 mM ABC to the sample reduced the urea concentration to < 2 M. Trypsin (from 1 µg·µL^{-1} stock, in 1:50 ratio, w/w) was introduced to the samples which were then incubated overnight at RT, followed by STop And Go Extraction (STAGE) TIPS desalting procedure, as described in[161]. The final crosslinked peptide samples were subjected to the same LC-MS/MS process and the results were analyzed as described above in the MS protein identification method. All data reported that resulted from mass spectrometry-related experiments is included in **Supplementary Table 2**. All plots, visualizing MS/XL-MS data were created with the xiVIEW online webserver[162].

2.2.8 Cryo-electron microscopy sample preparation and data collection

For structural characterization of the succinyl-CoA producing cell-free system, a 3.5 µL sample of final protein concentration of 0.3 mg·mL^{-1} was applied on a carbon-CoAted, holey support film type R2/1 on 200 mesh copper grid (Quantifoil, Germany) that was previously glow discharged under the following conditions: 15mA, grid

negative, 0.4mbar and 25 s glowing time with a PELCO easiGlow (TED PELLA, USA). The grid was then plunge-frozen with a Vitrobot® Mark IV System (ThermoFisher Scientific, USA) after blotting with Vitrobot® Filter Paper (Grade 595 ash-free filter paper ø55/20 mm). In the chamber, conditions were stabilized at 4 °C and 95% humidity, and after sample application, the grid was blotted for 6 s. The vitrified grid was clipped and loaded on a Glacios 200 keV Cryo-transmission electron microscope (ThermoFisher Scientific, USA) under cryo and low humidity conditions. Images were acquired with the Falcon 3EC direct electron detector and the EPU software (ThermoFisher Scientific, USA) in linear mode and total electron dose of 30 e-/Å^2. Before acquisition, the beam was aligned to be parallel and perpendicular to the sample, with a 2.5 µm diameter, while a 100 µm objective aperture restricted the objective angle. Complete acquisition parameters are listed in **Supplementary Table 3** and **Supplementary Table 4**.

2.2.9 Cryo-EM image processing

2.2.9.1 Exploratory CFS structural signatures EM reconstruction and identification

An exploratory analysis and reconstruction of the high molecular weight protein community members that were contained in the CFS was performed as follows, with all subsequent steps of cryo-EM data image analysis performed with the cryoSPARC 3.1 high-performance computing software[98] and its incorporated algorithms. A dataset of 2,808 raw movies was imported in the software suite and the incorporated patch motion correction (multi) and patch contrast transfer function (CTF) (multi) were used to correct for beam induced motion during acquisition and to estimate the micrographs CTF parameters respectively. From the motion/CTF corrected set of images, a 276,399 single particles initial set was selected with the blob picker, without using any starting reference. This particle set was then reference-free classified in 2D, with all single particles separated into 128 distinct classes. The classes that contained “junk” particles were discarded and the remaining particles were reclassified for 3 more rounds with the same parameters, repeating the process of discarding classes

containing noise, damaged particles or ice contaminations. After the final round of 2D classification, 4 distinct 2D structural signatures could be readily identified. These were then used as references for template picking in order to enrich each signature. The resulting particles belonging to each structural signature were reconstructed *ab initio* in 3D, separating the signature 1 and 4 particles in 2 classes, signature 2 particles in 5 classes and signature 3 particles in 3 classes. After *ab initio* reconstruction, the best reconstruction contained the following number of particles: signature 1 – 1,819 particles, signature 2 – 2,582 particles, signature 3 – 3,331 particles and signature 4 – 20,279 particles. The 3D reconstructed signatures were used as input for the Omokage search webserver[147], identifying them as the oxoglutarate dehydrogenase complex core (OGDHc, signature 1), pyruvate dehydrogenase complex core (PDHc, signature 2), fatty acid synthase complex (FAS, signature 3) and the pre-60S ribosomal subunit (pre-60S, signature 4) respectively. The 4 *ab initio* models were then 2D-projected and the resulting images were used again for another round of particle template picking in order to further enrich the particles contained in each signature, resulting in 1,819 particles for OGDHc, 7,825 particles for PDHc, 5,231 particles for FAS and 35,773 particles for the pre-60S respectively. The electron density maps belonging to OGDHc core, PDHc core and FAS were refined with the Homogeneous refinement (new) method, using the dynamic masking option. The pre-60S map was calculated and refined using the Local refinement method, employing the non-uniform refinement[163] and dynamic masking options. Final maps reached a resolution of 4.38 Å (FSC = 0.143) for the OGDHc, 3.84 Å (FSC = 0.143) for the PDHc, 4.47 Å (FSC = 0.143) for the FAS and 4.52 Å (FSC = 0.143) for the pre-60S maps respectively. Local resolution estimations for all maps were performed with the Local resolution estimation method of cryoSPARC[98].

2.2.9.2 Signature identification

The top-10 hits for all signatures were compared to the bottom-10 hits (N = 20, number of cross-correlation scores) from the first 100 hits returned by the Omokage search[147] with single factor analysis of variance (ANOVA) (P = 0.05) with the Analysis

ToolPak in Microsoft Excel and then visually inspected. All resulting *P*-values for signatures 1, 2, and 3 are as follows for the top-10 and bottom-10 comparisons: signature 1: 3.09E-14, signature 2: 2.20502E-12, signature 3: 4.23128E-17 ($P < 0.05$). Means and variance for the top-10 and bottom-10 groups for each signature are as follows: signature 1: 0.79026/0.69987 (top-10/bottom-10, means), 0.000172607/6.47789E-06 (top-10/bottom-10, variance); signature 2: 0.74119/0.59715 (top-10/bottom-10, means), 0.00074327/4.48056E-06 (top-10/bottom-10, variance); signature 3: 0.84787/0.70323 (top-10/bottom-10, means), 0.000215289/9.49E-07 (top-10/bottom-10, variance); As each time only 2 groups were compared, no *post hoc* test was performed. All boxplots were generated with BoxPlotR[136]. In the case of signature 4, due to the heterogeneity of the returned hits, top-10 hits were also projected in 2D in RELION 3.0[150], comprising of: *E. coli* MutS[164], *H. sapiens* ATM kinase ((EMD-9523) and [165]), equine infectious anemia virus EIAV CA-SP hexamer[166], *S. pombe* ATM/Tel1[167], *S. cerevisiae* SWI/SNF complex[168], *H. sapiens* γ-tubulin ring complex[169], *E. coli* pre-60S[170], *H. sapiens* CSN-N8-CRL4ADDB2 complex[171], along with the *ab-initio* reconstructed signature 4 map and then visually compared. All model fits were performed in ChimeraX 1.1[152].

2.2.9.3 Systematic fitting for the identification of signature 1

EMDB-0108[172] was lowpass filtered to 4.38 Å (contour level 0.71) and an electron density map was simulated from PDB ID: 2IHW[59] at the same resolution (contour level 0.36). Both maps were fitted in the reconstructed OGDHc E2 core map (contour level 0.825) 100 times ($N = 161$, total number of unique cross-correlation scores of fits), checking for statistically significant difference between the 2 fits, by comparing the two groups with single-factor ANOVA ($P = 0.05$), with the Analysis ToolPak in Microsoft Excel. The resulting P-value for the comparison is 0.38 (not significant at 95% confidence interval). Means and variance for each group are as follows: 0.676125/0.642242353 (vs EMD-0108/vs 2IHW means) and 0.067526/0.053296562 (vs EMD-0108/vs 2IHW variance). As only 2 groups were compared, no *post hoc* test was performed. All boxplots were generated with BoxPlotR[136].

2.2.9.4 Signature 1 final sequence identification

To resolve ambiguity between E2o and dihydrolipoamide acyltransferase (E2b) for signature 1 model reconstruction (sequence identity of 22%), both homotrimeric structures were predicted with AlphaFold2 using the ColabFold advanced notebook. Resulting models were fitted into EM reconstruction using "Jiggle-Fit this molecule with Fourier Filter" tool and refined in real space with all-molecule self-restraints at 5 Å distance cut-off using COOT version 0.9.2-pre. Finally, the model's backbones were used as input to the findMySequence[141] software to identify sequences in *C. thermophilum* proteome (taxonomic identifier 759272). The E2o sequence was selected based on calculated E-value. A similar analysis for FAS was performed and a sequence of both chains built into a map using findMySequence was unambiguously confirmed (E-values 6.1E-180 and 3.4E-93 for α- and β-subunits respectively). Sequence identification of the third reconstruction of the pyruvate dehydrogenase complex was also pursued. In this case, a relatively high resolution of the reconstruction of 3.8 Å allowed *de novo* interpretation of the map and an optimized modeling approach. For a manually selected map region corresponding to a trimeric subcomplex, a mainchain-only model was automatically built using ARP/wARP[135] with a sequence-independent loop building algorithm[173].

2.2.9.5 Pre-60S ribosomal subunit identification

For the identification of the ribosomal subunits included in the reconstructed pre-60S map (contour level 0.34), models PDB ID: 6LSS[174] and PDB ID: 3JCT[175] were fitted in the map. For visualizing the pre-60S, a superposition in ChimeraX via comparison to a bacterial pre-60S[170] and a complete yeast 60S subunit[176] was performed. All ribosomal subunits that were included in the pre-60S map, along with the rRNA, were isolated, combined in a new model and then a simulated electron density map was created at a resolution of 4.52 Å.

2.2.9.6 CFS core component EM reconstruction

After its identification in the exploratory dataset, structural analysis of OGDHc, the core component of the CFS with succinyl-CoA production capabilities, was implemented as follows, with all steps of image processing performed in the cryoSPARC high-performance computing software version 3.3.1[98]. A dataset of 25,803 movies with a pixel size of 1.568 Å/px was imported to a dedicated workspace. The in-software patch motion correction (multi) and patch CTF estimation (multi) algorithms were employed to correct for beam induced motion and calculate for the CTF parameters of the micrographs respectively. After dataset curation, 24,300 micrographs were selected for subsequent steps of the image analysis process. Templates were created from EMD-13844[123] with the create templates job (20 equally-spaced generated templates) and were employed for a picking job with the template picker, resulting in an initial dataset of 3,596,302 particles that was extracted with a box size of 326 Å and then reference-free 2D classified in 200 classes. From the initial 2D classification, 71,912 particles were selected from 4 classes and subjected to heterogeneous refinement, further refining the final particle set to 52,034 particles after discarding particles that resulted in mal-formed OGDHc E2o core structures. The final particle set was then symmetry expanded with octahedral (O) symmetry and employed for the final core reconstruction with the local refinement (new) job, resulting in a OGDHc E2o core map of 3.35 Å resolution (FSC = 0.143) (**Supplementary Figure 2A**). For the reconstruction of the complete OGDH complex, the 71,912 particles that were included in the core reconstruction were re-extracted with a larger box size of 452 Å in order to include signal for the subunits that are located in the periphery of the core. They were then re-classified in 20 2D classes and the classes that showed most prominent peripheral densities were used again as template for a new round of template picking, resulting in an initial particle set of 2,891,518 single particles. This particle set underwent 3 more rounds of 2D classification, always selecting towards class averages that displayed a robust core signal, but in addition to the core also displayed peripheral subunit signal, ending up with a set 52,551 particles that were finally used for a 3D classification job with 10 classes. Class 0, containing 5,178 particles and displaying the most well-resolved peripheral densities, was finally used for homogeneous refinement, resulting in a OGDHc map of 21.04 Å resolution (FSC

= 0.143) (**Supplementary Figure 2B**) which was then utilized for all structural analysis. All map visualization was performed with the ChimeraX[152] software package.

2.2.10 Atomic model building and refinement

2.2.10.1 Atomic models of structural signatures discovered during CFS probing

Atomic modeling of the three identified signatures in the CFS, namely signature 1: PDHc E2 core, signature 2: OGDHc E2 core and signature 3: FAS was based on initial complex structure predictions made by AlphaFold2[106] by running the, available online (https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb), ColabFold[137] "advanced" notebook . This version, in contrast to the "simple" notebook enables the user to perform model predictions of multimeric protein complexes. During setup of the ColabFold run, most parameters were kept at default, but the final-model AMBER[177] force-field relaxation parameter was enabled. Following generation of the predicted models for all three signatures, they were fitted into their corresponding EM maps using the "Jiggle-Fit this molecule with Fourier Filter" tool and real-space refined with all molecule self-restraints at 5 Å distance cut-off using COOT version 0.9.2-pre[138]. After initial refinement, the monomer of each model was extracted and then again expanded into a complete complex by using the symmetry operators identified directly from their corresponding EM map with the "phenix.map_symmetry" and "phenix.apply_ncs" tools that are implemented in the PHENIX suite[149]. The final, complete complex models of all signatures were then refined with the "phenix.real_space_refine" tool by using a starting model and non-crystallographic symmetry (NCS) restraints.

For the PDHc and OGDHc model reconstructions specifically, AlphaFold2 was again employed to obtain trimeric subcomplex predictions. The predictions were symmetry-expanded again and then an extracted monomer was checked for steric conflicts between residue sidechains manually in ChimeraX[152] and ISOLDE version 1.2.5[144]. After correction of any identified steric conflicts, the monomer was again

expanded to a complete complex and refined using the exact same approach described above.

Due to the limitations of the ColabFold advanced notebook, in order to generate predictions for the FAS model, the heterodimer was split in 6 overlapping fragments, show in **Table 8**:

Table 8: Sequence fragments of AlphaFold2-predicted FAS heterodimer models.

Uniprot ID	**Amino-acid Sequence Fragment**
G0S866 / G0S867	1-100 / 1,500-2,037
G0S866	320-1,640
G0S866	950-1,865
G0S867	1-600
G0S867	540-1,640
G0S867	1,100-2,000

The overlapping fragments were then fitted and refined, each one independently, in the EM map. After refinement, they were merged together in COOT as described above, using a homologous FAS structure as guidance (PDB ID: 6U5V[178]). Finally, the resulting FAS complex model was inspected for steric clashes using a representative heterodimer in ISOLDE[144], all clashes were resolved and finally the heterodimer was symmetry-expanded to the complete FAS complex once again and refined as described above.

2.2.10.2 Atomic model building and refinement of the high-resolution OGDHc E2o core

For refinement of the E2o core structure that was derived from the ~25,000 EM image data set at 3.35 Å resolution (FSC = 0.143), the initial model (PDB ID: 7Q5Q) was fitted into the cryo-EM density using ChimeraX and then refined using iterative manual refinement with COOT and real-space refinement with PHENIX with standard parameters. Visible density that could be mapped was extended towards the *N-ter* of the E2o, from Met195 to Glu188.

2.2.11 OGDHc-specific AI-based model generation and electrostatic surface calculation

A local installation of AlphaFold2-Multimer[134] was utilized in order to perform predictions of the E2o core vertex trimer that was modeled in the experimentally derived E2o map, the E1o dimer (Uniprot ID: G0RZ09) and E3 dimer (Uniprot ID: G0SB20) in complex with the Uniprot-annotated E2o LD domain (residue numbers 40 to 115, Uniprot ID: G0SAX9). All AI-predicted model validation metrics can be found in the **Supplementary Figure 3A, B**. In order to calculate and visualize the electrostatic surface potential maps the APBS electrostatics plugin[179] was used in PyMol (Schrödinger, USA). These models and the individual coordinate files of the predicted complexes were used for macromolecular docking calculations and refinements (see below).

2.2.12 Energetic calculations, macromolecular docking and interface residue frequency calculations

Two distinct protocols were applied for reported HADDOCK2.2 calculations: (a) HADDOCK refinement[180]. Here, the refinement protocol was applied to calculate and compare the energetics of interfaces between OGDHc components and its human

homologue as well as the AlphaFold2-generated interfaces formed by the LD. In this refinement procedure, only the water refinement stage of the HADDOCK protocol was performed that showed to qualitatively correlate calculated energetics with binding affinities for transient protein-protein interactions[181] skipping the docking step. For this, complexes were solvated in an 8 Å shell of TIP3P water. The protocol consisted of the following steps: (1) 40 EM steps with the protein fixed (Powell minimizer) and (2) 2X 40 EM steps with harmonic position restraints on the protein (k = 20 kcal·mol^{-1}·Å^{-2}). For the final water refinement, a gentle simulated annealing protocol using molecular dynamics in Cartesian space is introduced after step (2). It consists of: (1) Heating period: 500 MD steps at 100, 200 and 300 K. Position restraints (k = 5 kcal·mol^{-1} Å^{-2}) are applied on the protein except for the side-chains at the interface. (2) Sampling stage: 1,250 MD steps. Weak (k = 1 kcal·mol^{-1} Å^{-2}) position restraints are applied on the protein except for the backbone and side-chains at the interface; and (3) Cooling stage: 500 MD steps at 300, 200 and 100 K. Weak (k = 1 kcal·mol^{-1}·Å^{-2}) position restraints are applied on the protein backbone only except at the interface. A time step of 2 fs is used for the integration of the equation of motions and the temperature is maintained constant by weak coupling to a reference temperature bath using the Berendsen thermostat[182]. The calculations were performed with CNS[183]. Non-bonded interactions were calculated with the OPLS force field[184] using a cutoff of 8.5 Å. The electrostatic potential (E_{elec}) was calculated by using a shift function while a switching function (between 6.5 and 8.5 Å) was used to define the van der Waals potential (E_{vdW}). The structure calculations were performed on the HADDOCK web server at https://alcazar.science.uu.nl/ using the refinement interface. A total of 200 structures was generated for each complex. (b) HADDOCK flexible docking. The HADDOCK docking server was used, utilizing the guru interface. Here, distance restraints were used by applying the derived crosslinking data matching lysine residues of the LD and the E1, and E3, respectively. The distance restraints applied are included in the haddockparam.web files provided with this thesis. For the docking calculations, the guru interface was utilized by default, but search and scoring space were significantly expanded. This meant that it0 generated structures were increased to N = 10,000, and scored structures were increased to N = 400.

Calculations of frequent amino-acid residues involved in the binding of the LD were produced from the formed interfaces from the N = 400 final, water-refined

docking solutions. An interface residue is considered if it is in proximity of 5 Å from any other residue from the LD. Highly frequent residues reported are the ones that belong to the top quartile (above 25%) in the calculated frequencies per residue, and these were plotted with the BoxPlotR online webserver[136]. All obtained values that were used to generate the corresponding plots presented in this thesis are included in **Supplementary Table 5** for the HADDOCK scoring plots, and **Supplementary Table 6** for the interface residue frequencies.

2.2.13 Peripheral subunit fitting

The OGDHc complex map that was generated from the cryo-EM experimental data (see 2.2.9.6 above) was used to identify the placement of the peripheral E1o and E3 subunits, in complex with the E2o LD domain that were generated through AlphaFold2-Multimer predictions as follows: the map was displayed in ChimeraX and after fitting the E2o core in the center, it was segmented into a “core” region and an “external” density region with the segment map tool with SEGGER[151]. Then a fit search of 100 fits was performed for both E1o-LD and E3-LD complexes and cross-correlation (CC) values for each fit were listed and separated as fitting in the core or external density region. For statistical significance, the two CC fit groups were tested with single-factor analysis of variance (ANOVA) with the Analysis ToolPak in Microsoft Excel (Microsoft Corporation, USA) and values for the fits can be found in **Supplementary Figure 4A, B** and **Supplementary Table 7**. CC plots were plotted with the BoxPlotR online webserver.

2.2.14 Linker distance and rotational displacement calculations

C. thermophilum E2o linker distance calculations were performed as follows: (a) For the experimentally resolved distance measurements, after all peripheral subunits with bound LD were fit in the OGDH complex map, distance was measured with the “distance” command in ChimeraX, starting from the last resolved *N-ter* residue for each of the 3 E2o subunits of the visually inspected closest E2o core vertex trimer,

to the first *C-ter* residues of the modeled LD domain bound to either E1o or E3 dimers on the periphery. All values were then averaged, and individual measurements and standard deviations calculated can be found in **Supplementary Table 7**.

Theoretical calculations based on disordered linker length (73 a.a., Uniprot ID: G0SAX9) were then based on derived equations published by Marsh and Forman-Kay[185], Wilkins *et al.*[186] and George and Heringa[187] (**Supplementary Table 7, Supplementary Figure 5A**) Additionally, to further derive insights on disordered linker length based on experimentally resolved structures from the PDB, the complete PDB database, as of 1st Jun. 2022, was downloaded and a total of 191,144 mmCIF format files, containing more than half a million chains were analyzed. Missing linker regions with their corresponding sequences are identified from the "_pdbx_unobs_or_zero_occ_residues" entries in the mmCIF files and a total of 399,404 missing regions were recorded. For each missing region entry, the left and right observed amino-acids are identified using the atomic coordinate data in the mmCIF files. For every linker region, following properties are derived (a) the end-to-end Cα-Cα distance between the two observed residues, measured in Å, (b) the length of the linker region, and (c) the sequence of the linker region. Lengths of these missing regions varied from 1 amino-acid up to 3,736 amino-acids, and a sharp reduction of available PDB entries was observed upon increased length of linker sequence, with the trend visible in **Supplementary Figure 5B**. To derive better recapitulation and statistical analysis of collective properties, the files were grouped together by adaptively increasing the bin width of length of amino-acids (**Supplementary Figure 5C**). For 1 to 50 amino-acids, the bin width was kept at 1. For 50 onwards, the right bin-edge increases gradually, to 55, 60, 65, 70, 75, 100, 125, 150, 250 and 4,000. From entries falling in every bin, the mean value of the Cα-Cα distances was calculated, and distance at quantiles varying from 0.0 to 1.0 in steps of 0.2. The analyzed data is presented in **Supplementary Figure 5**. In **Figure 52** presented in **3.15**, the mean Cα-Cα distance, represented by black dots, was plotted against the length of the linker region and distances at quantile level (0.0 to 1.0, 0.2 intervals) was plotted as a color shade as annotated in the legend. For reference, a horizontal line (dotted blue) was plotted corresponding to 3.5 Å. Similarly, a line (solid blue) corresponding to 7 Å times the length of the missing region was plotted. The relation between the mean Cα-Cα distance (y) and the number of amino-acids of the missing

sequence (x) can be empirically characterized by a model function of the form: $y = A(1 - e^{-bx}) + 3.5$, where, constants A and b are observed to have values, 11 and 0.2, respectively. The model function was plotted as a red line. All plots related to PDB distance calculations (**Figure 52, Supplementary Figure 5B, C**) were generated with the Pandas package in Python 3.9. Values for all plots related to PDB linker distance calculations are included in the **Supplementary Table 8**. The bubble plot containing all theoretical and experimental distance calculations (**Supplementary Figure 5A**) was generated with the ggpubr 0.4.0 package in R.

Rotational displacement calculations of the *H. sapiens* E2o vertex trimer subunits vs. the experimentally resolved *C. thermophilum* E2o vertex trimer subunits (A, B, C) were performed as follows: first, *C. thermophilum* and *H. sapiens* (PDB ID: 6H05) E2o trimers were extracted and aligned upon subunit A in PyMol. Then, the "angle_between_domains" command was used to first calculate the angle between *C. thermophilum* subunit A and B. The same was done for the angle between *C. thermophilum* subunit A and *H. sapiens* subunit B. Then the values were subtracted. The same process was done for the subunit pair A and C, again keeping the rotation axis the same, aligned on subunit A. A visual representation of the measured domains can be seen in **Supplementary Figure 6**.

2.2.15 Network analysis and community identification

Identified proteins from the mass spectrometry were mapped in the provided higher-order assemblies described in the Kastritis *et al.* 2017[72]. These higher-order assemblies were further enriched utilizing homology search per Uniprot entry against the Protein Data Bank (PDB), to also include possible homologous subunits resolved in PDB structures after 2017. For this, the Blastp search algorithm was used (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins) by default against the PDB database. Next, networks and communities provided in Kastritis *et al.* 2017 acted as a basis for further expanding networks reported in this study. This was performed by mapping the proteins to those networks, identifying them in the STRING[188] database and expanding the identified interactions into a network with first- and second- shell interactors and communities. Then, *C. thermophilum*-specific local STRING network

clusters were retrieved and coverage per string cluster is reported by simple division of identified proteins over all proteins present in the STRING cluster. This was performed from these expanded networks via back-mapping to the MS and XL-MS data. Finally, networks were checked for biological significance utilizing the in-database enrichment detection criterion[189], showing that all 54 clusters (protein communities) derived here were significantly enriched in biological interactions. Visualization of the networks was performed with Cytoscape[139]. Recovery of the complete TCA cycle was inferred utilizing KEGG[190].

2.2.16 Multiple sequence alignment

Uniprot ID: G0S3G5 (*C. thermophilum* putative holocytochrome C synthase - sequence length: 382), Uniprot ID: Q7S3Z3 (characterized in [61] as the *N. crassa* Kgd4 protein - sequence length: 130), along with Uniprot ID: Q7S3Z2 (*N. crassa* putative holocytochrome C synthase - sequence length: 317) were downloaded in .fasta format from the Uniprot database and aligned with Clustal Omega[191] in two separate alignment pairs: G0S3G5 with Q7S3Z3 and G0S3G5 with Q7S3Z2 separately. The .aln files were downloaded and visualized with Jalview[145]. Alignment results are displayed in **Supplementary Figure 7**.

3 Results and discussion

3.1 From cell to cell-free: suitability assessment of *C. thermophilum* for cell-free system preparation via imaging and biochemical characterization

Imaging of *C. thermophilum* could display mitochondrial enrichment, as validated by laser scanning confocal and electron microscopy methods (**Figure 10A, B, C**). Mitochondria exhibit the expected lamellar, ribbon-like cristae, forming parallel stacks in cross section[192] (**Figure 10B, C**). A high abundance of mitochondria (**Figure 10B, C**) is consistent with the increased respiratory rates of thermophiles[193] and served as motivation to derive a cell extract enriched in mitochondrial activities where functional OGDHc presents itself in high abundance [79] with all known subunits (E1o, E2o, E3)[72].

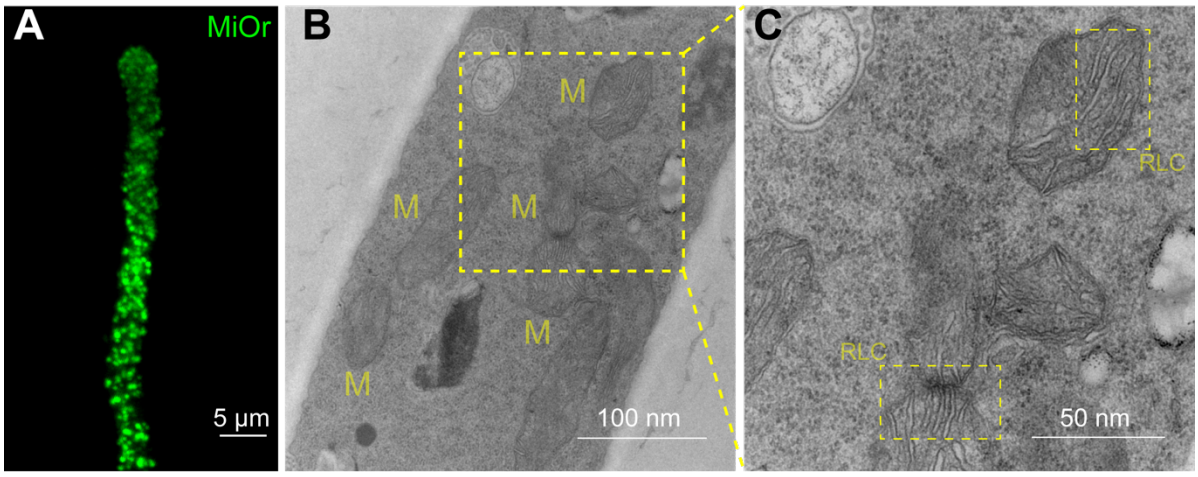


Figure 10: *C. thermophilum* mitochondria visualization.
(A) Fluorescent light microscopy of a single *C. thermophilum* filament, with the highly abundant mitochondrial content visible in green color, stained with MiOr. (B) Transmission electron microscopy image of cryo-fixated ultra-thin sections of a *C. thermophilum* filament. With "M" the mitochondria are annotated in the image. (C) Zoom-in into the mitochondria ultrastructures visible in (A) and (B). The ribbon-like cristae of the *C. thermophilum* mitochondria are visible (annotated as "RLC"). Figure reproduced from[194].

These results show that *C. thermophilum*, which was selected in this thesis as a model organism for understanding cell extract function during respiration is a valid choice due to its thermophilicity and mitochondrial abundance. Therefore, the next

step was to retrieve a cell extract with enriched functionality in respiration, in a similar fashion as performed in Kyrilis *et al.* 2021[79]. To this purpose, the thermophilic fungus *C. thermophilum* was grown at 52 ºC for 20 h before lysing and fractionating by size-exclusion chromatography (SEC) 8 g of cell fresh weight (**Methods**) (**Figure 11**). This SEC profile unambiguously shows minimal aggregation (low absorbance at 320 nm), and enrichment of biomolecular complexes (high absorbance at 260 and 280 nm).

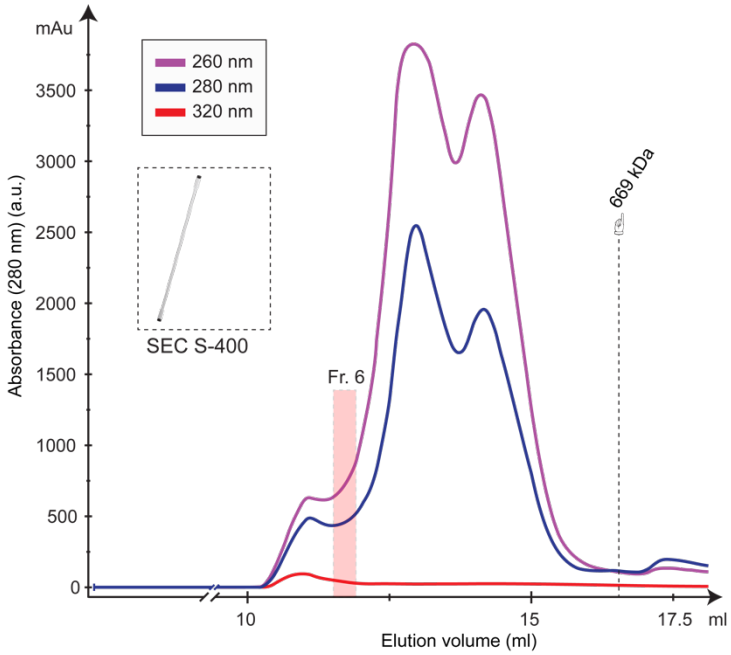


Figure 11: Size-exclusion chromatography profile of a native *C. thermophilum* cell extract.
Fraction 6 is located in the MDa molecular weight range. Figure reproduced from[123].

In the elution profile, all visible peaks can be attributed to protein complexes of molecular weights above 200 KDa, whereas the focus of subsequent analysis is given to fraction 6, which belongs to the MDa weight range of eluting fractions. In this fraction, initial cryo-EM screening could verify the presence of protein communities after improving image signal-to-noise ratio with denoising algorithms[195] (**Figure 12, Supplementary Figure 8**). These protein communities show higher-order assembly of cellular material that may form “beads-on-a-string” structures close to 100 nm. This result from cryo-EM shows that higher-order states of biomolecular complexes that are retrieved in the cell extract are retained, aggregates are rare, and it is amenable

to proteomic and cryo-EM characterization as well as subsequent deep analysis to identify its composition.

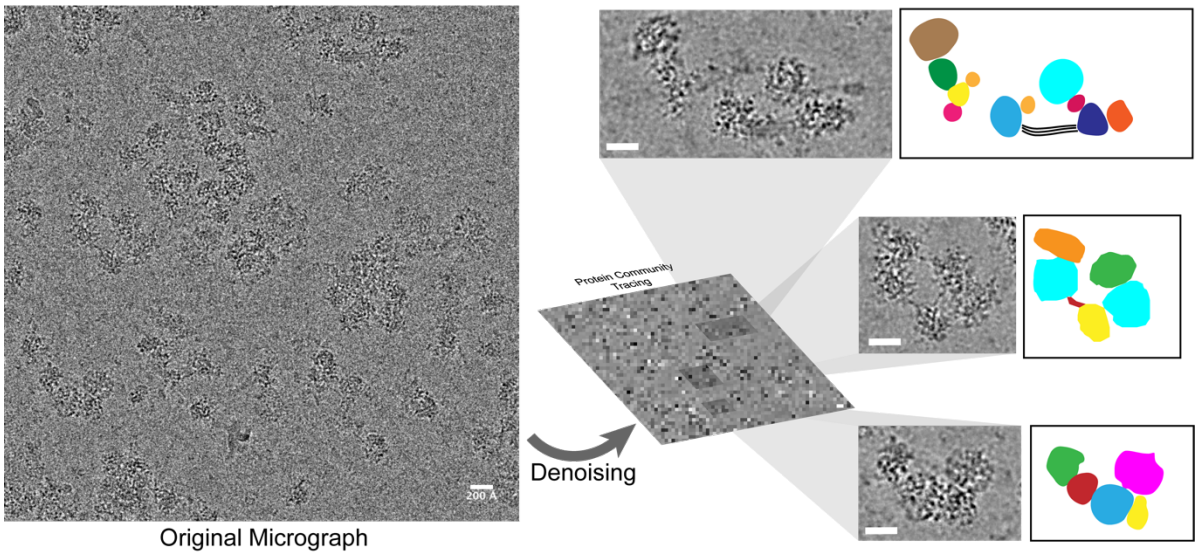


Figure 12: Detection of in-CFS protein communities.
Cryo-EM of fraction 6 allows, after denoising, to detect and trace various protein community assemblies that remain intact during fractionation and vitrification. Scale bars: 20 nm. Figure reproduced from[123].

Analysis of previously published mass spectrometry (MS) data[72] verified the in-fraction high abundance of 2-oxoacid dehydrogenase complexes that take part in the formation of protein communities with OGDHc specifically also belonging to this category. In general, the analysis revealed the in-fraction enrichment of five complexes that are involved in the formation of higher-order protein communities or metabolons, namely the:

- Pyruvate dehydrogenase complex (PDHc)
- 2-Oxoglutarate dehydrogenase complex (OGDHc)
- Branched-chain ketoacid dehydrogenase complex (BCKDHc)
- Fatty acid synthase (FAS)
- 60S ribosomal subunit and other ribosome-associated material.

MS-calculated abundance of the above-mentioned protein complexes in the fraction comprised 24.8% of the total protein abundance, from a total protein number of 1,281 different proteins that could be identified in the fraction and the high

abundance of in-fraction OGDHc designates fraction 6 as a promising candidate for employ as a succinyl-CoA producing CFS (**Figure 13**). This is particularly critical if the CFS is also active for this specific activity.

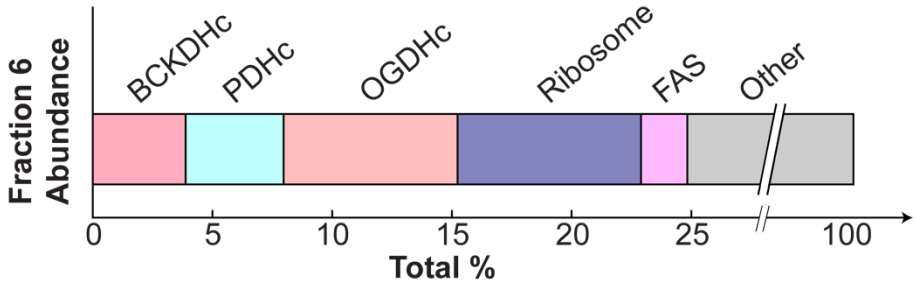


Figure 13: MS abundance of metabolons detected.
MS data allows for the calculation of the abundance of high molecular weight metabolons in the fraction, with OGDHc displaying high in-fraction abundance. Figure reproduced from[123].

Specifically, for targeting CFS Succinyl-CoA manufacturing capability of the retrieved CFS from the SEC, identification of all subunits with proteomics was known[72,79], but presence of OGDHc was further validated. Consequently, spurred by the initial analysis of MS-based abundance, the in-CFS presence of the E2 core-forming protein of PDHc (E2p) as well as all the enzymes that take part in the formation of the complete OGDHc (E1o, E2o, E3) with immunoblotting assays was identified (**Figure 14, Supplementary Figure 9A**).

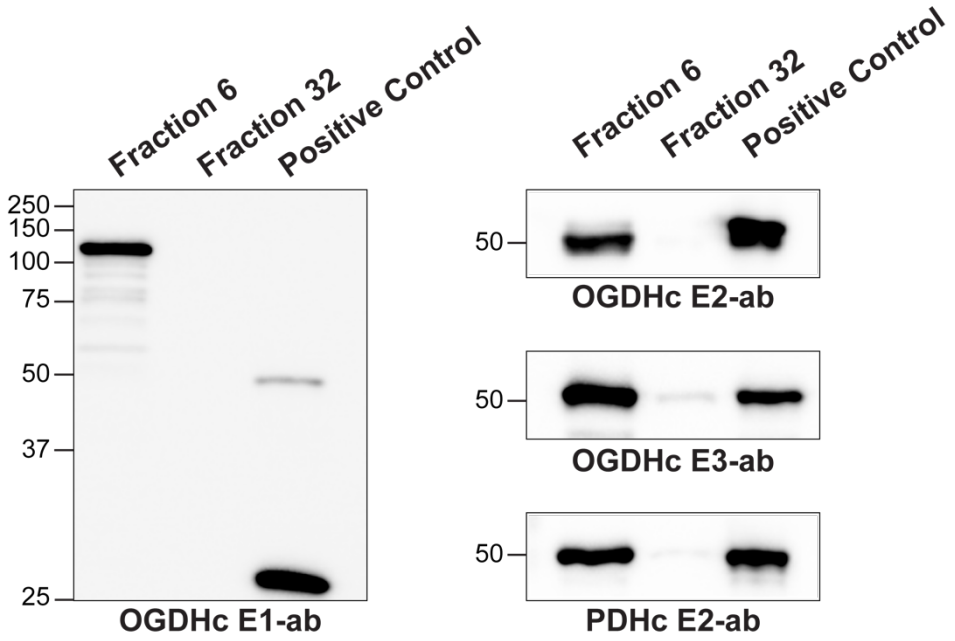


Figure 14: Immunoblotting identification of OGDHc and PDHc.
Western blots displaying the detection of the PDHc E2 core-forming protein and all OGDHc components in the cell-free system at the expected molecular weight (MW). A low-MW fraction was used as negative control, whereas the overexpressed protein that was used to create the antibodies was employed as a positive control. Due to the large size of the E1o protein, a fragment was employed for this reason and the same fragment was also used as positive control, hence the lower MW shown. Figure partly reproduced from[123,194].

Results in **Figure 14** unambiguously show that OGDGc is present as a full complex in the CFS that was retrieved. However, verifying the presence of the E2p, E1o, E2o and E3 proteins by themselves does not connotate the existence of an active complex, so a test for in-CFS enzymatic activity of the PDHc and OGDHc was performed by employing commercially available kits, thus verifying the non-compromised activity for the transformation of pyruvate to Acetyl-CoA and α-ketoglutarate to Succinyl-CoA (**Figure 15**).

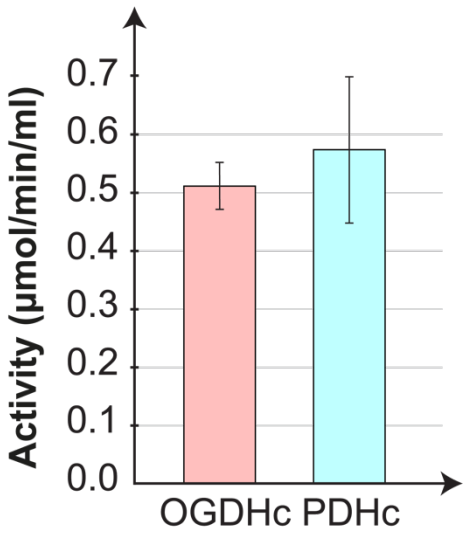


Figure 15: In-fraction OGDHc and PDHc activity assays.
Standard deviation is calculated for 3 technical triplicates. Figure reproduced from[123].

Results show that the CFS retrieved in this Thesis has dual activity (at least), producing both AcoA and succinyl-CoA, and therefore, comprises a promising system for storage of cofactors critical for protein acylating capacity[196]. Specifically, after ensuring that there is sufficient overall enzymatic activity for the OGDHc in-CFS, the complete kinetic parameters of all substrates involved in the enzymatic catalysis of α-ketoglutarate to succinyl-CoA were investigated, namely NAD+, α-ketoglutarate (AKG) and coenzyme A (CoA). OGDHc catalyzes the conversion of AKG to Succinyl-CoA through a multiple-step reaction involving different co-factors (**Figure 16**): thiamine diphosphate (ThDP) binds to the E1o, a lipoate covalently attached to lipoyl-binding domain (LD) of the E2o, whereas FAD binds to its respective site at the E3.

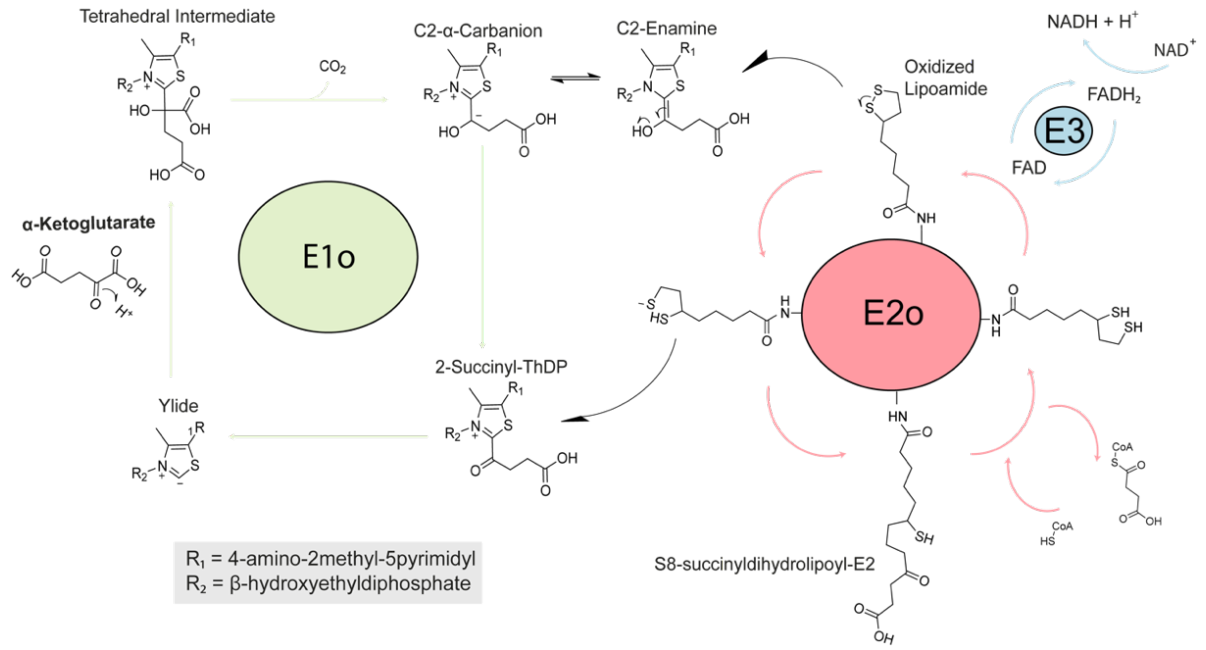


Figure 16: Complete reaction scheme of OGDHc.
The reaction that transforms α-ketoglutarate to succinyl-CoA is mediated by a lipoate that is covalently bound to the LD. The complete reaction requires the three distinct E1o, E2o, and E3 active sites in order to occur[197]. Figure reproduced from[194].

E1o and E2o are unique for this complex, but E3 is shared amongst all oxo-acid dehydrogenase complexes[64]. In-fraction K_M values were determined at [149.80 ± 41.78] µM, [146.11 ± 46.15] µM and [22.81 ± 11.93] µM respectively (**Figure 17, Supplementary Figure 9B, Supplementary Table 1**) and are comparable to those from other mesophilic counterparts[198,199]. Therefore, the efficient recovery of an active CFS is shown with succinyl-CoA manufacturing capability which is also accessible for determining exact kinetic constants.

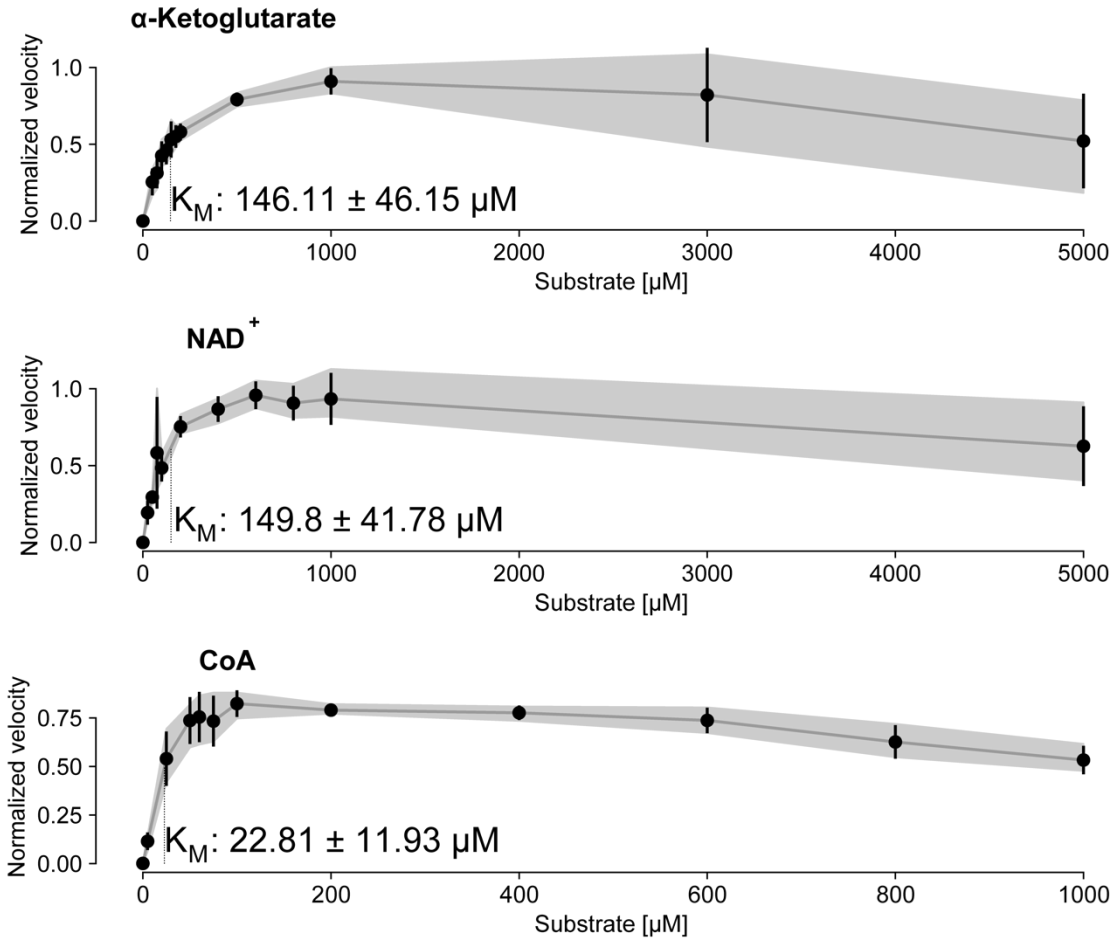


Figure 17: Enzymatic characterization of OGDHc present in the native cell-free system.
α-Ketoglutarate, NAD^+ and CoA were used at the concentrations shown in each plot accordingly, the velocity was normalized from 0 to 1 and a line is connecting each point. The K_M values shown in the plot were obtained by the Burk-Lineweaver plots shown in **Supplementary Figure 9B** and the gray background for each graph represents the standard deviation derived from N = 3 independent biological replicates and 2 technical duplicates for each replicate. All values shown here are listed in **Supplementary Table 1**. Figure reproduced from[194].

To accurately display the advantage of a CFS derived from a thermophilic eukaryote, the characterization was repeated and a comparison of the reaction velocity for AKG in a temperature gradient between a *C. thermophilum* and a *S. cerevisiae* equivalent sample (**Figure 18, Supplementary Table 1**) was also performed.

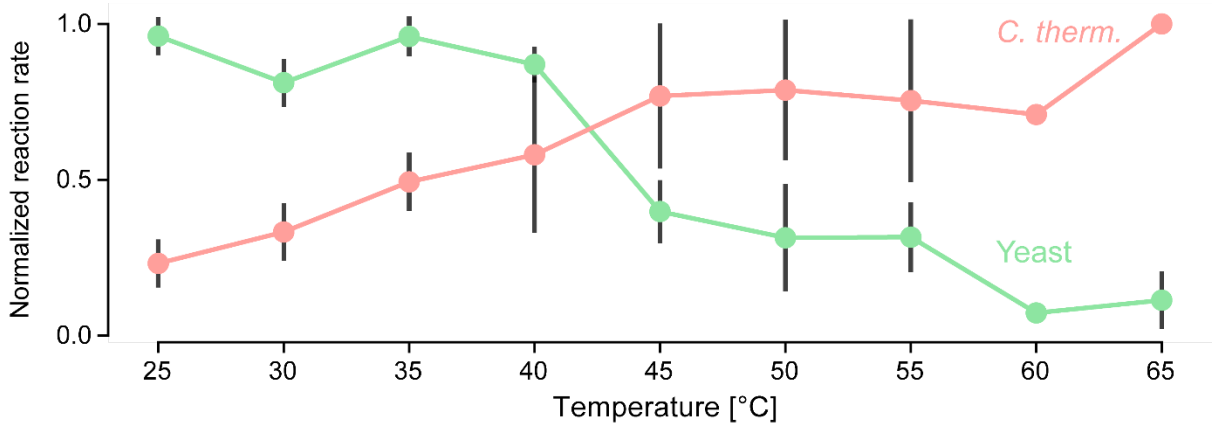


Figure 18: Comparative analysis of the E2o reaction velocity for AKG of a thermophile and a mesophile.
Change in reaction velocity (after normalization from 0 to 1) in relation to temperature for *C. thermophilum* as compared to a yeast equivalent sample. The black bars for each graph represent the standard deviation derived from N = 3 independent biological replicates and 2 technical duplicates for each replicate. All values shown here are listed in **Supplementary Table 1**. Figure reproduced from[194].

The derived data clearly displays an increase of reaction velocity for the *C. thermophilum*-derived CFS as contrasted by the velocity decrease of the yeast equivalent sample, showing the suitability of the CFS derived from a thermophilic organism for biotechnological application schemes that rely in high temperatures. The findings in this thesis demonstrate that a single cellular fraction recovered by SEC can be well-employed in CFS pipelines, is scalable due to its biochemical nature and is exploitable for product formation without the need for further purification or enrichment schemes. Most importantly, thermophilicity is retained in terms of activity within the cell extract conferring major advantages for Succinyl-CoA production at a wide range of temperatures.

3.2 A cryo-EM pipeline for the identification and *ab initio* map reconstruction of in-CFS heterogenous protein structural signatures

An initial ~2,800 movies dataset (**Supplementary Table 3**) of the CFS was collected in order to initially investigate the presence and recovery of protein structural signatures. This development is of critical importance, because its architectural characterization is essential to inform on its future improvement. For this purpose,

during data collection, a pixel size of 1.568 Å/px was selected to account for in-CFS metabolon and in general protein community recovery and improvement of per-image statistics. Template-free automated particle picking in the dataset verified the retrieval of numerous heterogenous single-particle 2D classes. A convolutional neural network (CNN) approach was utilized both for the picking and 2D classification of single-particles, which resulted in distinct structural signatures for multiple complexes, displaying high signal-to-noise ratios (SNR) (**Figure 19**).

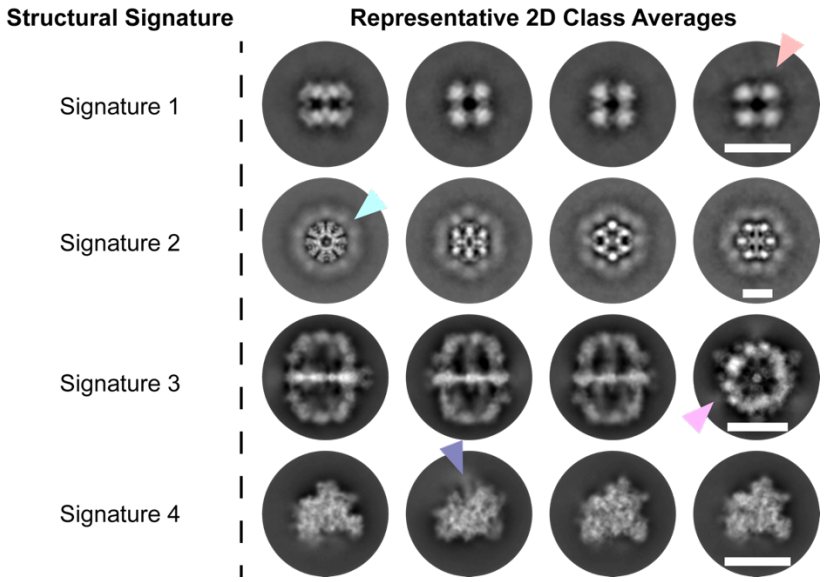


Figure 19: Representative 2D class averages of the most prominent in-fraction structural signatures.
Arrows denote diffused densities of lower signal, highlighting structural flexibility or potential binders. Scale bars: 20 nm. Figure reproduced from[123].

In these 2D class signatures, diffused, lower-resolution densities can be observed at the periphery of the central, high-contrast densities, demonstrating the presence of either conformational plasticity, external binders of high flexibility or other, non-readily identifiable subunits for each signature.

Proceeding with *ab initio*, asymmetric 3D reconstructions of the particles belonging to each 2D structural signature, again by employing CNNs, EM density maps with discrete and uniform densities that belong to four distinct large biomolecular assemblies are revealed (**Figure 20**).

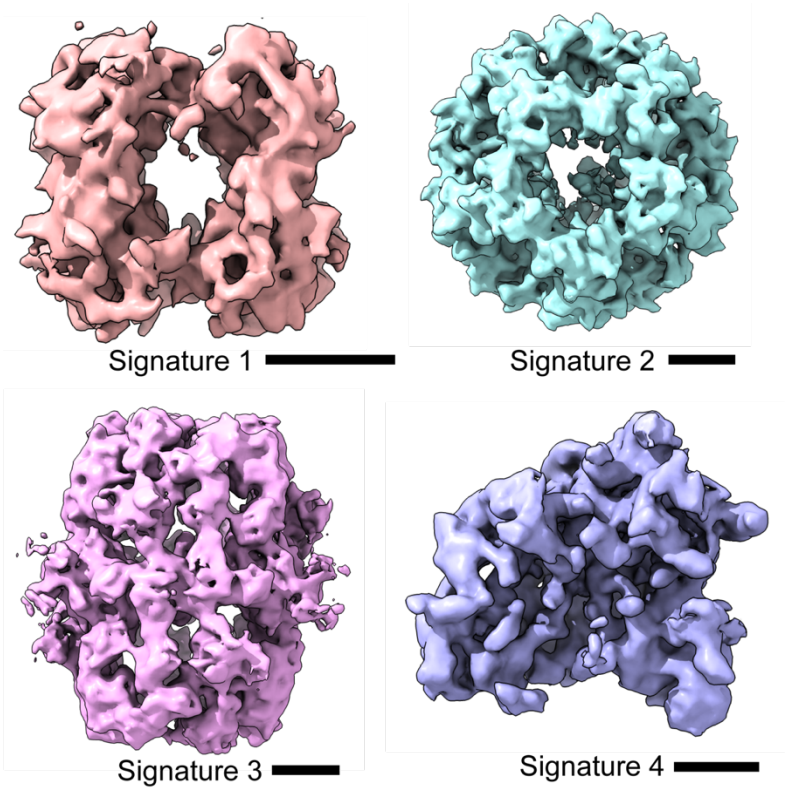


***Figure 20: Ab-initio* reconstruction of all four distinct structural signatures.**
Scale bars: 10 nm. Figure reproduced from[123].

The 3D *ab initio* reconstruction of each group of 2D class signatures can be described as:

- Signature 1: near-cubic (octahedral symmetry group – O)
- Signature 2: near-icosahedral (icosahedral symmetry group – I)
- Signature 3: near-dihedral (dihedral symmetry group – D2)
- Signature 4: asymmetric-like (lack of apparent symmetry group – C1).

Symmetry of the ab initio reconstructed maps was evaluated both by visual inspection and utilizing the ChimeraX function “measure symmetry”. All signature *ab initio* 3D reconstructions reached a resolution of ~20 Å (FSC = 0.143) (**Supplementary Figure 10**).

Accounting for the initial number of particles that were picked during the first round of template-free particle picking (N = 276,339), particles that belong to signatures one, two, three and four represent only 0.65%, 0.93%, 1.2% and 7.30% respectively

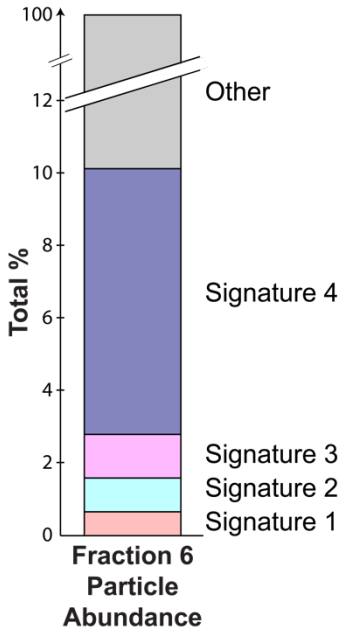


Figure 21: Signature particle abundance. Particle abundance for each signature is compared to total particles initially picked. Figure reproduced from[123].

(**Figure 21**), meaning that even complex structural signatures of very low particle abundance (as are the ones that are often encountered in complex cellular mixtures such as native cell extracts) can effectively be 3D asymmetrically reconstructed by utilizing cutting-edge image processing algorithms and software packages, with cryoSPARC being a prime example[98].

For the molecular identification of the four structural signatures (*i.e.,* to which exact protein complex each of these correspond to), two complementary strategies were applied:

1) The Omokage shape similarity search[147] was utilized to search for similar proteins included in available databases that matched the overall shape of the *ab initio* reconstructed 3D structural signatures, necessary without prior knowledge requirements of protein biochemistry, function, or stoichiometry. Statistical comparison between the top-10 versus the bottom-10 of the top-100 hits returned by the Omokage database in the case of signatures one, two and three reveals a significant identification (**Figure 22**).
2) Due to the asymmetric nature of signature four an extended identification scheme was required. This was because in the other cases the matching by Omokage, performed utilizing the principal component, was multiplied due to their inherent symmetry. However, lower cross-correlation is expected when matching an asymmetric cryo-EM map. All top-10 hits returned by the Omokage database and the signature four *ab initio* reconstruction were projected in 2D and visually compared to account for the heterogeneity of the returned results. Visual inspection, combined with MS data integration allowed for the signature's unambiguous identification (**Figure 22**, **Supplementary Figure 11**).

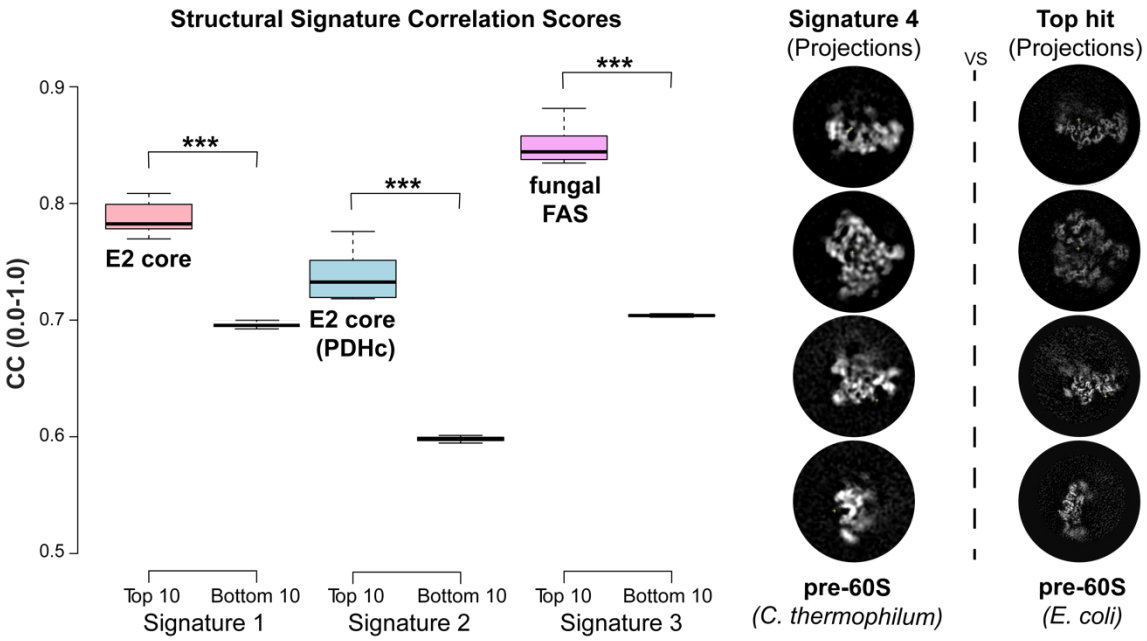


Figure 22: Cross-correlation comparison among top-10 and bottom-10 of the top 100 hits returned from the Omokage search for each signature.
2D projections of signature 4 compared to the matching top-10 hits returned from the Omokage search. P-values for the top-10 and bottom-10 comparisons: Signature 1: 3.09E-14, Signature 2: 2.20502E-12, Signature 3: 4.23128E-17 ($P < 0.05$). Figure reproduced from[123].

In detail, through external database integration and statistical analysis of returned results, the four structural signatures were identified as:

1) Signature 1: hybrid 2-oxoglutarate/branched chain ketoacid dehydrogenase complex E2o/b core
2) Signature 2: pyruvate dehydrogenase complex E2p core
3) Signature 3: fatty acid synthase
4) Signature 4: pre-60S ribosomal subunit

with another layer of validation of their identity provided by their MS abundance (**Figure 13**) and the in-fraction activity and kinetic assays cross-validating the presence and active state of PDHc and OGDHc.

3.3 Reconstruction of in-CFS identified protein community members at high-resolution

After identification, the EM maps belonging to the four signatures were refined at high resolution, spanning from 3.84 Å to 4.52 Å (FSC = 0.143) resolution (**Figure 23, Figure 24**). The high-resolution reconstructions for each set of single-particles was achieved by the application of straightforward refinement protocols without any need for convoluted post-processing. After refinement, all four maps displayed highly-resolved secondary structural features, such as α-helical pitch (**Figure 23A**), resolved interfaces between α-helical structural elements (**Figure 23B**), β-strand separation (**Figure 23C**) and identification of non-protein structural elements, *e.g.,* the ribosomal RNA components of the pre-60S ribosomal subunit (**Figure 23D**).

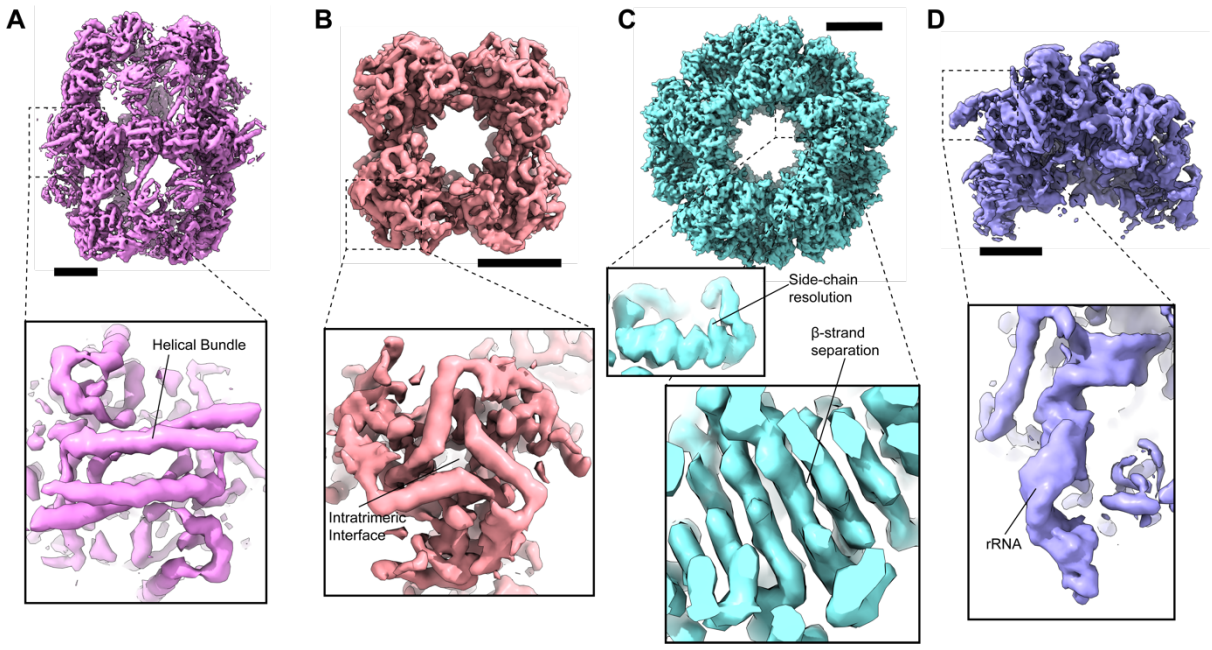


Figure 23: High-resolution signature reconstructions and visible features.
(A) Reconstruction of fatty acid synthase complex. α-Helical bundles and pitch are clearly visible. (B) Reconstruction of the hybrid oxoglutarate dehydrogenase/branched chain ketoacid dehydrogenase complex E2 core, where the intra-trimeric interfaces at the edge of the core are recapitulated. (C) Reconstruction of the pyruvate dehydrogenase complex E2 core. High-resolution structural features, such as side-chain densities and β-strand separation are identifiable. (D) Among other features, in the reconstruction of the pre-60S ribosomal subunit densities belonging to the rRNA structural elements are visible. Scale bars: 5 nm. Figure reproduced from[123].

Apart from overall resolution calculations for each signature EM map, a local resolution estimation was also performed. All maps displayed a fairly uniform resolution distribution (**Figure 24A-D**) with regions reaching a resolution of 3.4 Å. If the selected pixel size for data acquisition is taken under consideration (1.568 Å/px), this denotes that these regions have almost reached the data's resolution limit (Nyquist frequency).

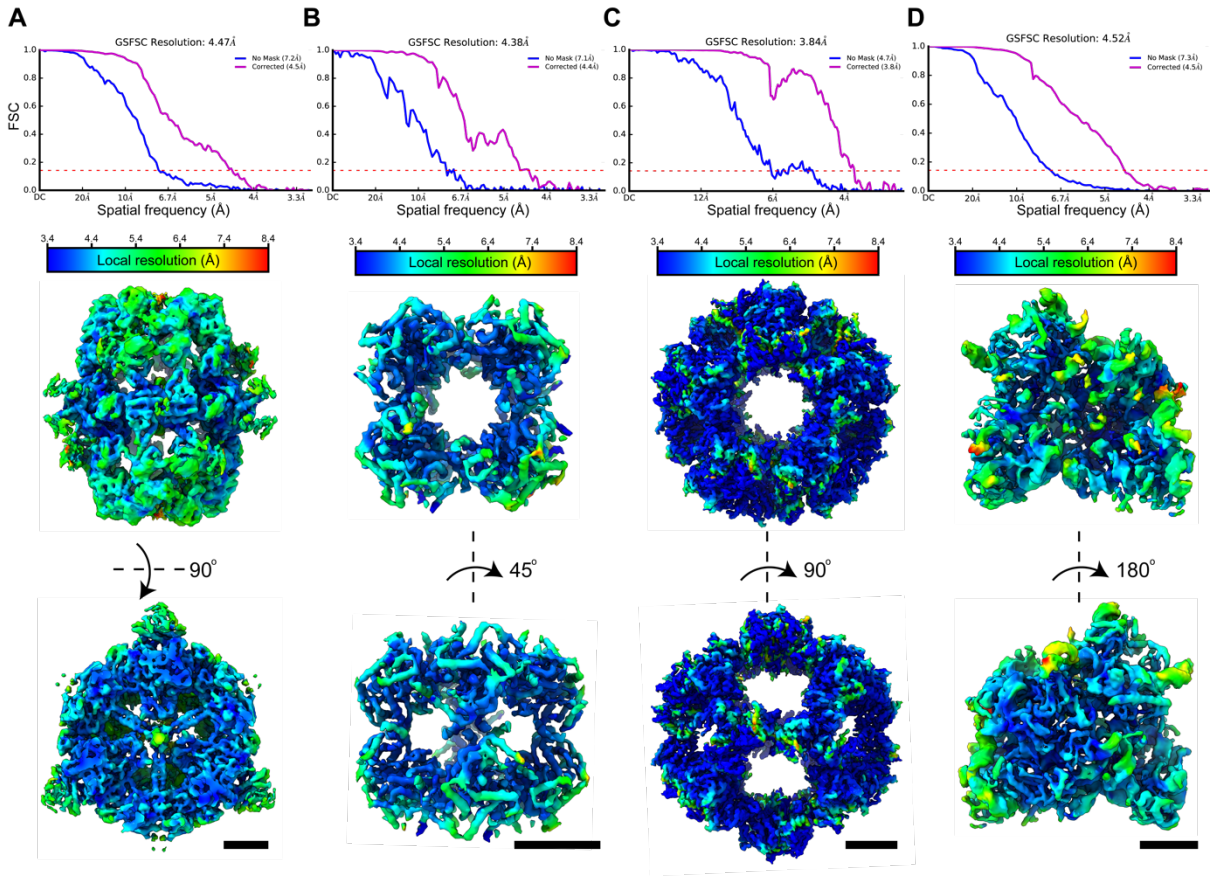


Figure 24: FSC plots and local resolution distributions for all reconstructed maps. (A) FSC resolution plot and local resolution distribution for FAS. Central core of the complex demonstrates an overall higher resolution when compared to the external, more flexible densities. (B) FSC resolution plot and local resolution distribution for OGDHc E2 core. External densities demonstrate more flexibility compared to the inner part of the complex' core. (C) FSC resolution plot and local resolution distribution for PDHc. The reconstruction demonstrates high and uniform resolution distribution. (D) FSC resolution plot and local resolution distribution for the pre-60S ribosomal subunit. More flexible, external densities corresponding to the rRNA components are recognizable. Scale bars: 5 nm. Figure reproduced from[123].

Nevertheless, regions can also be observed that, based on the local resolution estimation, display relatively lower resolution. The comparison between the two types

of regions allows for the differentiation between very stable, rigid regions and highly flexible regions since the former would be expected to display higher local resolution whereas the latter relatively lower due to their conformational (or chemical) heterogeneity. This can be exemplified by the FAS EM map (**Figure 24A**), where densities that are located along the horizontal symmetry axis on the periphery appear significantly more flexible when compared to the main “dome” ultrastructure. The lower local resolutions of flexible domains are not sufficient for direct protein modeling but nevertheless can provide significant insights concerning structure, function, domain interactions and are amenable to integrative modeling approaches, a fact that is applicable to all four reconstructed signatures (**Figure 24A-D**). Another intriguing fact to consider is the single-particle number that is included in the sets used for each reconstruction. In the case of signature four, the pre-60S, despite the comparatively high number of single-particles that were employed for its reconstruction (N = 35,773), it displays the lowest resolution out of all the reconstructed signatures, a clear sign of its high heterogeneity and flexibility. In contrast, a very low number of particles (N = 1,819) and octahedral symmetry allowed for a high-resolution reconstruction of the hybrid E2o/b core, clearly recapitulating secondary structural elements. The above-mentioned observations highlight the possibility of high-resolution, in-CFS reconstructions even for protein community members of very low particle abundance.

3.4 Conformational adaptations of the *C. thermophilum* PPT acetyl-CoA binding domain of FAS

The native *C. thermophilum* FAS has been the subject of previous studies by Kastritis *et al.*[72], but, due to the lower overall resolution of the reconstruction reported by the authors, the acetyl-CoA binding domain of the phosphopantetheinyl transferase (PPT) was not observable (**Figure 25**).

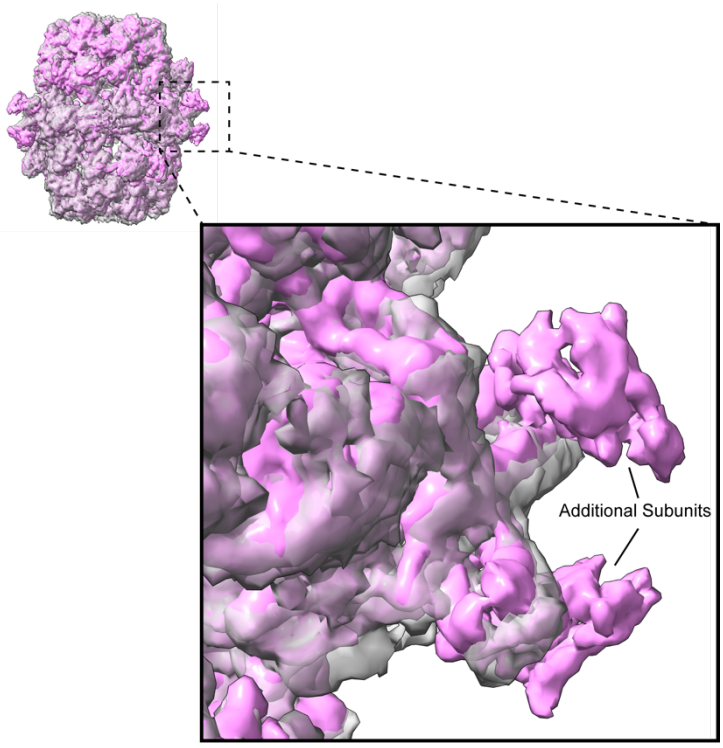


Figure 25: Structural insights into the FAS reconstructed map.
When compared to previously resolved *C. thermophilum* FAS, new densities can be identified. Figure reproduced from[123].

In the reconstruction that was obtained by analyzing the in-CFS collected data, the resulting map was of sufficiently high resolution to unambiguously resolve the domain itself, as well as its location in relation the overall FAS structure. In a work published by Singh *et al.*[200], the researchers were also able to resolve the PPT acetyl-CoA binding domain of the *S. cerevisiae* FAS. Since both FAS structures (*C. thermophilum* and yeast) display a similar overall organization, a direct comparison between the mesophilic and thermophilic FAS is possible, resulting in the observation of significant differences concerning the PPT acetyl-CoA domain. Sequence-wise, the domain is part of the FAS α-subunit *C-ter* and it is highly conserved. After real-space refining the *C. thermophilum* domain in the in-CFS derived EM map corresponding density and comparing it to the yeast ortholog, a distinct conformational change could be observed (**Figure 26**).

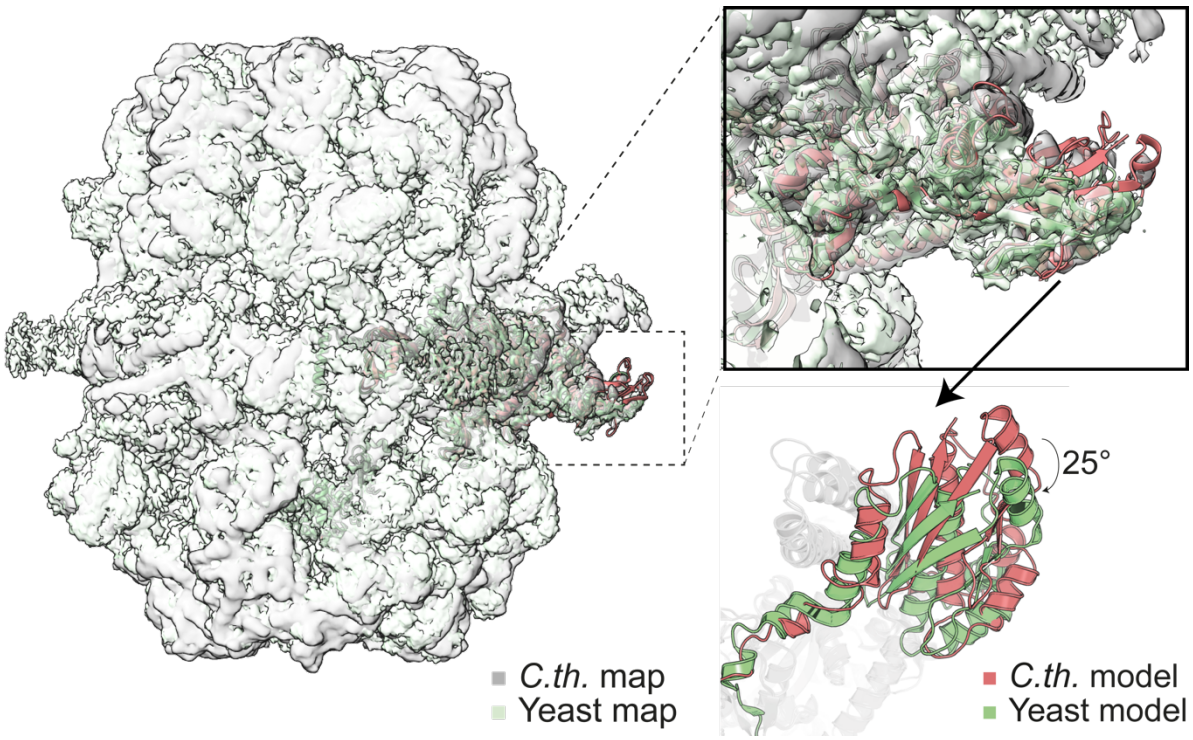


Figure 26: *C. thermophilum* FAS PPT domain displays a conformational change. Comparison of the Acetyl-CoA binding PPT domains of the *C. thermophilum* (*C. th.*) and *S. cerevisiae* (Yeast) reveals the lateral movement of the domain. Figure reproduced from[123].

This conformational variation can be attributed to a helix-turn-helix linker region that is accessible to the solvent and spans residues Ser1730-Lys1761 in yeast and Asn1709-Arg1737 in *C. thermophilum* FAS α-subunit (**Figure 27**). Even though both domain equivalents display a similar local fold, the *C. thermophilum* sequence of the linker region appears to be more flexible, allowing broader mobility for the linked domain. This larger movement space may translate to facilitated interactions between the domain and the ACP, improving the efficiency of the ACP 4'-pphosphopantetheine modification.

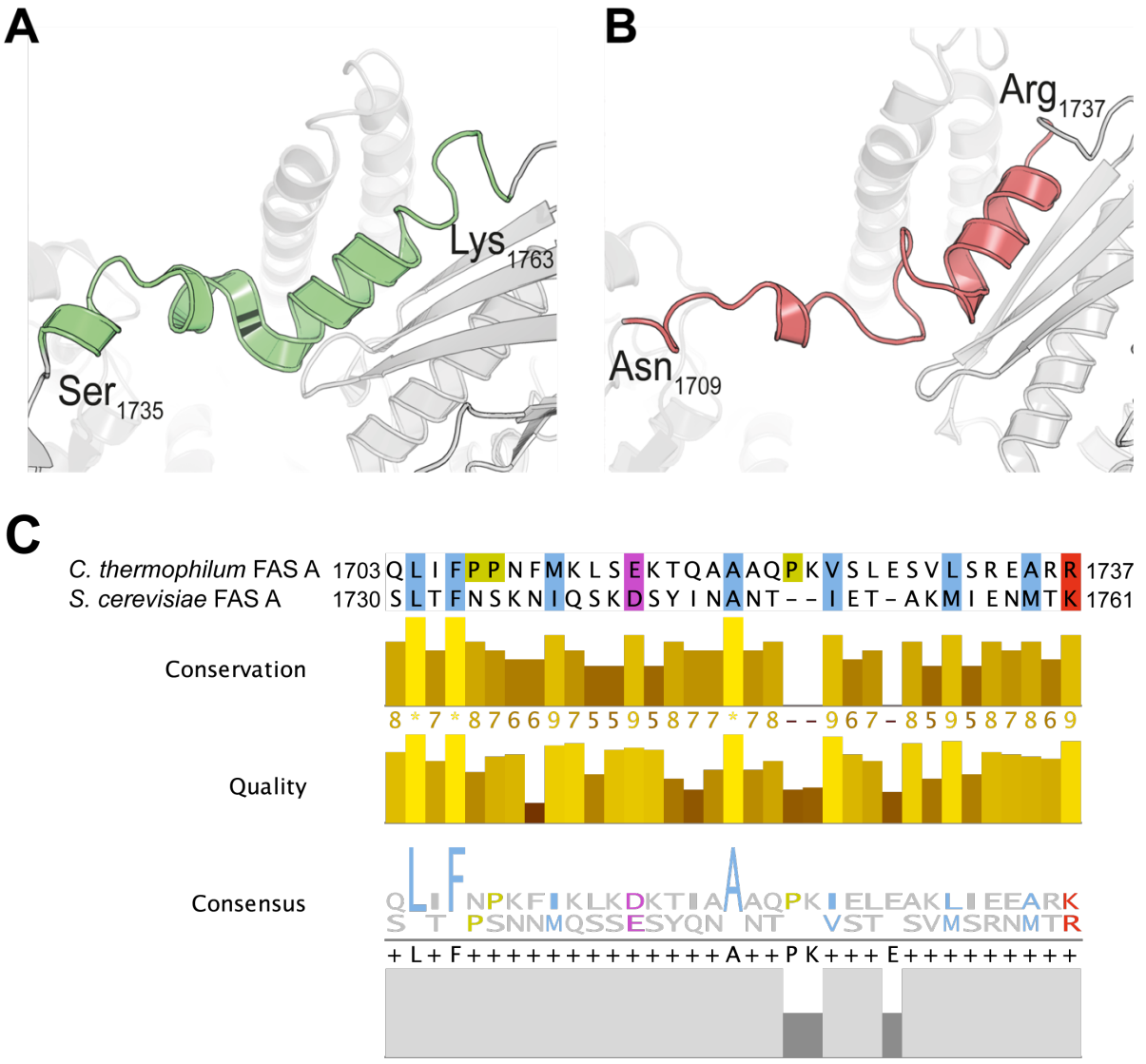


Figure 27: Comparison of linker regions between PPT acetyl-CoA binding domains of yeast and *C. thermophilum*.
(A) Flexible linker region of the Acetyl-CoA binding PPT domain of yeast. (B) Flexible linker region of the Acetyl-CoA binding PPT domain of *C. thermophilum*. Longer length of unstructured region may provide explanation for the domain's higher flexibility. (C) Alignment of flexible linker region sequences of the Acetyl-CoA binding PPT domain of *C. thermophilum* and yeast. Sequences were retrieved from Uniprot[201] and visualized in Jalview[145]. Figure partly reproduced from[123].

3.5 High-resolution symmetric reconstruction of PDHc E2 core reveals possible E3BP anchor points

Through careful investigation of the in-CFS reconstructed E2p core EM map, persistent electron densities are discernible on the inside of the core's ultrastructure, proximal to the E2p trimers (**Figure 28**).

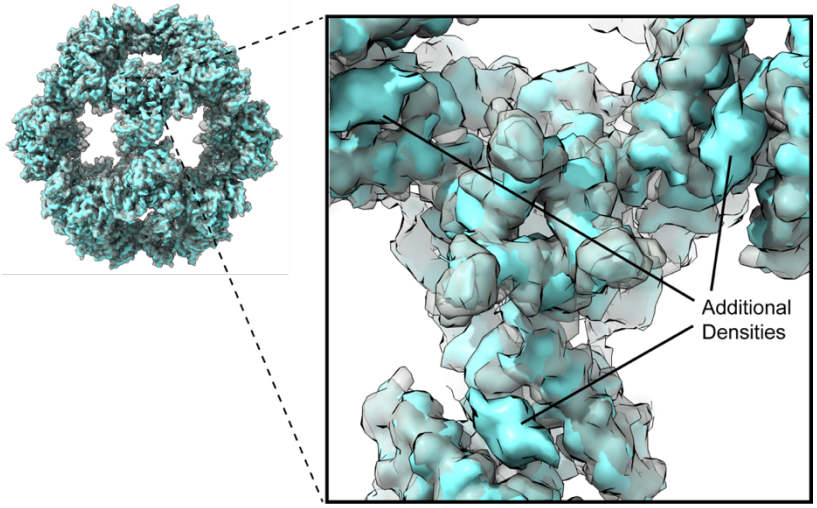


Figure 28: Identifying an E2p-E3BP interface.
Comparison to the previously resolved *C. thermophilum* PDHc E2 core reveals the existence of additional densities on the inside of the core, possibly indicating the anchor points of the E3BP. Figure reproduced from[123].

In previously reported EM reconstruction of the thermophilic E2p icosahedral core, similar densities were described as belonging to the E2-interacting interface of the E3BP[79], and were also observed in the mesophilic E2p core[202]. The current E2p core reconstruction contains a similar interface and acts not only as validation of the previously published observations but also hints at the presence of an E2p-E3BP internal interface that is conserved across fungal species. Interestingly, the previously published cryo-EM *C. thermophilum* E2p core map[79] did not display this internal interface density, something that was attributed to the symmetry imposed during reconstruction. In this reconstruction, even though again symmetry was imposed, sufficiently high resolution revealed the averaged feature. This contradiction could prove hopeful for the future structural investigations of protein communities as, even

by imposing symmetry during EM reconstructions, at higher resolutions it will not pose a barrier for the discovery of lower-symmetry or asymmetric structural elements that leave such traces in the final reconstruction. Of course, this would also require careful evaluation of any observation, as to verify that any observable density is not the result of symmetry artifacts.

3.6 A native, low-abundant, α-ketoacid dehydrogenase hybrid E2 core can be recovered in the CFS

Utilizing the notably low number of just 1,819 single-particles, a cubic EM reconstruction belonging to a native protein community member was possible. The achieved resolution of 4.38 Å (FSC = 0.143) for the reconstruction revealed a clear cubic shape that could be identified as belonging to an E2-like protein assembly with octahedral symmetry, comprised of 24 copies of the same protein. The current reconstructed map also displays an ordered *C-ter* domain that matches both to the *in vitro* reconstituted resolved maps of the human OGDHc E2 core [172] and the bovine BCKDHc E2 core[59].

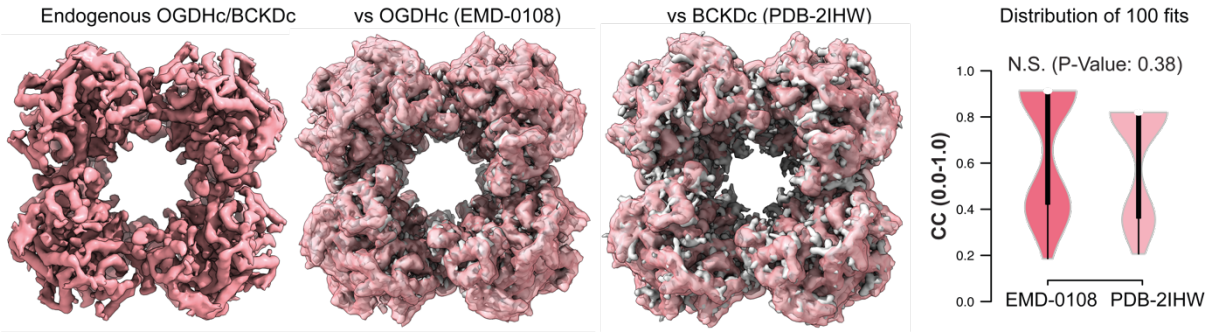


Figure 29: Signature 1 resolution is insufficient for fit-based identification. Distribution of fits with overexpressed OGDHc and BCKDHc E2 core maps (gray) does not allow the unambiguous identification of the endogenous reconstructed hybrid OGDHc/BCKDHc map (salmon). Figure reproduced from[123].

Despite the fairly high resolution of the reconstruction, the map itself could fit with high cross-correlation scores both the placement of the E2o core (CC_{maxE2o} = 0.91) and the E2b core (CC_{maxE2b} = 0.8) and statistical analysis of a fit-score distribution

between the two possible targets could not reach a statistically significant conclusion (**Figure 29**), despite the relatively higher CC values of the E2o fits. The native fungal α-ketoacid dehydrogenase hybrid core displays a cubic architecture and a lack of inner core densities is apparent, as contrasted by the native PDHc E2 core reconstruction.

3.7 Organization of the flexible, thermophilic pre-60S ribosomal subunit

The in-CFS derived asymmetric reconstruction of the thermophilic pre-60S ribosomal subunit reached a resolution of 4.52 Å (FSC = 0.143), and included the highest number of single-particles attributed to each signature (N = 35,773). As an additional validation measure, the *S. cerevisiae* and *H. sapiens* previously resolved pre-60S ribosomal subunits[174,203] were fitted in the current map and displayed acceptable fits (**Supplementary Figure 12**). Based on overall shape insights derived from a previously published yeast equivalent[204], the *C. thermophilum* pre-60S appears to be at the nucleolar assembly state. The sufficiently high resolution allowed for the unambiguous identification and localization in the EM map not only of the rRNA component, but also for 22 distinct ribosomal proteins, 4 of which were identified from the human equivalent pre-60S[174] and 18 from the yeast equivalent[203] (**Figure 30**).

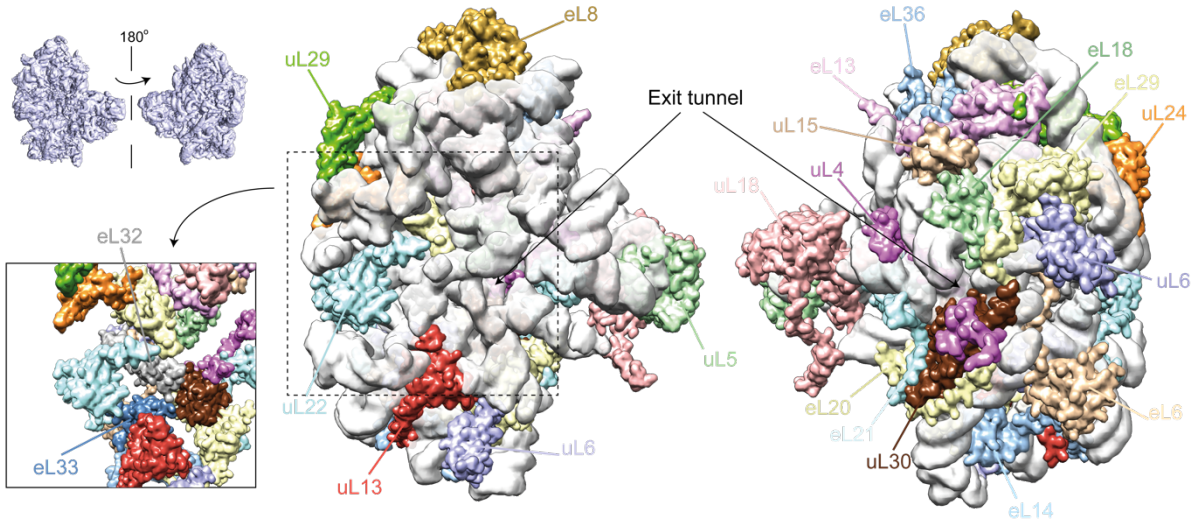


Figure 30: Mapping the components of a thermophilic pre-60S ribosomal subunit. Identifiable ribosomal subunits taking part in the assembly of the endogenous pre-60S ribosomal subunit. The exit tunnel for newly synthesized polypeptide chain is one of the first features visible during the 60S assembly sequence. The rRNA component of the pre-60S is shown with a transparent white color. Figure reproduced from[123].

As per the standardized ribosomal protein nomenclature system[205], the *C. thermophilum* pre-60S ribosomal subunit is assembled by the uL4, uL5, uL6, uL13, uL15, uL18, uL22, uL24, uL29, uL30, eL6, eL8, eL13, eL14, eL15, eL18, eL20, eL21, eL29, eL32, eL33 and eL36 ribosomal proteins. Looking further into the gene ontology (GO) terms associated with each of the ribosomal proteins that comprise the thermophilic pre-60S, uL24, uL30, eL14 and eL33 seem to be associated with processes involved in ribosomal biogenesis (GO:0042273), whereas uL5, uL18 and eL6 appear to be associated with the yeast large subunit ribosomal assembly (GO:0000027). After fitting and annotating all possible ribosomal proteins in the CFS-derived pre-60S EM map (**Figure 30**), they appear to organize centered around the construction of the protein exit tunnel, hinting that its early formation may be of importance for the overall assembly as it is possibly the first ribosomal "ultrastructure" that appears during its immature assembly.

3.8 A pipeline for the *de novo* identification of native protein community members

Despite the utilization of global shape similarity search algorithms (*e.g.,* Omokage, **Results 3.2**, **Figure 22**), which compare an unknown structure to all deposited known structures and maps in PDB and EMDB, the ideal scenario would be the deployment of a statistically independent measure for the direct identification of cryo-EM maps. In this scenario, the only information that would be used is the inherent characteristics of the map, such as density to residue-specific matching, and the identification would be performed completely *de novo*, without the need for external database searches. As it was shown in the case of the signature 1 identification, the final resolution of 4.38 Å (FSC = 0.143) reached for the reconstruction would still not be sufficient for an unambiguous identification based on shape matching. The reason would be that, even if primary sequences may differ significantly, the overall fold of two different protein complexes can still remain quite similar, thus occluding their distinction based on shape. This problem was exemplified during signature 1 identification, as the shape similarity search could not statistically distinguish between the possible targets of the human oxoglutarate dehydrogenase complex core (E2o)[172]

and the bovine branched chain ketoacid dehydrogenase complex core (E2b)[59]. Looking specifically into the *C. thermophilum* equivalent protein sequences (Uniprot[201] G0SAX9 and G0S0D3 respectively), their sequences present an identity of ~22 %, but the 3D structures of the proteins are very similar.

To tackle shape-based identification ambiguity, an identification pipeline that relies solely on the map's intrinsic information was developed (**Figure 31**).

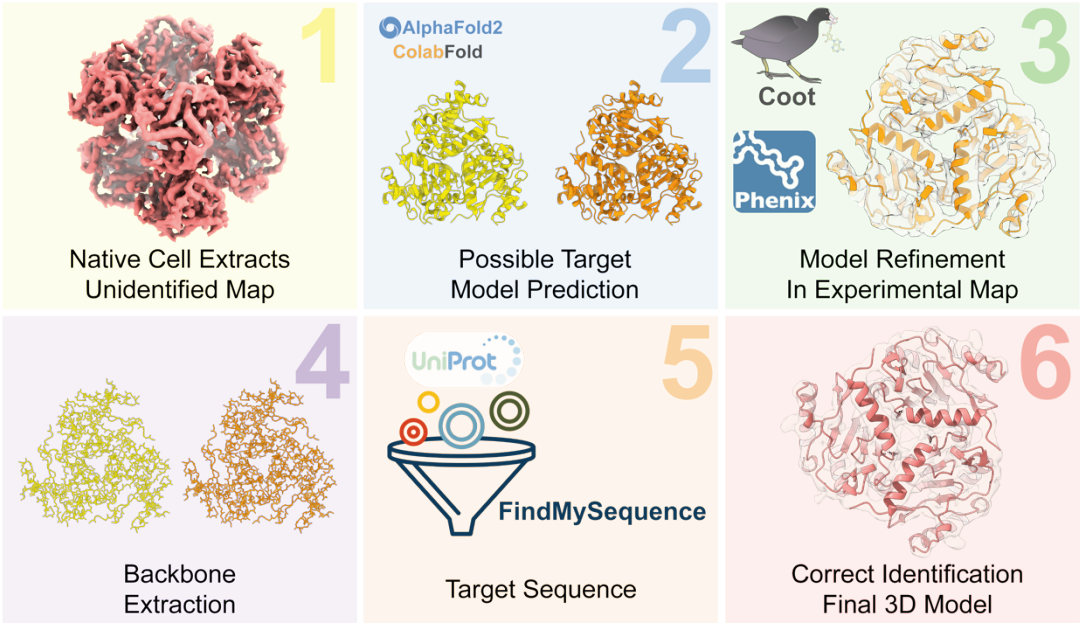


Figure 31: Scheme illustrating the workflow employed for the *de novo* identification and reconstruction of the OGDHc E2 core model derived from native cell extracts. Figure reproduced from[123].

For the first step of the process, the protein structure prediction AI software AlphaFold2[106] was employed in order to predict the trimeric structures of both possible identification targets (E2o and E2b) in the ColabFold advanced notebook[206]. After prediction of both models, they were real-space refined into the CFS-derived EM map of signature 1. The models' protein backbone were then extracted and, along with the EM map, were fed into the findMySequence[141] algorithm which could identify the corresponding sequences in the publicly-available proteome of *C. thermophilum*, selecting in the end as a definite match the E2o sequence with a significantly higher scoring in comparison to the E2b sequence (E-value 67.9E-30 for the E2o and 1.7E-3 for E2b, lower is better).

The unambiguous identification of E2o is an interesting observation as this means that E2b particles are completely absent from the final reconstruction, despite the MS abundance results. This could be attributed to various parameters, such as the BCKDHc metabolon stability, preferential orientation and presence in vitreous ice due to sequence-based surface properties or increased flexibility that would contribute to the non-inclusion of E2b core particles in the high-resolution classes and reconstructions during the cryo-EM image analysis process. Nevertheless, this result shows that identification and separation of proteins with very similar structural features can be achieved even at resolutions above 4 Å with high confidence, improving identification of EM reconstructions derived from complex mixtures.

3.9 Insights into the identification and structural characterization of protein community members by employing AI for the atomic modeling of cryo-EM maps.

The CFS-derived cryo-EM maps for protein community members PDHc E2 core, OGDHc E2 core and FAS reached resolutions that allowed *de novo* atomic modeling by employing the latest advances in AI-guided protein structure prediction[106,107]. Due to their multimeric nature, AlphaFold2[106] within the ColabFold[206] advanced notebook was used to build their multi-subunit atomic models and unambiguously fit them into their corresponding EM maps. Validation metrics show that the AI-built and refined final models are of sufficient quality (**Supplementary Figure 13, 14, Supplementary Table 3**).

Despite the validated accuracy of the AI-derived model reconstructions, differences exist between the before- and after-experimental refinement models. In the case of PDHc and OGDHc E2 core trimeric sub-complex predictions, the models were steric clash-free and recapitulated the most important intra-and intermolecular interfaces. Comparison of the before and after refinement models for the PDHc E2 trimer reveal an interesting feature. The pre-refinement model has a backfolded *N-ter* helix on top of each monomer of the E2p trimer, which is completely absent from the experimental cryo-EM map (**Figure 32**). This means that either the element is

completely absent from the native active structure or it is only transiently present during acetylation of coA performed by the E2p (**Figure 32**).

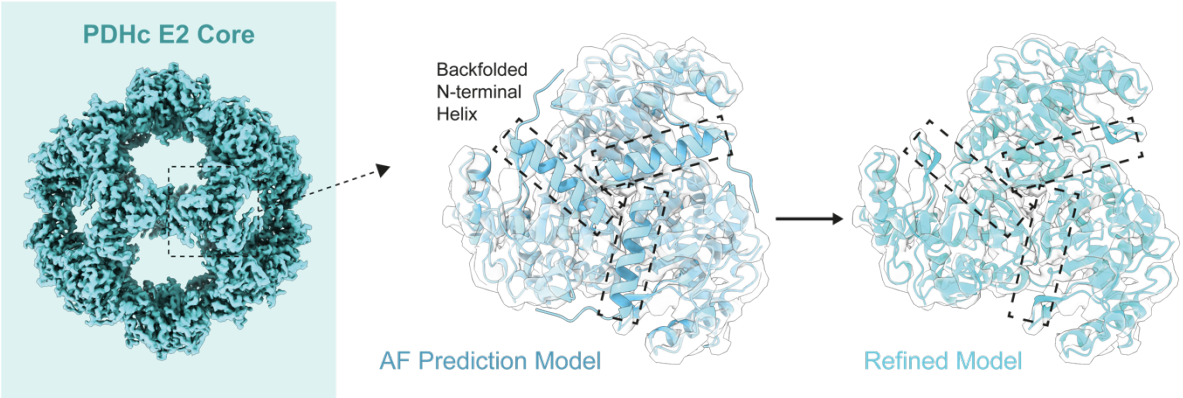


Figure 32: AI-prediction vs. experimental data: PDHc.
In the case of PDHc E2 core trimer, AlphaFold2 predicts a backfolded N-ter helical structural element (left) which is absent from the model that is refined in the experimental map (right). Figure reproduced from[123].

In the AI-derived OGDHc E2 core trimeric sub-complex, AlphaFold2 predicts the *N-ter* lipoyl domain (LD) stably bound in close proximity to the core (**Figure 33**). The LD is responsible for the transfer of the reaction intermediates amongst the protein subunits that comprise the OGDHc metabolon, meaning that this interaction is reasonable from a biological perspective and must happen for the reaction to proceed. AlphaFold2 correctly predicts their direct interaction interface but density to recapitulate it is absent from the experimental cryo-EM map, hinting at the transient nature of the interaction, a fact also supported by the diffused signal visible on the periphery of the E2o core (**Figure 33**). The LD-E2o interaction could possibly represent either a low-stability encounter complex or a stable, bound conformation only until the finalization of the catalytic intermediate's succinylation by the E2o active site.

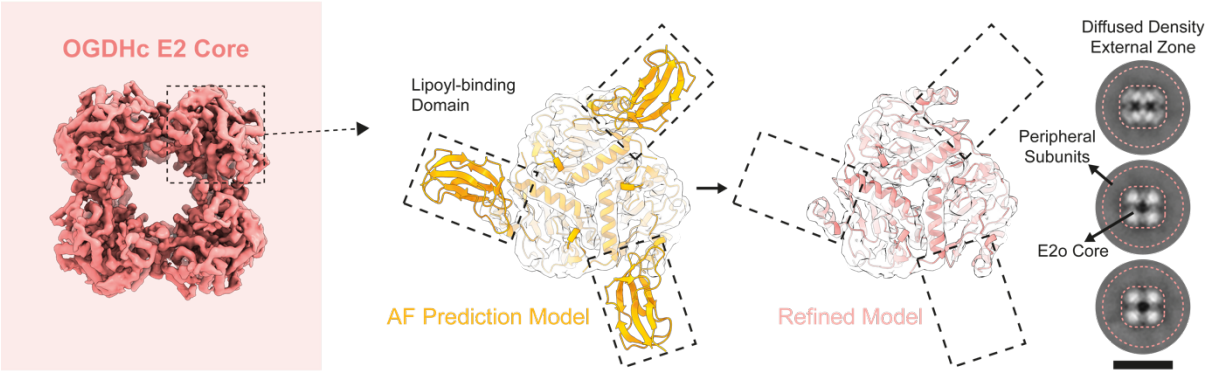


Figure 33: AI-prediction vs. experimental data: OGDHc.
A lipoyl-binding domain is predicted to be bound on the OGDHc E2 core trimer by AlphaFold2, where the same is clearly absent from the experimental map data. 2D class averages of OGDHc particles show the same, where there is a strong signal for the E2 core but diffused and weak signal for the peripheral subunit densities. Scale bar: 20 nm. Figure reproduced from[123].

The fatty acid synthase (FAS) basic multimeric unit is a heterodimer, which, due to its sheer size and memory limitations cannot be modeled as a whole and had to be split into six overlapping fragments in order to be predicted, then refined and merged into the complete structure. Despite the difficulty of inter-chain interface prediction, a confident model of the MPT domain was obtained. This domain would be very difficult to model due to the map's resolution constraints, as it is comprised by tightly interacting α- and β- subunits. At this domain, after D2 group symmetry application on the heterodimer to retrieve the complete complex, a central interface within the FAS "cage" is completely recapitulated (**Figure 34**), and can be almost perfectly fitted in the map without drastic refinement schemes.

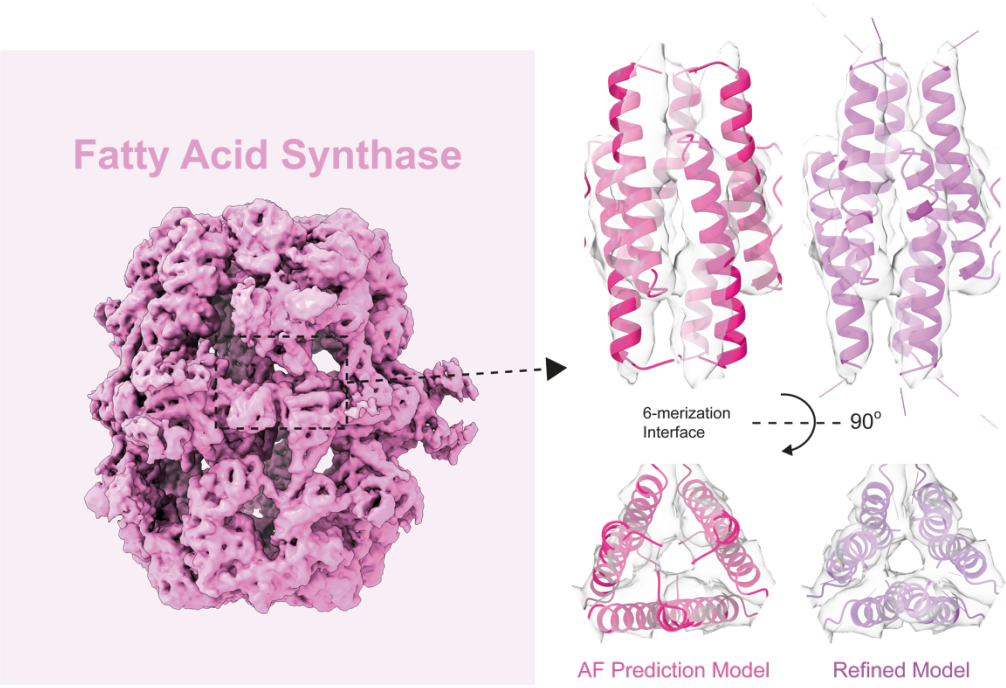


Figure 34: AI-prediction vs. experimental data: FAS.
AlphaFold2 accurately predicts the overall fold of the hexameric sub-complex of AT domains of the *C. thermophilum* FAS, revealing possible interface information, even before refinement in the experimentally obtained map. Figure reproduced from[123].

This observation strengthens the notion that even though the AI software was not trained on multimeric protein datasets[134], the information necessary to recapitulate higher-order protein-protein interactions within a complex still lies in its logic circuits.

As part of the AI-derived modeling of *C. thermophilum* FAS, the PPT acetyl-CoA binding domain (**Figure 26**) was also structurally derived. Interestingly, prior to experimental refinement, AlphaFold2 predicts a domain placement very close to the conformation present in the yeast homolog[200] (**Figure 35**).

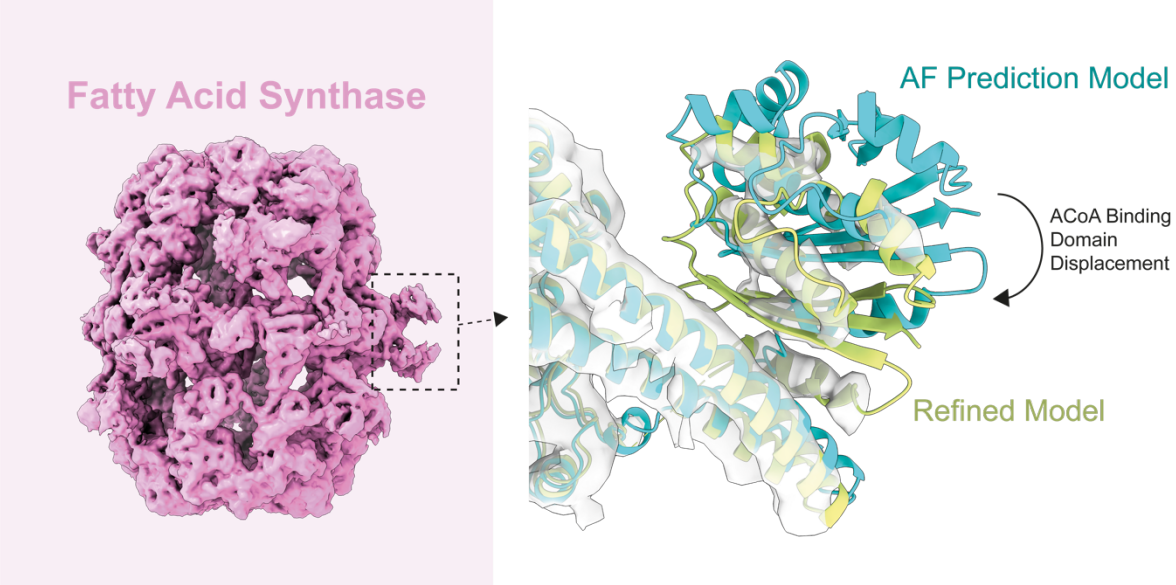


Figure 35: AI-prediction vs. experimental data: FAS (2).
AlphaFold2-predicted model of the acetyl-CoA binding domain of FAS PPT is once again shown to be displaced when compared to the model that was refined in the experimental map. Figure reproduced from[123].

This domain orientation was previously shown in **Figure 26** as not optimal, with cryo-EM data revealing a domain rotational displacement of ~25º in relation to the yeast equivalent. Domain proximity predictions of AlphaFold2 are of low confidence, pointing at the instrumental role that the experimental data plays in the validation and refinement of AI-predicted protein structure models.

3.10 The features revealed by the cryo-EM structural determination of the native, eukaryotic, in-CFS OGDHc core at 3.35 Å

After the successful identification of the CFS-derived OGDHc E2 core component during the cryo-EM exploratory analysis of protein communities, a more intensive cryo-EM data acquisition of the characterized CFS was performed in order to structurally characterize the OGDHc metabolon in detail that confers the succinyl-CoA manufacturing capability of the studied CFS. Overall, ~25,000 cryo-EM movies were acquired, and acquisition was followed by image processing and single-particle analysis (**Appendix Figure 2**). This dataset provided extremely well-resolved 2D class averages from 52,034 particles, showing multiple views of the OGDHc E2o core.

Image processing of the refined particle set (see **Methods**) resulted in a high-resolution map reconstruction – the cryo-EM map determined for the E2o core reached 3.35 Å resolution (**Supplementary Figure 2A, Supplementary Table 4**), allowing for *de novo* model-building and signifies a considerable resolution improvement over the core that was reconstructed after initial signature detection[123] (**Supplementary Figure 15A**).

High-order structural features are observable in the E2o-24mer map (**Figure 36A**), *e.g.,* inter-, and intra- trimeric interfaces (**Figure 36B**), secondary structure elements are captured, and accurate placement of amino-acid side chains was possible (**Supplementary Figure 15B**). After identification of the catalytic active site of the E2o, accurate modeling could also be extended to the side chain conformations of the participating amino-acids.

Modeling results indicate that the localization of the amino-acids participating in the active site, along with the placement of their side chains, is conserved amongst members of the 2-oxo-acid dehydrogenase family of enzymes. The CoA is accommodated by the outward orientation of Phe247, facilitating, through π-stacking, the interaction and stabilization of its 3', 5'-adenosine diphosphate group **(Figure 37)**. The flexibility of the endogenous, bound coenzyme A is underlined by the partly resolved density of the coenzyme A components, i.e., β-alanine, cysteamine and pantoic acid, also denoting its functional implications[207].

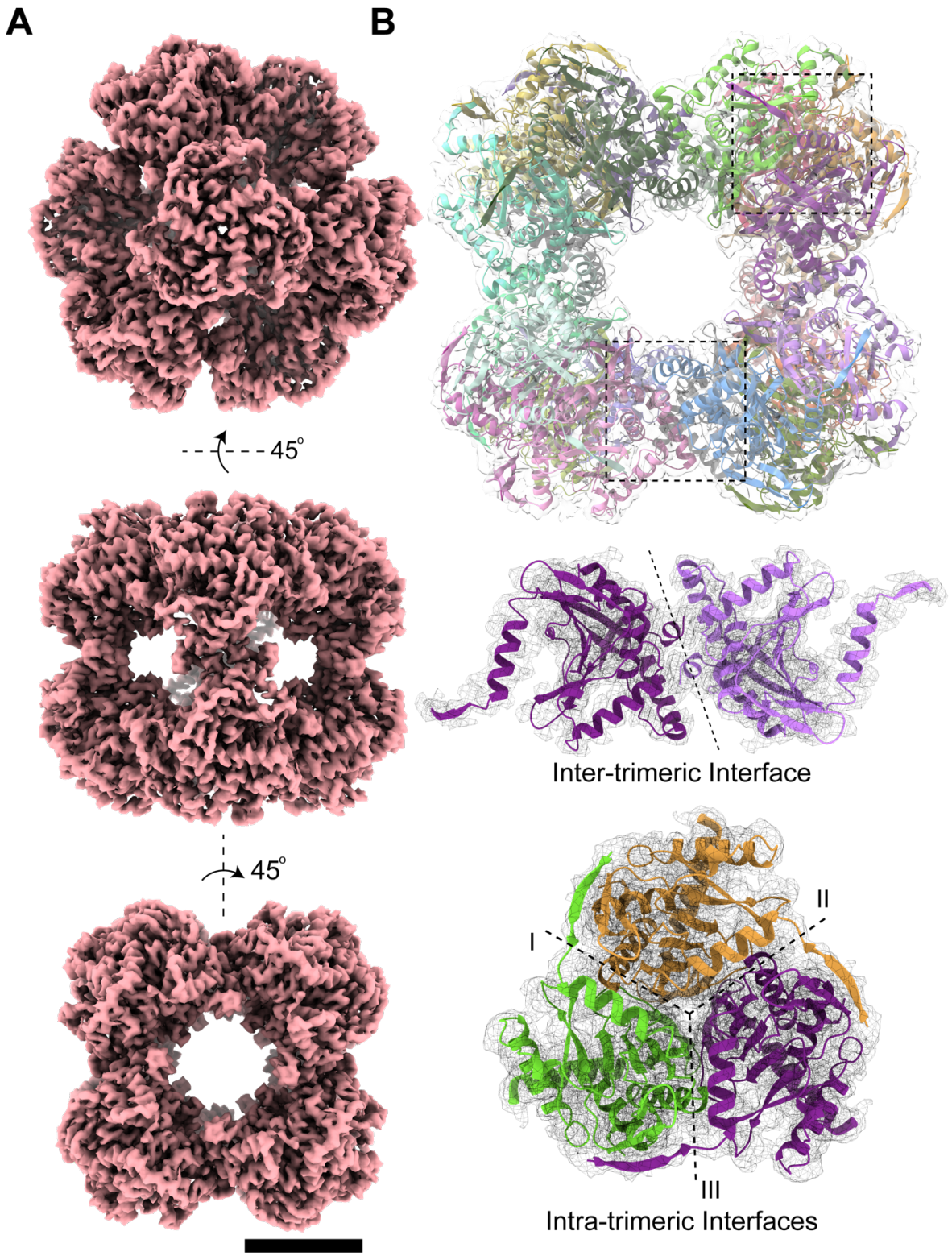


Figure 36: The high-resolution cryo-EM structure of the OGDHc E2o 24-mer core. (A) The 3.35 Å (FSC = 0.143) cryo-EM map of the *C. thermophilum* OGDHc E2o core. Scale bar: 5 nm. (B) The reconstructed atomic model of the *C. thermophilum* E2o core shown in cartoon representation, fitted in the cryo-EM map where inter- and intra-trimeric interfaces are clearly observable. Figure reproduced from[194].

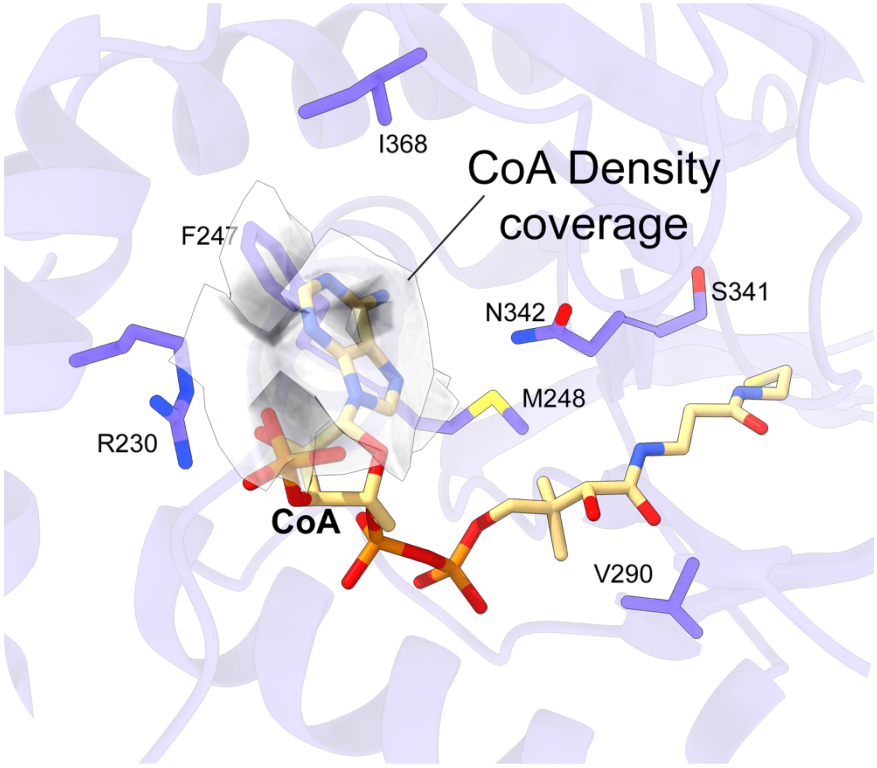


Figure 37: The CoA binding region of the E2o.
Amino acids that participate in the CoA binding region. In the vicinity, a partly resolved density may be attributed to bound CoA. Figure reproduced from[194].

In proximity to the E2o lipoyl- binding site, a partially resolved density could be detected (**Figure 38**); in the case of the bacterial pyruvate dehydrogenase complex, whose core shares the same cubic architecture – a common arrangement across kingdoms for keto-acid complexes – a similar density was attributed to the E2 lipoyl-binding domain (LD)[208]. In another work, an LD bound to the prokaryotic PDHc E2 core active site was refined, presenting a similar conformation[209]. This density also appears, although in lower resolution, in the native, active high-resolution OGDHc core that is presented in this thesis. Based on this density, the LD could be modeled and accommodated (**Figure 38**). When comparing this model to the previously published bacterial PDHc in respect to LD placement and modeling, a major difference is immediately apparent. In the case of the bacterial PDHc, a model was proposed for its resting state where all LDs were bound to the E2 core lipoyl- binding sites. In contrast, the model presented in this thesis suggests an alternative resting state for

the complex, in which the LDs are interacting with the core in a sub-stoichiometric and transient fashion.

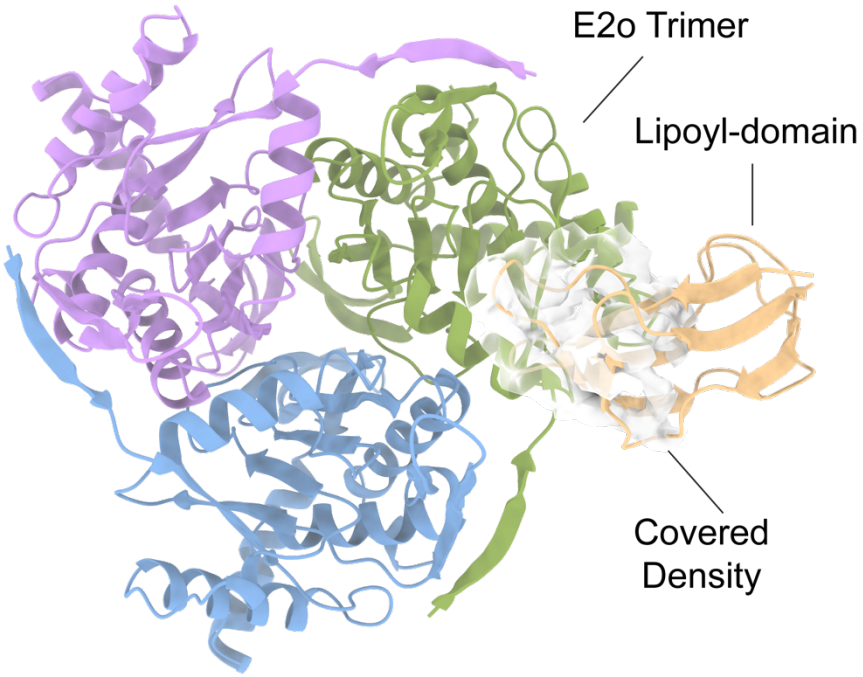


Figure 38: An LD in the proximity of the E2o core.
A partially resolved density in proximity to the E2o vertex trimer could possibly belong to a bound E2o lipoyl- domain. Figure reproduced from[194].

To conclude, the proposed OGDHc E2o core reconstruction constitutes the highest resolution model to date for a protein community member that has been characterized in the context of a native, active cell extract. A wide list of high-resolution features can be observed, such as intra- and inter- trimeric interfaces, as well as active site residue side chains. In addition, its endogenous state reveals lower resolution densities for the flexible, bound CoA and, in contrast to previously published results, indicates a sub-stoichiometric binding of the LD. In combination, the above-mentioned results represent an accurate recapitulation of the previously structurally unresolved endogenous OGDHc core.

3.11 The native, metabolon-embedded dihydrolipoyl succinyltransferase E2o core displays a higher degree of compaction

The overall fold of the native OGDHc E2o core reconstruction, when compared to its overexpressed and inactive human counterpart, is highly similar but, upon closer inspection, displays distinct adaptations. In previously published structures, the *N-ter* region of the E2o adopts a conformation similar to an extended loop and does not display specific folding (**Figure 39**).

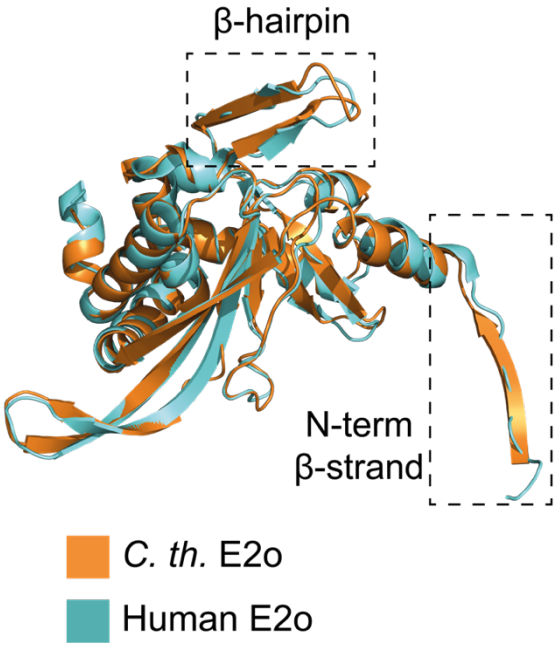


Figure 39: Comparison of a mesophilic and thermophilic E2o monomer.
One of the *C. thermophilum* E2o core monomers (orange) is aligned with its human counterpart (blue). Despite the overall fold similarity, specific differences are observable: a tighter turn conformation allowing better alignment of the *C. thermophilum* β-hairpin model and the β-strand conformation of the *C. thermophilum N-ter*. Figure reproduced from[194].

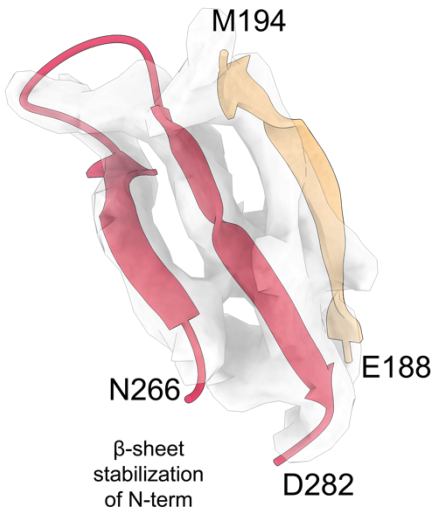


Figure 40: A novel secondary structural element of OGDHc E2 core.
The β-hairpin motif of N266-D282, along with the *N-ter* β-strand E188-M194 form a β-sheet that contributes to the core's overall stability. Figure reproduced from[194].

This is not the case for the *C. thermophilum* element, as it is observed acquiring a β-strand conformation (E188-M194) and, combined with the β-hairpin that is comprised of residues N266-D282, participates in the formation of an extended β-sheet (**Figure 40**). This extended conformational variation is possibly critical for inter-trimeric subunit association and may contribute to the increased stability of the complete 24-meric E2o cubic core through its extensive network of hydrogen bonds (**Figure 53B**). Additionally, upon comparison with the published human counterpart[210], this fold may be the main reason for its increased compaction, also visible by the lower displayed centroid distances amongst the subunits of the E2o core trimers (**Figure 41, Supplementary Figure 6**). This effect can also be quantified in energetics terms, by comparative refinements of both sub-complexes with the HADDOCK macromolecular docking platform. HADDOCK scoring of both human[210] and C*. thermophilum* inter- and intra- trimeric interfaces (**Supplementary Figure 16A, B, Methods**) results in consistent, markedly higher scores for the *C. thermophilum* E2o interfaces, a fact that could, along with the novel locally observed fold, be attributed to the sample's native and thermophilic nature. In principle, a complex's buried surface area usually correlates to its dissociation constant (K_D), given that it is not often subjected to large conformational changes[181], and the results of this analysis suggest that *C. thermophilum* E2o enzymes bind more strongly. In agreement, intra- and inter- subunit interfaces of *C. thermophilum* are approximately 1.5 and 1.7 times larger than the

human E2o core equivalent (**Supplementary Figure 16A, B**), including substantial charged interactions (**Supplementary Figure 16A, B**).

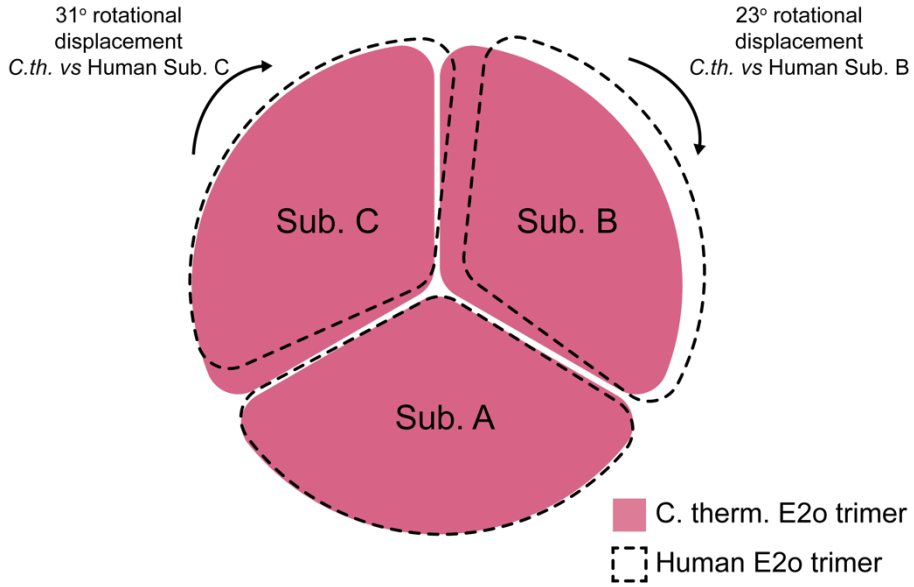


Figure 41: The thermophilic E2o core is more compact when compared to a mesophilic counterpart.
Schematic representation of the human E2o vertex trimer rotational displacement in comparison to the experimentally resolved *C. thermophilum* E2o trimer, displaying a "loose" conformation for the mesophilic counterpart. Figure reproduced from[194].

The core compaction discussed above has an additional effect: it can induce the confinement of the LD located at the E2o *N-ter*. Antiparallel β-sheets generally present higher stability in comparison to the loop conformation that was resolved previously[211], as they can withstand both exposure to different solvents and distortions (β-bulges or higher torsion forces); to this effect, the structural element that was newly identified can act as an anchor point for reducing the conformational space available to the flexible region that connects the two highly structured E2o domains, *i.e.,* the intermediate-carrying LD and the E2o core domain with succinyltransferase activity.

3.12 The OGDHc reaction interfaces are governed by comparable electrostatic complementarity

In the catalytically active OGDHc metabolon (**Figure 16**, **Figure 17**), the succinyl- intermediate is shuttled by the flexible E2o arm that includes the LD from the active site of E1o to the one of E2o and is then prepared for the next cycle through re-oxidation by the E3[212]; therefore, this transfer requires the unhindered access of the LD's lipoylated lysine to each of the complex's different active sites. To elucidate the structure of these transient reaction interfaces, *i.e.,* E1o-LD, E2o-LD, E3-LD, AlphaFold2-Multimer[134] was applied, resulting in the generation of the three metabolon component interfaces (**Figure 42A-C, Methods**)**.** Quality metrics for all three complex interfaces reveal their high quality and confident ranking in regard to experimental error (**Supplementary Figure 3A, B**). As additional validation of the predicted model's quality, (a) the AI-derived predicted positional fit of the LD to the E2o core (**Figure 42B**) highly resembles the one that was observed in the cryo-EM-derived E2o core density revealed above (**Figure 38**), and (b) lipoylated lysine distances of the cofactor to active sites included in the generated interfaces ranges from 10 to 25 Å (**Supplementary Figure 17**), with this range corresponding to the length of the lysine side chain and the lipoate. This observation highlights the intrinsic flexibility of the interacting proteins[155], also highlighted in a recently-published bacterial PDHc E2o-LD interface[209].

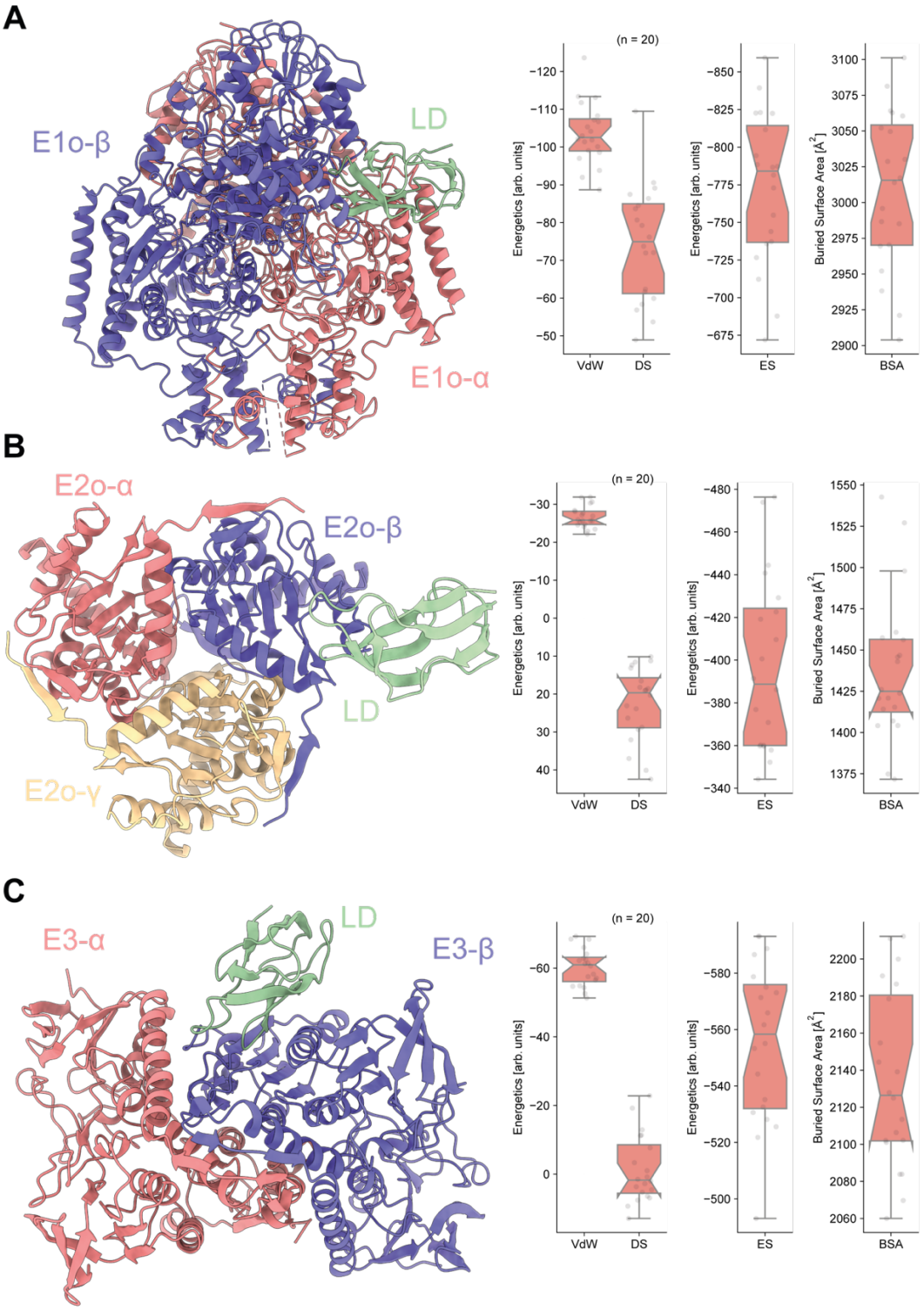


Figure 42: AI models and energetics of the OGDHc components.
(A) AI-derived model of the E1o-LD interaction. Overall model, energetics (vdW = van der Waals interactions, DS = Desolvation energy, ES = Electrostatics) and buried surface area (BSA) can be seen in left and right respectively. (B) AI-derived model of the E2o-LD interaction. Overall model, energetics and buried surface area can be seen in left and right respectively. (C) AI-derived model of the E3-LD interaction. Overall model, energetics and buried surface area can be seen in left and right respectively. (D) AI-derived model of the ordered LD domain of E2o. Lys42 carries the lipoyl-moiety. A highly negatively charged interaction interface is observable. Figure reproduced from[194].

E1o assembles into a homodimer (**Figure 42A**), and can bind simultaneously two thiamine diphosphates (ThDP), and possibly two LDs simultaneously, with their binding site formed by the E1o dimerization interface (**Figure 42A**). In the process of α-ketoglutarate decarboxylation, the lipoylated lysine must be in a biochemically feasible distance to the ThDP C2 carbon in its binding site, a distance that is recapitulated in the AI-generated model (*d* = 14 Å) (**Supplementary Figure 17A**). Additionally, a single LD is bound per E2o monomer (**Figure 42B**) and again, the lipoyl-lysine is positioned close to the CoA that is bound to its respective binding site, with a biochemically feasible distance between the lysine Cα and the CoA thiol group of *d* = 10 Å (**Supplementary Figure 17B**). Finally, a conserved disulfide bridge, along with a histidine, is located in the E3 active site, where the LD re-oxidation reaction takes place[213-215]. The E3-LD complex model places the lipoylated lysine of the LD in a distance of *d* = 23 Å that is accessible for disulfide bonding, proximal to the E3 active site (**Supplementary Figure 17C, D**). It is of interest that, during model prediction, for both E1o and E3 in complex with LD, models containing alternate conformations were generated (**Supplementary Figure 17C, D**) but the previously mentioned biochemical constraints were not satisfied, violating the required distances that are necessary for the active metabolon (**Figure 17**).

After biochemical (in the cases of E1o-LD, E2o-LD and E3-LD) and cryo-EM (in the case of E2o-LD) validation of the generated metabolon subcomplexes, the models were subjected to energetic refinement. After the refinements, a common characteristic was revealed to be shared between the metabolon-embedded interfaces of the AI-generated subcomplex models (**Figure 42A-C**): Strong, comparable electrostatic interactions appear to govern all the validated interfaces (**Figure 42A-C**), a signature characteristic of interfaces that are of transient nature and display high on/off rates[216]. Metabolons that rely on reaction intermediate shuttling via the “swinging arm” mechanism[46], as well as other metabolic processes that require rapid on/off rates[217,218], display complementary electrostatics (**Figure 43, Supplementary Figure 18**) in their interfaces. After calculation of electrostatic potential maps, an extensive, positively charged region is highlighted for each of the LD-binding interfaces of the OGDHc metabolon (**Figure 43**). The function of these surfaces lies in the attraction and accommodation of the corresponding negatively charged surface of the LD and additionally act as another measure of structural

compaction, a fact that is also observable in other oxo acid dehydrogenase complexes[219], which also employ "swinging arm" mechanisms for substrate channeling[220].

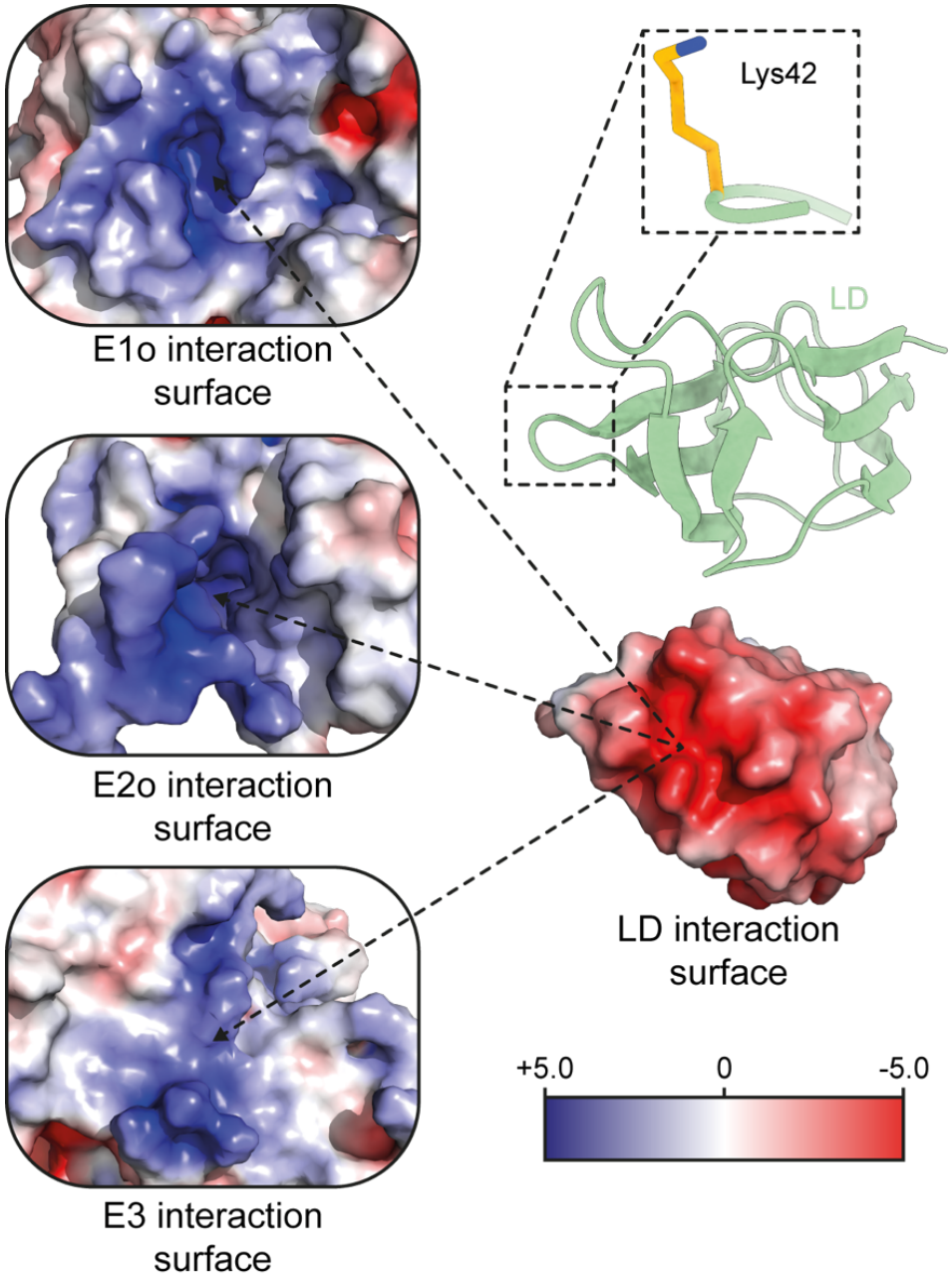


Figure 43: Electrostatic interactions between the LD and the main OGDHc components.
On the left, the interaction interface of E1o (top), E2o (middle) and E3 (bottom) with the LD (right) can be observed, along with the LD AI-predicted atomic model, where Lys42 carries the lipoyl-moiety. A highly negatively charged interaction interface is observable. Figure reproduced from[194].

3.13 Mapping of the CFS-embedded protein communities reveals the E3 binding protein of the OGDHc

In-extract mass spectrometry and crosslinking mass spectrometry experiments were performed (**Methods**) in order to characterize the complete proteome and interactome of the CFS. To this purpose, the optimal crosslinker concentrations were also benchmarked (**Supplementary Figure 1C, D**). The experiments retrieved a total of 4,949 residue-residue crosslinks (3,632 intra- and 1,317 inter- residue crosslinks), and 99.4% mapped to *C. thermophilum* polypeptide chains (**Supplementary Table 2**) with an FDR of 2% recovery level (**Methods**). Across two biological and two technical replicates (**Supplementary Table 2**), 505 out of 2,091 total polypeptide chains were crosslinked, showing that a 29.1% of the retrieved *C. thermophilum* proteome appears to be organized in protein communities[221], a considerably larger percentage that what was previously reported[72].

During analysis of the crosslink-retrieved protein interaction network, 54 different cellular communities could be identified, stemming from multiple cellular compartments (**Supplementary Table 2**), including 1,488 distinct participating members. Across communities, significant subunit coverage can be observed (67% +/- 24%), including examples such as the complete mitochondrial (N = 74, 99% of subunits) as well as cytoplasmic (N = 128, 96% of subunits) translation machinery.

Additionally, the complete PDHc/TCA cycle with all of its component subunits could be identified in the CFS (N = 25, 100% of subunits, **Supplementary Table 2**). The crosslink-retrieved interaction network revealed interactions both within and across (122 intra- and 169 inter-links, **Supplementary Table 2**) the eleven enzymes that participate in the PDHc/TCA cycle.

Interestingly, one of the most interconnected proteins that were identified during network analysis of the resulting crosslinks appears to be the E3 protein (N = 88 inter-molecular crosslinks). E3 was found to be in proximity to seven different community members (**Figure 44**), highlighting its critical participation in the mitochondrial metabolism. Apart from its interactions within the PDHc (with subunits E1p, E2p, E3BP) and within the OGDHc (with subunits E1o and E2o) metabolons, it was observed to interact with a putative holocytochrome c synthase (HCCS) (**Figure 44**)

that could be part of the hypothesized heme metabolon[222] found in the mitochondrial matrix.

However, sequence alignments revealed that there is a fused, misannotated protein sequence at the *N-ter* of HCCS (**Supplementary Figure 7**) that can be attributed to the eukaryotic KGD4 subunit of OGDHc[61], whose function is to tether the E3 to the E2o core of the metabolon (residues 1-130). The *N-ter* sequence presents high similarity and confidently aligns to the eukaryotic KGD4 (**Supplementary Figure 7**). KGD4 showed significant co-elution along with the other subunits that comprise the OGDHc metabolon, after analysis of previously published MS-derived data[72] in three biological replicates and across all involved fractions (**Supplementary Table 2**). As this co-eluting protein performs essentially the same role as the corresponding E3BP subunit of PDHc, *i.e.,* tethering the E3 to the E2p core, it was designated as the OGDHc E3 binding protein of *C. thermophilum* (E3BPo).

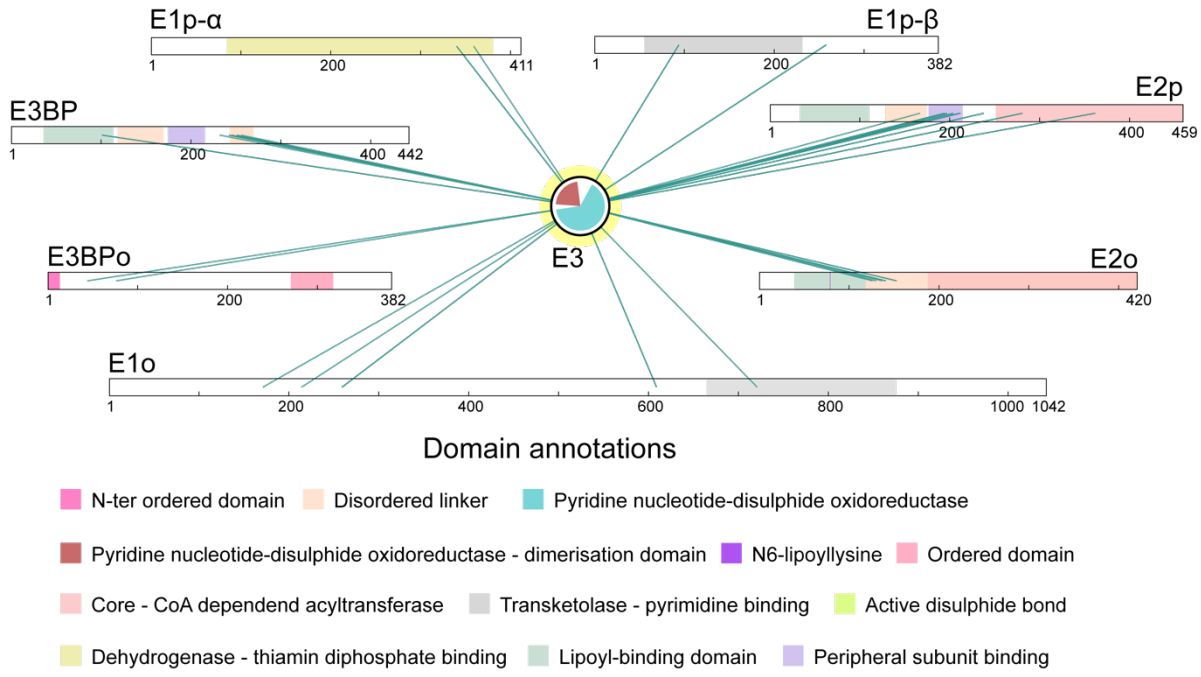


Figure 44: Inter-crosslinks of E3 reveal a plethora of interaction partners in proximity. The E3 protein is found in the vicinity of multiple proteins that belong to different metabolons (PDHc, OGDHc), critical to mitochondrial metabolic pathways. Figure reproduced from[194].

3.14 The OGDHc metabolon interactions are elucidated by a crosslinking-derived network

Amongst the component proteins that make up the OGDHc metabolon (E1o, E2o, E3, E3BPo) 81 intra- and 44 inter-molecular unique crosslinks could be identified (**Figure 45**).As there is a number of inter-crosslinks (**Figure 45**) between subunits that are not known to physically interact and form direct interfaces (*i.e.,* E1o/E3 and E1o/E3BPo), OGDHc presents a high level of compaction in the cell-free system.

Based on the relatively close distance of the E1o and E3 proteins to the E2o core, it is highly possible that the E2o *N-ter* region downstream of the LD (**Figure 45**) may be also involved in peripheral subunit organization.

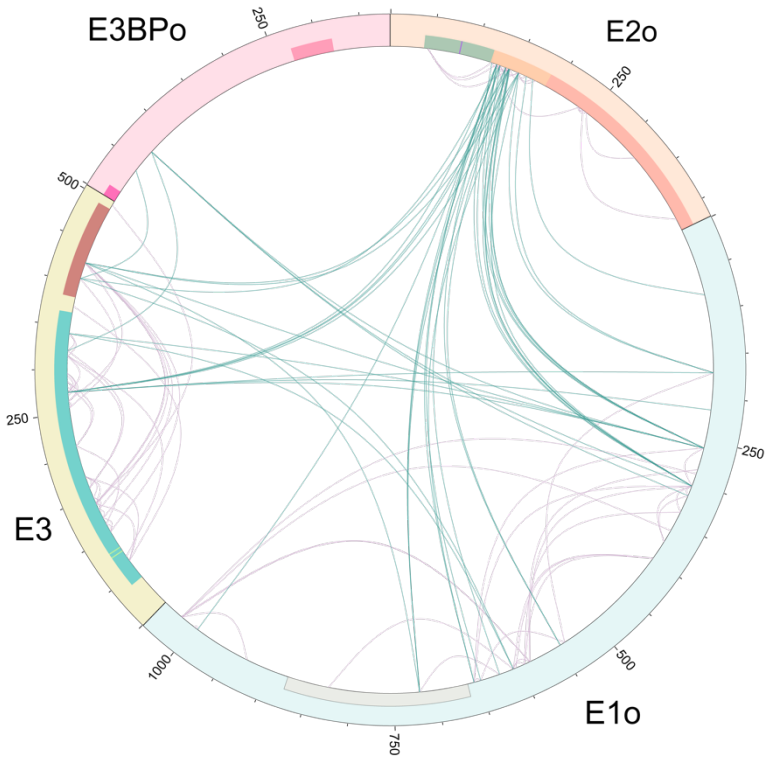
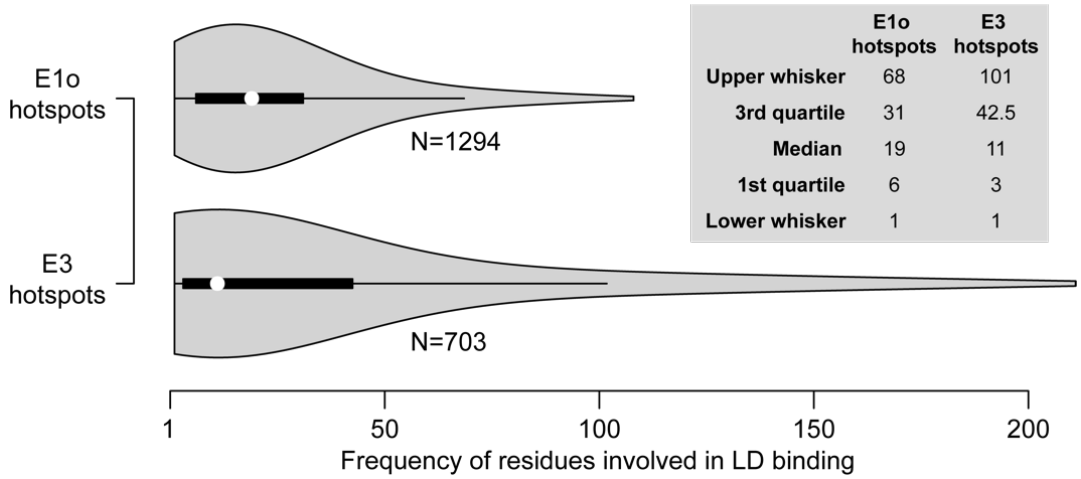


Figure 45: The OGDHc protein components are highly interconnected.
Blue lines represent the inter-crosslinks and purple the intra-crosslinks identified between the subunits that comprise the OGDHc. Figure reproduced from[194].

Driven by this observation, crosslinking-informed flexible molecular docking was employed on the previously validated E1o-LD and E3-LD interaction models, in

order to obtain a better understanding of the interaction interfaces that the LD can sample during reaction intermediate transfer (**Figure 46, Figure 47, Figure 48**).

An adapted docking protocol that could incorporate crosslinking information was devised, where crosslinks could be mapped to the disordered *N-ter* regions of the E2o that are in proximity to the ordered domains participating in the OGDHc protein-protein interactions. The informed docking results were used to obtain the frequencies by which E1o and E3 residues are involved in the binding of the LD (**Methods,**[181]).



	E1o hotspots	E3 hotspots
Upper whisker	68	101
3rd quartile	31	42.5
Median	19	11
1st quartile	6	3
Lower whisker	1	1

Figure 46: Frequency plots of residues involved in the binding interface between the LD and E1o and the LD and E3.
E1o residue numbers that participate in the LD binding are higher, with lower overall individual frequencies in comparison to the corresponding E3-LD interacting residues. Figure reproduced from[194].

The HADDOCK-refined in explicit-solvent molecular models (N = 400) were used to map the top-25% of residues that were observed to be involved in the binding of the LD (**Figure 46**). The results showed that there is a highly distinctive attraction surface at each respective protein (E1o, E3) that may assist in the “guiding” and binding to each respective active site (**Figure 47, Figure 48**).

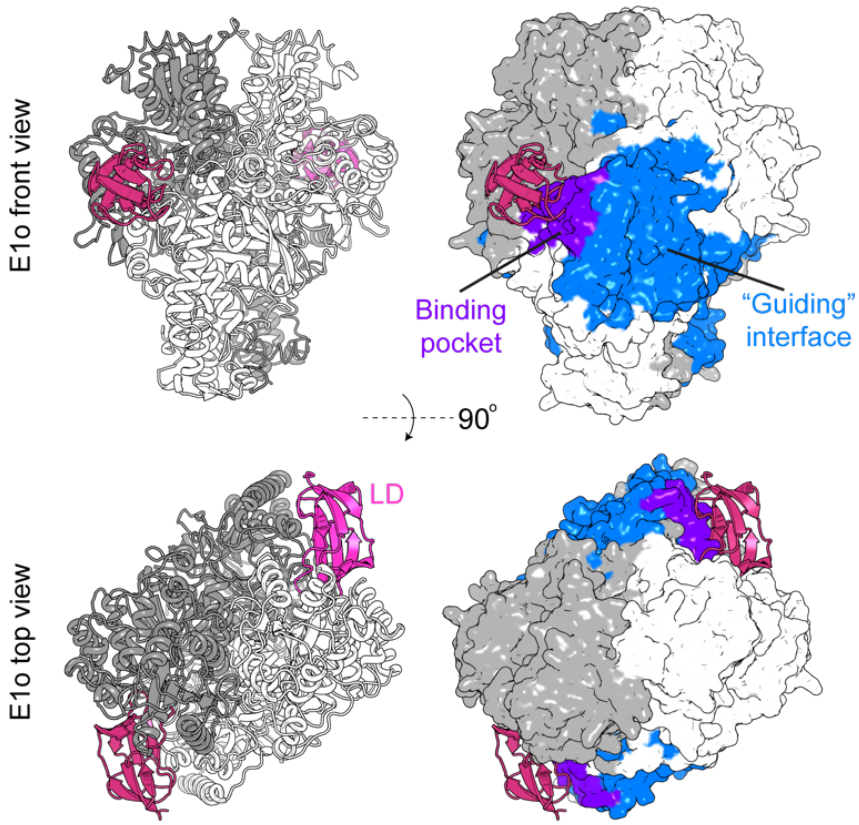


Figure 47: An E1o-LD "guiding" interface.
Residue frequencies of each residue involved in the interaction between the LD (pink) and the E1o (gray) reveal a "guiding" interface (blue) between the two that orients the LD towards its binding pocket (purple). Figure reproduced from[194].

In both the case of the E1o-LD (**Figure 47**) and E3-LD (**Figure 48**) interactions, a residue of either E1o or E3 is considered a "guiding" interface residue when it comes frequent contact with an LD residue. It is quite interesting that in the case of E1o, its attracting surface appears to be more diffused in comparison to the E3 calculated surface (**Figure 47, Figure 48**) in terms of recovered residue hot-spots.

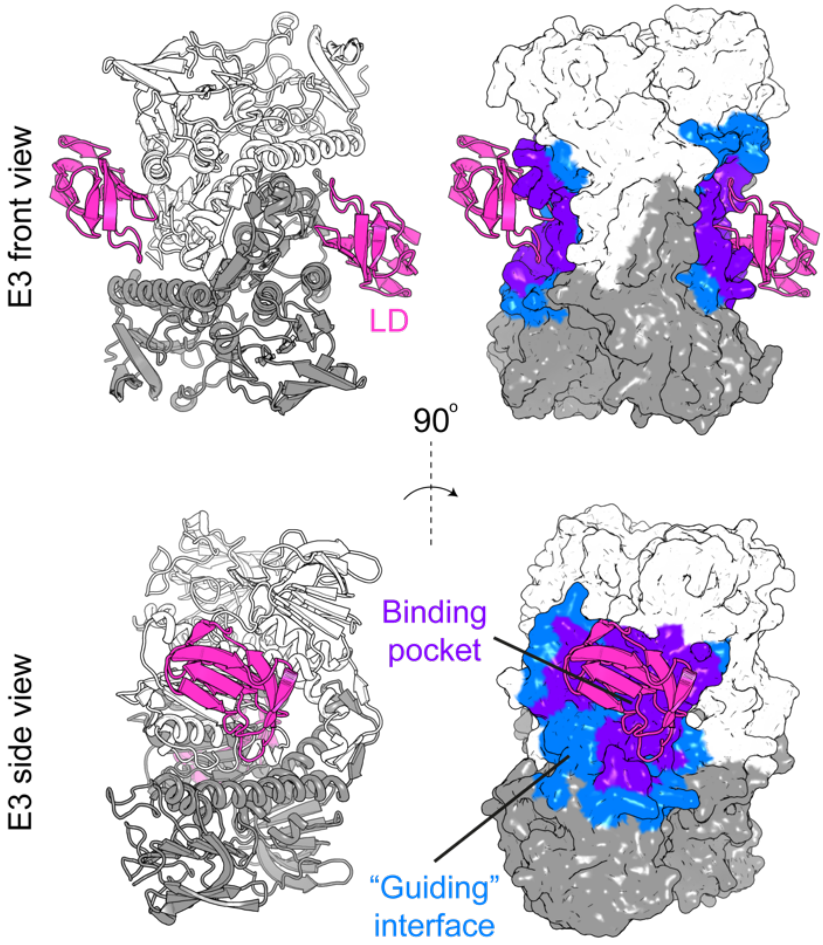


Figure 48: An E3-LD "guiding" interface.
Residue frequencies of each residue involved in the interaction between the LD (pink) and the E3 (gray) reveal a "guiding" interface (blue) between the two that orients the LD towards its binding pocket (purple). Figure reproduced from[194].

For the LD binding to the E1o, the signal is more highly diffused, suggesting that the resulting surface defined by the LD-E1o residue interactions may indeed act as a "guiding" surface, to direct the LD to the quite deeply buried E1o binding pockets (**Figure 47**). In contrast, the corresponding resulting "guiding" surface of the E3 is more limited (**Figure 48**), and appears to be only ~5 Å extended from the more easily accessible E3 active site. In combination, these observations reveal a discrete mechanism for the attraction of the LD to the E1o and E3 active sites respectively, with the flexible regions downstream of the LD further directing the overall interaction.

The intra-crosslinks that were identified serve another purpose apart from validating the presence of the OGDHc subunits: they confirm that these proteins form multimers. Nevertheless, this is not the case for the E3BPo, as no intra-crosslinks

were identified for it, strengthening the previously published hypothesis that it exists only in a monomeric form, tethering one E3 per single copy to the E2o core[61], via its highly flexible *N-ter* region (**Figure 45**). Following previously performed work for the PDHc[79], the iBAQ scores that resulted from the MS experiments were translated into qualitative stoichiometries for the OGDHc's participating proteins. Calculations revealed the presence of 24 E2o subunits (a fact that was also validated by *de novo* model building in the high-res, cryo-EM E2o core) and approximations for the stoichiometries of the rest of the components totalled to <10 E1o dimers, and ~4 E3BPo which tether 4 E3 dimers (**Figure 49A, B**). It is a possibility that more E3s could be accommodated, as according to the identified inter-crosslinks, the E3 often lies in proximity to the disordered regions of the E2o *N-ter*, downstream of the LD. However, theoretically, the E2o core may be able to recruit up to 48 different molecules, leading to an upper limit of 96 different polypeptides (broken down to 24 E2o and E3BPo monomers, each able to bind 24 E1o and E3 dimers respectively). In contrast to the theoretical maximum stoichiometries, the relative stoichiometries that were calculated reveal the substoichiometric binding of peripheral OGDHc subunits. This can be interpreted in two, possibly co-existing ways: (a) that overall availability of E1o and especially E3 is under tight regulatory control and (b) a sub-population of flexible *N-ter* E2o lipoyl- arms always remains unbound to cycle reaction intermediates.

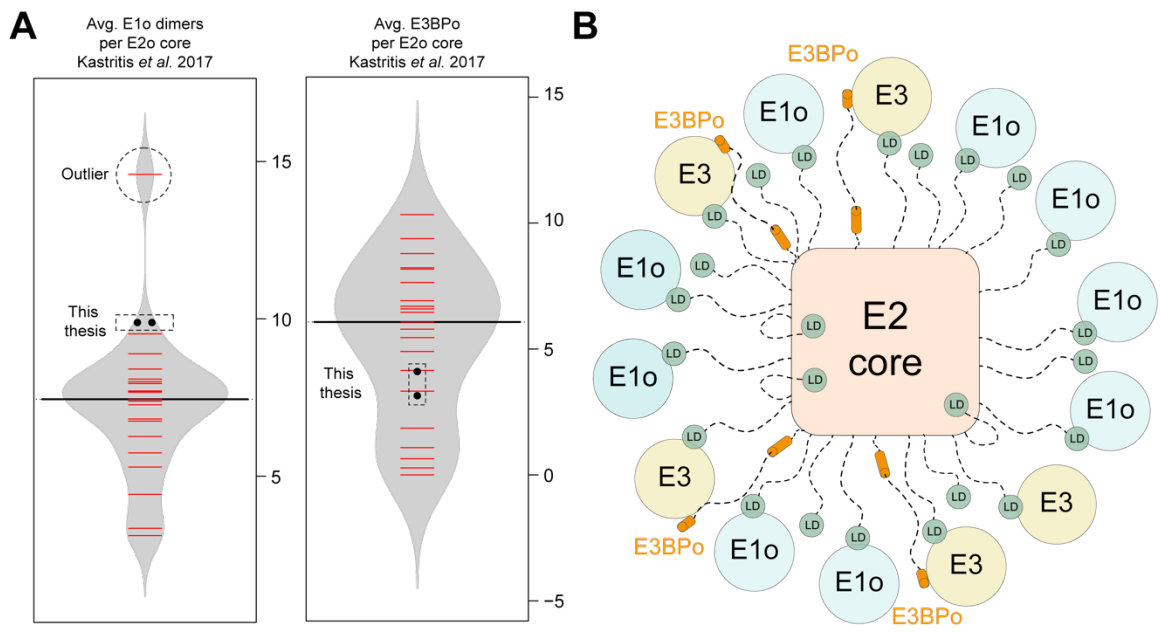


Figure 49: MS stoichiometric calculations of OGDHc protein components.
(A) Stoichiometric calculations of previously published MS data[72] and newly derived MS data reveal the stoichiometry of the E1o and E3BPo components that take part in the formation of the fully active, native OGDHc. (B) Schematic representation based on the stoichiometries derived from the MS data presented in this article. The 24-mer E2o core is surrounded by at least 4 E3 dimers that are bound to 4 E3BPo, joined by 9 E1o dimers. Figure reproduced from[194].

3.15 The proximity of E1o, E2o and E3 proteins in the context of an active OGDHc metabolon is revealed by cryo-EM and computational analysis

In order to gain further insights into the higher-order, CFS-embedded architecture of the OGDHc metabolon, single particle data was used to perform asymmetric reconstructions. Initial particle classifications led to 2D projections that display visible external densities, along with a strong signal in the center (**Figure 50**).

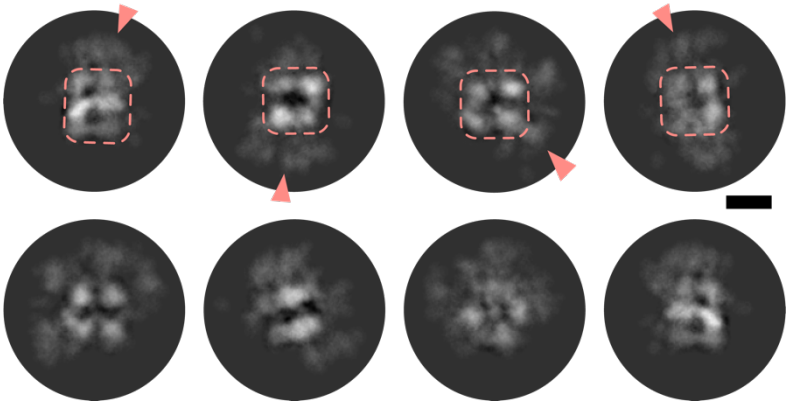


Figure 50: Distinct signal zones of OGDHc 2D classes.
2D class averages of the OGDHc present a core of high signal (designated by the dotted line in the center), surrounded by more diffused but still well-defined signal on the periphery (shown with arrowhead). Scale bar: 5 nm. Figure reproduced from[194].

These external densities should represent the locations of the, peripheral to the core, E1o and E3 dimers. Each E1o and E3 dimer remains tethered to the E2o core via the flexible *N-ter* region of the E2o (**Figure 45**) and the highly flexible E3BPo (**Figure 45**, [61]), respectively.

The asymmetric 3D reconstructions lead to a map that was determined at a resolution of 21 Å (FSC = 0.143) (**Supplementary Figure 2B, Supplementary Table 4**). This map exhibited clustered densities in the periphery of the E2o, while, with increasing density threshold visualization weaker, more diffused densities could additionally be observed (**Figure 51**). The AlphaFold2-derived models of the E1o-LD and E3-LD sub-complexes of *C. thermophilum* were systematically fitted in those densities (**Figure 51, Supplementary Figure 4A, B**). In both cases, the models fit in peripheral densities with higher cross-correlation scores when compared to the core densities (**Supplementary Table 7**), enabling their confident placement in the periphery and allowing for the additional fitting of the previously *de novo* resolved E2o core in the central, core densities. In total, two distinct E1o and three E3 dimers could be confidently fitted in the recovered strong external densities of the asymmetric reconstruction (**Supplementary Table 7, Figure 51**).

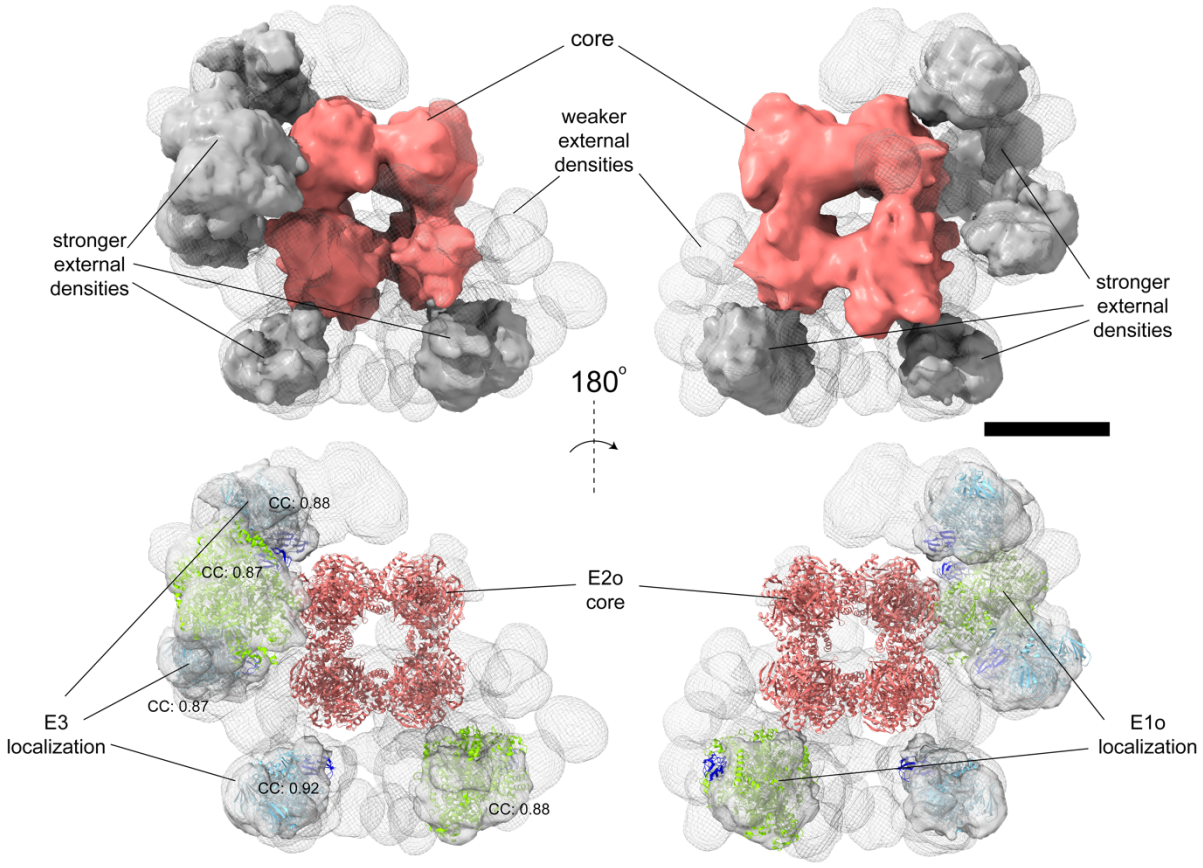


Figure 51: Localizing all OGDHc components in an asymmetric reconstruction. The asymmetric 3D reconstruction of the OGDHc displays a strong density corresponding to the E2o core, and a mix of weaker and stronger external densities. In the stronger densities, E1o and E3 dimers can be localized with high confidence. Scale bar: 10 nm. Figure reproduced from[194].

The E1o and E3 enzymes are asymmetrically distributed in the periphery of the E2o core and remain tethered in relatively equal spatial distance to the core itself (**Figure 51**). While the linker region of the E2o is inherently flexible, its apparent spanned distance comes in contrast to theoretical calculations based on physical-chemical properties of other categories of flexible protein regions (*e.g.,* IDP, pre-molten globule, unfolded; **Supplementary Figure 5A**). Based on direct measurements of the E2o linker length after systematic fitting of all models in the asymmetric 3D reconstruction (**Figure 51**), the flexible regions of the OGDHc display properties that do not coincide either with the theoretical calculations for an "extended" IDP (as defined by Marsh and Forman-Kay[185]) or with "restricted linkers (as defined by George and Heringa[187]) (**Figure 52, Supplementary Figure 5A**).

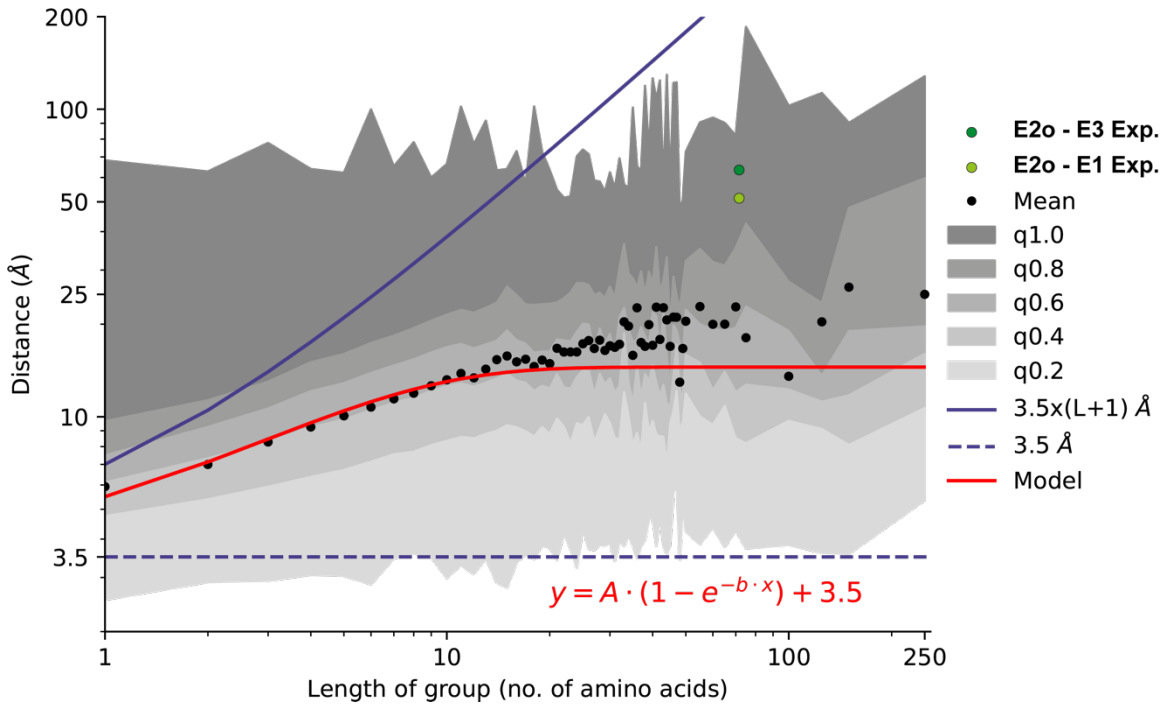


Figure 52: Flexible linker distance analysis.
Graph representing the mean distance values of the binned groups of all unresolved amino-acid sequences belonging to all protein structures deposited in the PDB. Black dots represent the mean value of a length of amino-acids group, with varying scales of gray the standard deviation after the integration of each quartile of total data, as denoted in the plot legend. The red line represents a fitted model that describes the relationship between the distance and the length of amino-acid groups. With light green the experimentally measured average distance between the N-ter of the resolved core domain of the E2o and the C-ter of the LD that is bound to a fitted E1o dimer on the periphery, while dark green represents the same average distance experimentally measured between the N-ter of the resolved core domain of the E2o and the C-ter of the LD that is bound to a fitted E3 dimer. The blue line represents the theoretical maximum distance of an amino-acid sequence when completely stretched, while the dashed blue line represents the theoretical lower limit of any amino-acid sequence. Figure reproduced from[194].

To better understand the significance of the measured distances that the flexible linkers span between the ordered domains of the enzymes that participate in the formation of the complete OGDHc metabolon, a systematic search was performed in all available structural model data. With this data, it was possible to calculate the distance between protein residues that are structurally resolved, but sequence-wise are distant due to a non-resolved residue span in-between. While this non-resolved stretch may be attributed to technical reasons, it is often considered as evidence for protein disorder[223]. An analysis of the complete structural data that has been deposited in the Protein Data Bank (PDB)[224] as of June 2022 (191,144 available structures)

showed that distances between the structurally resolved residues that precede and follow a non-resolved stretch appear to be significantly confined (**Figure 52**). The analysis made apparent that data on interactions or flexible linkers such as the ones reported in the previous sections are rarely included in the so-far deposited structural data in the PDB (**Figure 51, Figure 52**). Concerning the PDB-recapitulated non-ordered region distances, a “flattening” of their average distance is apparent when the non-ordered stretch spans more than 25 protein residues (**Figure 52**). Maximum Cα-Cα distances average at a maximum of ~30 Å, a distance that is significantly shorter than the ones observed between the ordered components of the OGDHc metabolon.

3.16 A CFS-derived integrative model for the architecture of OGDHc

In combination, the results of this thesis present a suggested architecture for the complete oxoglutarate dehydrogenase complex, investigated as part of a succinyl-CoA-producing cell-free system (**Figure 53**). In the center of the complex lies a compact cubic core, comprised by 24 E2o subunits. The E2o trimers that make up the cube’s vertices show an increased level of compaction that results from the core domain *N-ter* β-sheet formed from two neighboring E2o subunits. This secondary structure element appears to have an additional function in confining the available conformational space of the flexible *N-ter* linkers of the E2o that tether the LD. Critical for the correct function of the OGDHc, the LD is responsible for the transport of reaction intermediates between the complex’s serial active sites. The distances that the LD has to span range from 50 to 100 Å for the LD-E1o interaction and from 60 to 70 Å for the interaction between LD and E3. Based on the derived XL-MS data, the LD, along with its flexible linker region, could also serve as an additional “anchor” that could assist in maintaining the E1o and E3 subunits proximal to the E2o core, given the absence of a defined peripheral subunit binding domain included in the E2o *N-ter* sequence.

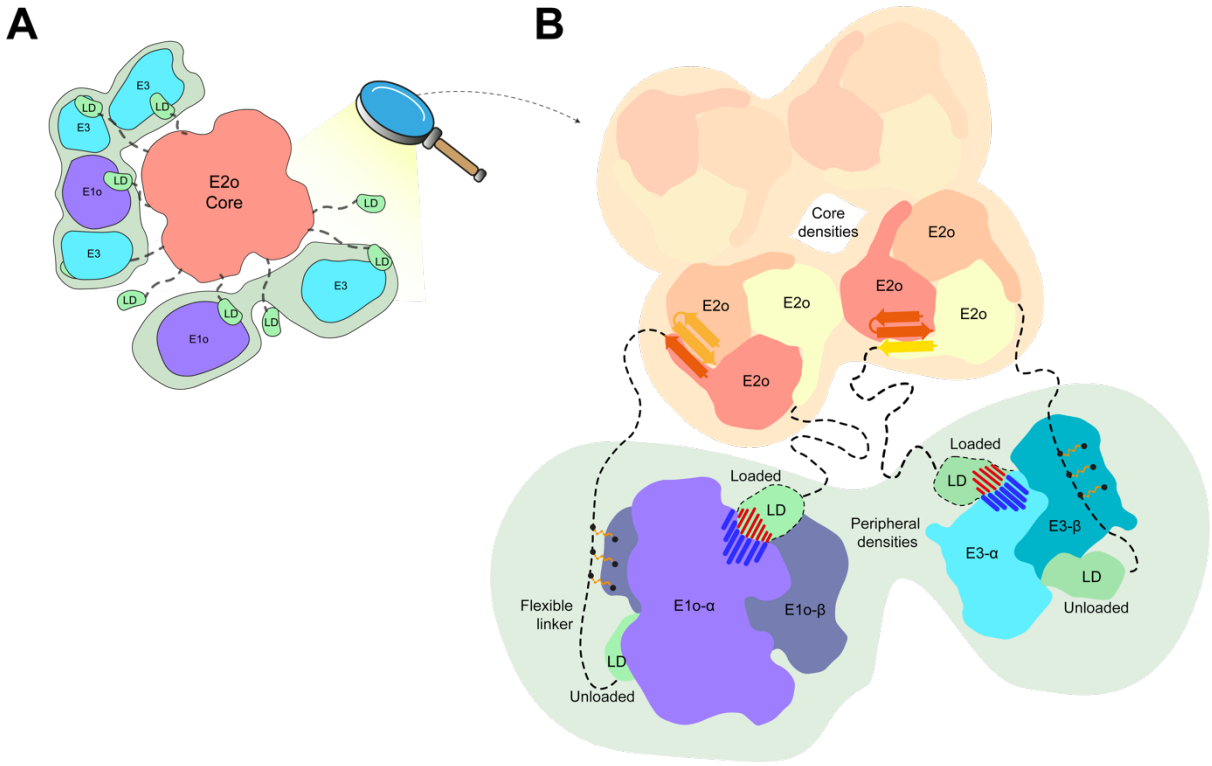


Figure 53: A comprehensive model for the organization of the OGDHc.
(A) Schematic representation of the peripheral subunit organization of the resolved OGDHc. (B) Integrative model describing all the novel observations concerning the organization of the OGDHc. The N-ter β-sheet conformation of the E2o core vertex trimer helps stabilize and compact the 24-mer E2o core, while simultaneously orienting the flexible linker that connects the LD domain. Crosslinking data reveals a fairly stable interaction between the flexible linker and the peripheral subunits, possibly hinting at a structural role of an unloaded LD domain, while another, loaded LD domain performs the reaction cycling. The interaction between the LD and the peripheral subunits is governed by strong electrostatic forces. Figure reproduced from[194].

Recently, the *N-ter* disordered E1o domain was resolved and was observed to interact with the LD-E1o binding site[63]. This interaction could be important for the organization of the E1o in the vicinity of the E2o core, as it is possible that it may interact with other of the metabolon's LD-binding interfaces. As E1o and E3 are homodimeric assemblies, they include two distinct LD binding sites. This could lead to a distinct function for each one, with the first actively participating in the reaction and the second acting as an "anchor point" for an "unloaded" LD that will act as the anchor that maintains the homodimer in the vicinity through complementary, highly-charged electrostatic tethering. Additional, weaker interactions between the E1o/E3 surface residues with the flexible *N-ter* linker region of the E2o and, specifically in the case of the E3, with the highly flexible E3BPo protein may assist in further stabilization of the LD "anchoring" interaction. The substoichiometric relationship between the E1o-E3-

E3BPo peripheral proteins and the E2o core (that was quantified in the MS data presented in this thesis) could further hint at a structural role for the available LDs.

4 Conclusions and outlook

Fractions derived from native cell lysates could act as a, formerly uncharted, framework for identifying, characterizing and optimizing multiple enzymatic reactions, enriching our understanding and applicability of cell-free biotechnology. They also represent a tool by which medically-relevant protein complexes that are involved in critical metabolite production can be explored, as their amenability to direct biochemical and structural investigation techniques is much higher when compared to single-cell *in vivo* approaches, and their "native" context provides unique information that is not accessible during the investigation of overexpressed, purified protein samples.

So far, utilization of cell-free systems was mostly limited to production of a narrow selection of biotechnologically relevant products[115]. Extensive metabolic engineering was employed *a priori* in the specific organism used to produce the CFS, aimed at enhancing the efficiency of the targeted enzymatic pathway[126]. Additionally, the enzymes that are part of the employed metabolic pathways are usually characterized *in vitro*[225], missing critical information about the enzyme's "native" context, both structural and functional.

In most biotechnological applications, cell-free systems are often treated as "black boxes", as the extracts that are employed to produce the desired product, are not characterized or structurally interrogated, missing critical information about their internal structure and function. To gain the necessary insight that will elucidate cell-free system organization, diverse methods are necessary, allowing for better understanding of CFS content, both structurally and chemically. Moreover, it is critical that the "native" principles of organization are retained in a cell lysate, through a single-step fractionation. The preservation of protein-protein interactions and macromolecular assemblies lead to further in-depth understanding of a reaction pathway, both in biochemical and structural terms. This knowledge can then not just

be directly employed in further biotechnological applications, but also be correlated to *in vivo* conditions.

4.1 The importance of biochemical characterization of the CFS

The CFS presented in this thesis constitutes a very complex mixture of protein communities, which requires a rigorous and systematic approach for its characterization, especially when the analysis is focused on the identification, separation and reconstruction of protein community members. This type of analysis pipeline can contain an inherent bias towards easy-to-identify protein assemblies. There are ways to tackle this through specific strategies during image analysis, such as non-template-guided particle picking or "blob picking", but still particles can be missed, as centering in a specific signature that is part of a community will be inaccurate, due to their intricacy. Thus, a system's biochemical characterization acquires even greater importance, such as the work performed in this thesis, concerning the identification and measurement of the pyruvate and α-ketoglutarate reactions, as well as the calculation of the full kinetic parameters and temperature adaptations of the reaction performed by the OGDHc. Kinetic characterization is a prerequisite for the biotechnological application of a CFS but activity assays can also be utilized solely as a discovery- and verification- based approach on complex samples. Kinetic assays can verify that a protein community member under investigation is captured in its active form or even discover new biochemical functionalities in the sample that can then be used as a guide for structural biology-based approaches and supplement novel data analysis schemes that employ machine learning strategies[226].

4.2 Metabolite availability is defined by flexible regions

In this thesis, the main three large enzymatic complexes that were investigated (OGDHc, PDHc and FAS) are the main regulators of the cellular availability of three key metabolites, α-ketoglutarate, acetyl-CoA and palmitic acid respectively. All three

metabolites, aside their critical role in multiple biosynthetic and metabolic pathways, can enact changes and control the flow of multiple cell signaling pathways[227]. They can affect cell survival and proliferation pathways[7], control autophagy[9,228], inflammation[33,229] responses and tumor progression and patient survival[230] can be negatively affected due to their deregulation. It is quite intriguing that their function seems to be not only reliant on the pathway they are associated with, but also on cell type, as the same metabolite concentration change can affect different cell types in different ways, with the mTOR signaling pathway[35,231] being a prime example of this effect, hinting at the existence of another regulatory layer relative to cell type.

Despite the seeming rigidity of the three metabolons, the results that were presented here point to a considerable amount of inherent flexibility, which seems to regulate conformational space and subunit interactions, thus facilitating substrate channeling among subunits, substrate accessibility, regulating protein-protein interactions and controlling overall reaction speed[232]. Structural disorder has been shown to manifest in proteins and protein complexes that require all these characteristics. Disordered regions provide the flexibility and conformational range necessary for large enzymatic complexes to finely tune their activity and regulate the availability of metabolites of critical importance[233].

The OGDHc and PDHc E2 subunits that make up the core of each metabolon contain flexible regions that tether important domains to the core structure. These flexible linker regions display low conservation in regards to sequence length or identity, but seem to contain multiple alanine / serine / proline residues, amino-acids that are linked with structural flexibility and disorder. It can be speculated that, unlike active sites that are defined by a very conserved primary sequence pattern, the overall amino-acid composition of a region is responsible for its flexible properties and function.

The flexible regions of the E2o and E2p have another interesting feature, which is their fairly high (~10%) lysine content. Lysine residues have been shown to be spontaneously modified in the presence of acetyl-CoA[234], and this process can act as a “hidden” regulation layer of their function, as the acetylation of lysine residues could contribute in the decrease of available conformational space that the flexible linkers can explore, with implications on substrate channeling and availability. Since the

process is spontaneous and no other enzymes are involved, it would work based on a concentration-dependent manner, acting as a self-feedback regulatory loop.

4.3 A native-derived CFS provides insights into metabolon features, organization and function

A structural-oriented investigation of large biomolecular assemblies and especially of how they are organized or how flexible regions govern their assembly and function are limited severely by the very characteristics under investigation. This type of structural features has been proven quite elusive to characterize when utilizing samples derived from *in vitro* protein expression and purification. In this regard, native cell lysate fractions provide a previously unexplored framework for the identification, characterization and optimization of enzymatic reactions in the context of cell-free biotechnology. Previously, their utilization has mostly been confined to the production of specific biotechnologically relevant products[115], and only after extensive metabolic engineering of the desired production pathway[126] in order to maximize efficiency in specific organisms. The majority of the characterization of enzymes involved in the employed metabolic pathways is performed *in vitro*[225], lacking information concerning the enzyme's "native" context. It is evident that identification, characterization and structural interrogation of the derived extracts is lacking, hindering understanding of the function within, therefore, utilizing these systems as "black boxes".

Thankfully, recent advances in structural and computational biology, especially in cryo-electron microscopy[72] and computational modeling[106,107] have enabled researchers to probe protein complexes in a native setting[79-81], revealing characteristics that could not be observed with *in vitro* studies of protein assemblies. The integrative results presented in this thesis, provided with an array of methods, prove that cell-free systems, derived from a fractionated native cell extract, can be interrogated with multiple approaches to eventually understand better their content. Importantly, a single-step fractionated cell lysate retains the "native" principles necessary to fully dissect a reaction pathway, both biochemically and structurally, making it possible to directly employ it in biotechnological applications. Moreover, revealing the organization and characteristics of the ordered components of native

protein assemblies can provide insights on how the intrinsic disorder of flexible regions contributes to the regulation of enzyme kinetics, a critical parameter of any CFS targeted towards biotechnological application.

Previous attempts at a cell lysate fraction analysis had shown that recovery of native macromolecular complexes at high-resolution is feasible[235], but they were not targeted towards protein community member structural investigations. Identification and reconstruction of protein community members was limited between resolutions of ~5 to 7 Å[79], despite complementing information from mass-spectrometry and other integrative structural biology approaches. In this thesis, high-resolution structural information for four different structural signatures was retrieved, with all signatures corresponding to protein community members, via the utilization of contemporary cryo-EM technology and AI-based algorithms geared towards protein modeling and image analysis. All four structural signatures were retrieved from a CFS derived from a native cell extract fraction and after refinement, all signatures reached a resolution in the sub-5 Å regime.

4.4 Artificial intelligence aids in the *de novo* modeling of native protein community members

The recent advancements in AI-based model prediction were employed in two different processes during the *de novo* reconstruction of protein community members that were presented in this thesis. Based on mainchain AlphaFold2 identification, and assisted by external database searches, structural signature maps were able to be characterized, without the need for very high-resolution map reconstructions, as the full sequence could be fit in even the hybrid E2o/b core reconstruction (~4.5 Å). Until recently, the majority of *de novo* polypeptide chain modeling was performed in either highly purified or overexpressed endogenous protein species[141,236,237]. The results presented here prove that *de novo* modeling of protein community members at near-atomic resolution is feasible, despite limitations such as low particle abundance, high sample complexity and the inherent flexibility of large biomolecular assemblies.

Another interesting application for the AI-derived structural models of protein community members is to extract native structural adaptation information after refining them in experimental data. In this process, some intriguing observations became

apparent. Firstly, before refinement in the experimental data, the AI models contained some regions that could be interpreted as “over-folded”. This tendency to predict stable protein secondary structure which cannot be observed in the experimental data can hinder an investigator’s ability to derive meaningful insights concerning protein function based solely on the predicted model, as this type of flexibility or structural disorder was shown to be critical for protein function. This could be observed especially in the case of the predicted as backfolded *N-ter* helical region of the E2p, which is absent in the experimental data. Secondly, AI-predicted models tended to misinterpret, or predict as stable, transient domain conformations and interactions. In the case of OGDHc, the LD was predicted as stably bound to the E2o trimer, whereas this interaction must be transient for the reaction to occur, a fact that was also supported by the absence of stable density for this interaction in the experimentally resolved map. In the high-resolution reconstruction of the OGDHc E2 core, partial density for the LD is visible, but due to the symmetry imposed during its reconstruction, it could represent the capturing of a transient state. In the case of the PPT acetyl-CoA binding domain of FAS, its native conformation was not captured in the AI-derived or homology-based models, before experimental refinement. Apart from the linker region difference between the *C. thermophilum* and yeast sequences, which allows the *C. thermophilum* domain to explore a larger conformational space, there can be another exciting explanation for this adaptation. It could be a result of protein community-specific adaptations, facilitating acetyl-CoA channeling from a neighboring community member. If this holds true, this kind of adaptation could never be observed in FAS structure that was derived from a pure protein sample, as is the case of the yeast equivalent published structure.

In general, the results that were derived from AI-produced and experimentally refined models point to the fact that such algorithms are trained on the publicly available structural models, which are in their vast majority derived from protein samples that were overexpressed and highly purified, then crystallized and far removed from their native context. AlphaFold2 accurately predicts individual domain folds but correct recapitulation of domain placement and domain-domain interactions remain elusive. It was observed here that this lack of information can be supplemented by experimental information that stems from a vitrified native cell extract, which can

preserve the protein community architecture and offer insights into community-specific structural adaptations.

4.5 Limitations in the study of a native, cell extract-derived CFS and its biotechnological application

Cell extract fractionation has been proposed by previous studies[85,238] as an investigation method for the understanding of protein community architecture and function, by retaining their "native" character[235]. However, the apparent complexity of such samples requires a multi-technique approach for their elucidation. Contemporary integrative structural biology methodologies are geared towards the analysis of such samples and different structural and bioinformatics approaches can be used in a complementary fashion to obtain meaningful results. In the data obtained in this thesis, protein community members do not appear in a linear assembly of protein complexes, a fact visible in the original micrographs. This arrangement of structural signatures complicates image processing, especially during the initial steps of particle 2D classification. Cryo-tomography could be a solution for this issue by providing *in situ* 3D structural models, but, despite recent advances in throughput[239], remains quite inefficient as compared to single-particle cryo-electron microscopy in terms of throughput. The two techniques should be used in a complementary fashion in order to supplement each other's drawbacks.

Cryo-EM, despite its obvious suitability for the study of native, heterogeneous cell extracts, also displays apparent limitations. During sample preparation, ice thickness of the vitrified sample can already introduce bias in regards to protein community member visibility and analysis. Thin ice, which is sought after during sample preparation that will lead to high-resolution reconstructions, can possibly act as a barrier for the inclusion of very large protein assemblies, as it will introduce preferential particle orientations and even cause denaturation of large biomolecular assemblies due to their contact with the air-water interface. If the goal is the capturing of the complete protein community, and not just a single protein community member, alternative vitrification protocols should be considered that will be optimized for this purpose. These can include lower blotting times that will result in thicker ice, one-sided

grid blotting that can produce different particle distributions on the grid and the development of new nanospray-type sample application methodologies.

As mentioned above, a sample derived from native cell extracts is inherently complex. This means that particle abundance in the images is extremely low, even limited to a single-digit particle number per micrograph and to reach resolutions necessary for *de novo* modeling an extreme amount of data should be collected, as is shown in this thesis for the high-resolution OGDHc E2 core reconstruction. This particle scarcity can also be translated into an inability to capture conformational variability and different structural states. With higher particle abundance, investigation of conformational states would not require *in silico* approaches, as was the case with the FAS PPT acetyl-CoA binding domain, and would be performed directly in the primary experimental data, without any need for symmetry application, by employing techniques such as the 3D variability analysis[104]. For now, particle abundance can most effectively be tackled with rigorous data acquisition and advancements in the field of cryo-EM instrumentation (*e.g.,* 300 keV microscopes with in-column or post-column energy filters, combined with the latest direct electron detectors that can display a ten-fold increase in data throughput[240]) will contribute in resolving this type of limitations.

A critical aspect that should be considered before the biotechnological application of native CFS system is its cost. In this thesis, a native CFS with succinyl-CoA production capability was suggested as a model system in order to ascertain that indeed it can be applied for such a purpose without *a priori* metabolic engineering of the organism used to produce it. Nevertheless, despite the significant results presented here, the product generated in this case does not currently counter-balance the cost of its application. However, this work clearly shows the advantages of understanding how a CFS system works in a multi-scale level and with a variety of experimental and computational methods. This work provides the basis for future applications where a native CFS can be targeted with higher biotechnological value. This CFS could be selected, after careful planning and cost-benefit analysis, *e.g.,* for either small scale production of rare metabolites or as an investigative platform for biochemical and structural characterization of biomacromolecules in native conditions.

Modern cryo-EM sample preparation protocols, such as the ones described above still represent harsh conditions that could adversely affect a sample, possibly

resulting in dissociation and inactivation of protein communities and their members. Choosing the correct model system for sample preparation can counteract these phenomena. In this thesis, the proteins of a eukaryotic, thermophilic organism were investigated. Thermophilic proteins retain their higher-order organization principles through the complete sample preparation protocol application, from lysis, to fractionation and subsequent vitrification, displaying high stability. Furthermore, native, thermophilic adaptations could be explored and may be another explanation, *e.g.,* for the species-specific conformational variations that were observed for the FAS PPT acetyl-CoA binding domain and its flexible linker.

4.6 A new pipeline for the characterization of in-CFS native protein community members

This thesis is an example of the feasibility and inherent potential that is held by the integrative structural and biochemical investigation of native cell extracts, how they can contribute on the elucidation of native protein community and in-extract organizational principles and their suitability as a tool with biotechnological applications. An unobtrusive sample preparation was instrumental in the maintenance of native protein community interactions, a fact that was also supported by the sheer number of recovered protein communities and protein-protein interaction networks during crosslinking mass spectrometry analysis. Four different structural signatures were identified and reconstructed via state-of-the-art cryo-EM data collection and analysis protocols. The complete pipeline was extended by incorporating AI-driven model prediction and experimental refinement, as well as *de novo*, machine learning-based sequence identification algorithms. Mass spectrometry, crosslinking mass spectrometry and traditional biochemical enzyme characterization, along with robust statistical analysis not only provided additional layers of verification to our findings, but contributed by filling in the “gaps” that a single structural biology approach, such as cryo-EM could never provide information about, painting all together an integrative, complete map of the protein interactions and structural adaptations that are encountered in a CFS with succinyl-CoA producing capabilities. Moreover, results contributed to the complete biochemical characterization and the structural

recapitulation of the reaction's main player, the OGDHc and its components, E1o, E2o, E3, E3BPo. The workflow presented here can act as a proof-of-concept and a guide on how a comprehensive native protein community member analysis can be achieved. This type of characterization can provide necessary insights on protein community identity, in-community protein interactions, organizational principles and community member structural adaptations. This information, when combined, can contribute to the closer-to-native functional interpretation of large biomolecular assemblies, with implications that span multiple research areas, from metabolite regulation and influence of cellular processes targeted towards disease drivers, to biotechnological applications specialized for product generation.

4.7 Outlook and future goals of native protein community research

There are still open questions that future work needs to address: Do the multiple protein communities that are included in the fraction have in-fraction interplay? How can the biochemical and structural characterization of protein community inter-connections be made possible? Answers to these questions will most likely be obtained through the application of high-throughput cryo-EM, applied to a high-density, fully crosslinked CFS. The extreme heterogeneity, complex flexibility and relative low abundances could possibly be overcome by very large-scale data collection schemes. Their implementation will be possible with the significant advances observed in instrumentation[240], data collection[239] and analysis[104,163] , supplemented by the rapid advancements of AI-driven, high fidelity, protein structure prediction algorithms[106]. New technologies will streamline the characterization of native cell extract-derived cell-free systems and, in the near future, the approach described in this thesis could act as a guideline for the complete characterization and optimization of native CFS pipelines, leading to increased yields in biotechnological production.

5 Summary

In recent years, the multifaceted roles of metabolites have come to light. The classic signal transduction paradigm is being enriched with metabolic signaling, effected by products of the cell's primary metabolism, such as acetyl-coenzyme A, α-ketoglutarate and palmitic acid. These metabolites, via changes in their concentration and compartmentalization can affect cellular communication, control inflammation and even be involved in malignancy. Large enzymatic complexes, termed "metabolons", are the main regulators of metabolite availability, thus controlling with their function critical aspects of cellular fate. In turn, metabolons require multiple different proteins to carry out diverse reaction pathways, forming protein communities by maintaining all reaction partners in proximity. Additionally, in order for a metabolon to accurately move reaction intermediates to and from multiple active sites, flexibility is necessary. This flexibility comes in form of unstructured regions that, with their movement, control the availability of reaction intermediates in order to carry out the complete reaction.

The metabolites mentioned above, acetyl-coenzyme A, α-ketoglutarate and palmitic acid are produced and regulated by the large enzymatic complexes of the pyruvate dehydrogenase complex, the oxoglutarate dehydrogenase complex and the fatty acid synthase respectively. All three represent large macromolecular assemblies at the megadalton weight range and contain a significant amount of unstructured, disordered regions that provide the necessary flexibility for reaction intermediate transfer and regulation. To better structurally understand and characterize these large assemblies, there are two prerequisites: (a) a system that can provide access to them and maintain the native context through multiple different biochemical techniques and (b) the correct structural methodology to visualize and interpret them. The first prerequisite can be satisfied through the use of a cell-free system that is derived from fractionated native cell extracts. Cell-free system technologies have been used in recent years to produce various products of biotechnological interest, but are usually far removed from the real native conditions, as they are extensively engineered to optimize biotechnological yield, while simultaneously being used as "black boxes", lacking any kind of structural characterization. A native, fractionated extract can be used in this context and is accessible by multiple different techniques that can

disentangle its complexity, providing a better understanding for the underlying biochemical mechanisms. The second prerequisite can be leveraged through the use of cryogenic electron microscopy, a structural technique that has been proven in recent years to be able to visualize protein samples of various scales, reaching resolutions that rival other established structural techniques. Combined with classic biochemical assays, integrative computational structural biology techniques, mass spectrometry-based proteomics and robust artificial intelligence-based protein model prediction, tremendous insight can be gained on the organization of large enzymatic assemblies and the protein communities that they are part of, leading to better understanding of metabolite production and availability.

In this thesis, a cell-free system with succinyl-CoA-producing capability is employed, derived from the thermophilic, filamentous fungus *Chaetomium thermophilum*, in order to first understand the structure of multiple different protein community members of the megadalton range, namely the oxoglutarate dehydrogenase complex, the pyruvate dehydrogenase complex, the fatty acid synthase and the pre-60S ribosomal subunit. An AI-guided pipeline is devised in order to fully, structurally characterize these community members *de novo,* enabling insights into their organization and how flexibility in the structure can possibly affect their function. AI-generated models are robust, being able to predict, apart from protein secondary structure, critical structural elements such as subunit interfaces, but were also observed to often overinterpret and introduce folded elements that are absent in the experimental data, demonstrating the necessity for predicted model validation through the use of experimental data from multiple sources.

To continue, this thesis delves further into the organization of the oxoglutarate dehydrogenase complex. In higher detail, previous knowledge described the oxoglutarate dehydrogenase complex as comprised by three different proteins, the E1o, the E2o and the E3. High-resolution cryogenic electron microscopy allows for the *de novo* modeling of the thermophilic, 24-meric E2o core, revealing important structural features, such as increased core compaction and higher energetic stability when compared to a mesophilic counterpart. In parallel, a complete kinetic characterization is performed for all the enzymatic reactions that are carried out by the metabolon, revealing the kinetic parameters for all three different substrates used, α-ketoglutarate, NAD^+ and coenzyme A. Mass spectrometry and crosslinking mass

spectrometry-based proteomics reveal a new, flexible subunit that assists in the complex's assembly, the E3BPo which tethers the E3 in proximity to the E2o core. The relative stoichiometry of the complex (<10 E1o dimers : 24 E2o monomers : ~ 4 E3 : ~ 4 E3BPo) is also calculated and biochemically/structurally validated. Integrative, AI-based modeling provides models for all of the complex's different proteins, which are then systematically fit into a lower-resolution cryo-EM model of the complete complex. Crosslinking and cryo-EM guided energetic refinements of the complex's interfaces reveal electrostatic interactions that play a great part in the stabilization and cycling of the flexible region of the E2o protein that is responsible for transfer of reaction intermediates, but also possibly reveals a new structural role, the tethering of the peripheral subunits to the vicinity of the complex's core ultrastructure. Finally, all distances that the flexible region can travel are calculated based on the produced integrative model, revealing the peculiar character of the specific flexible regions, a character that is rarely captured in other protein structures. In combination, all the above findings are combined to propose a new integrative model for the structure and function of the oxoglutarate dehydrogenase complex in the context of a native, fractionated, cell-free system.

5.1 Zusammenfassung

In den letzten Jahren ist die vielfältige Rolle von Metaboliten ans Licht gekommen. Das klassische Paradigma der Signaltransduktion wird um metabolische Signale erweitert, die durch Produkte des Primärstoffwechsels der Zelle, wie Acetyl-Coenzym A, α-Ketoglutarat und Palmitinsäure, bewirkt werden. Diese Metaboliten können durch Veränderungen ihrer Konzentration und Kompartimentierung die zelluläre Kommunikation beeinflussen, Entzündungen steuern und sogar an Malignität beteiligt sein. Große Enzymkomplexe, die so genannten Metabolone, sind die Hauptregulatoren der Metabolitenverfügbarkeit und steuern mit ihrer Funktion kritische Aspekte des zellulären Schicksals. Metabolone wiederum benötigen mehrere unterschiedliche Proteine, um verschiedene Reaktionswege auszuführen, und bilden Proteingemeinschaften, indem sie alle Reaktionspartner in in räumlicher Nähe halten. Damit ein Metabolit Reaktionszwischenprodukte präzise zwischen mehreren aktiven

Stellen transportieren kann, ist außerdem strukturelle Flexibilität erforderlich. Diese Flexibilität kommt in Form von unstrukturierten Regionen zum Tragen, die durch ihre Bewegung die Verfügbarkeit von Reaktionszwischenprodukten kontrollieren, um die vollständige Reaktion durchzuführen.

Die oben genannten Metaboliten Acetyl-Coenzym A, α-Ketoglutarat und Palmitinsäure werden von den großen Enzymkomplexen des Pyruvat-Dehydrogenase-Komplexes, des Oxoglutarat-Dehydrogenase-Komplexes bzw. der Fettsäure-Synthase hergestellt und reguliert. Alle Komplexe drei stellen große makromolekulare Einheiten im Megadaltonbereich dar und enthalten eine beträchtliche Menge unstrukturierter, ungeordneter Bereiche, die die notwendige Flexibilität für den Transfer und die Regulierung von Reaktionsintermediaten bieten. Um diese großen Einheiten strukturell besser zu verstehen und zu charakterisieren, sind zwei Voraussetzungen erforderlich: (a) ein System, das den Zugang zu ihnen ermöglicht und den nativen Kontext durch mehrere verschiedene biochemische Techniken beibehält, und (b) die richtige strukturelle Methodik, um sie sichtbar zu machen und zu interpretieren. Die erste Voraussetzung kann durch den Einsatz eines zellfreien Systems erfüllt werden, das aus fraktionierten nativen Zellextrakten gewonnen wird. Zellfreie Systemtechnologien wurden in den letzten Jahren zur Herstellung verschiedener Produkte von biotechnologischem Interesse eingesetzt, sind jedoch in der Regel weit von den realen nativen Bedingungen entfernt, da sie in großem Umfang zur Optimierung der biotechnologischen Ausbeute entwickelt wurden, während sie gleichzeitig als "Black Box" ohne jegliche strukturelle Charakterisierung verwendet werden. Native, fraktionierte Zellextrakte können in diesem Zusammenhang verwendet werden und sind für verschiedene Techniken zugänglich, die ihre Komplexität entschlüsseln können und ein besseres Verständnis der zugrunde liegenden biochemischen Mechanismen ermöglichen. Die zweite Voraussetzung kann durch den Einsatz von kryogenen Elektronenmikroskopie erfüllt werden, einer Strukturtechnik, die in den letzten Jahren bewiesen hat, dass sie in der Lage ist, Proteinproben in verschiedenen Größenordnungen zu visualisieren und dabei Auflösungen zu erreichen, die mit anderen etablierten Strukturtechniken konkurrieren. In Kombination mit klassischen biochemischen Assays, integrativen computergestützten strukturbiologischen Techniken, massenspektrometriebasierter Proteomik und robuster, auf künstlicher Intelligenz basierender

Proteinmodellvorhersage können enorme Einblicke in die Organisation großer enzymatischer Proteinkomplexe und der Proteingemeinschaften, zu denen sie gehören, gewonnen werden, was zu einem besseren Verständnis der Metabolitenproduktion und -verfügbarkeit führt.

In dieser Arbeit wird ein zellfreies System mit der Fähigkeit zur Succinyl-CoA-Produktion verwendet, das von dem thermophilen, filamentösen Pilz Chaetomium thermophilum abgeleitet ist, um zunächst die Struktur mehrerer verschiedener Mitglieder der Proteingemeinschaft im Megadalton-Bereich zu verstehen, den Oxoglutarat-Dehydrogenase-Komplex, den Pyruvat-Dehydrogenase-Komplex, die Fettsäure-Synthase und die prä-60S ribosomale Untereinheit. Es wird eine KI-gesteuerte Pipeline entwickelt, um diese Mitglieder der Gemeinschaft de novo vollständig strukturell zu charakterisieren, was Einblicke in ihre Organisation und in die Frage ermöglicht, wie Flexibilität in der Struktur ihre Funktion möglicherweise beeinflussen kann. Die von der künstlichen Intelligenz erzeugten Modelle sind robust und können neben der Sekundärstruktur der Proteine auch kritische Strukturelemente wie die Schnittstellen der Untereinheiten vorhersagen. Es wurde jedoch auch beobachtet, dass sie häufig überinterpretiert werden und gefaltete Elemente einführen, die nicht vorhanden sind, was die Notwendigkeit einer Validierung der vorhergesagten Modelle durch die Verwendung von experimentellen Daten aus verschiedenen Quellen zeigt.

In dieser Arbeit wird die Organisation des Oxoglutarat-Dehydrogenase-Komplexes näher beleuchtet. Bisherige Erkenntnisse beschreiben den Oxoglutarat-Dehydrogenase-Komplex als Komplex, der sich aus der E1o, E2o und E3 Untereinheit zusammensetzt. Hochauflösende kryogene Elektronenmikroskopie ermöglicht die de novo Modellierung des thermophilen, 24-meren E2o-Kerns und offenbart wichtige strukturelle Merkmale wie eine erhöhte Kernverdichtung und eine höhere energetische Stabilität im Vergleich zu einem mesophilen Gegenstück. Parallel dazu wird eine vollständige kinetische Charakterisierung aller enzymatischen Reaktionen durchgeführt, die von dem Metabolon ausgeführt werden, wobei die kinetischen Parameter für alle drei verwendeten Substrate, α-Ketoglutarat, NAD+ und Coenzym A, ermittelt werden. Massenspektrometrie und Proteomik auf der Grundlage von chemischen Quervernetzungen zeigen eine neue, flexible Untereinheit, die den Zuammenbau des Komplexes koordiniert, das E3BPo, welches E3 in der Nähe des

E2o-Kerns fixiert. Die relative Stöchiometrie des Komplexes (<10 E1o-Dimere : 24 E2o-Monomere : ~4 E3 : ~4 E3BPo) wird ebenfalls berechnet und biochemisch/strukturell validiert. Integrative, KI-basierte Modellierung liefert Modelle für alle verschiedenen Proteine des Komplexes, die dann systematisch in ein Kryo-EM-Modell des gesamten Komplexes mit geringerer Auflösung eingepasst werden. Chemische Quervernetzungen und Kryo-EM-geführte energetische Verfeinerungen der Grenzflächen des Komplexes offenbaren elektrostatische Wechselwirkungen, die eine große Rolle bei der Stabilisierung und dem Kreislauf der flexiblen Region des E2o-Proteins spielen, die für den Transfer von Reaktionszwischenprodukten verantwortlich ist, aber möglicherweise auch eine neue strukturelle Rolle offenbaren, nämlich die Bindung der peripheren Untereinheiten an die Nähe der Kern-Ultrastruktur des Komplexes. Schließlich werden alle Entfernungen, die die flexible Region zurücklegen kann, auf der Grundlage des erstellten integrativen Modells berechnet, was den besonderen Charakter der spezifischen flexiblen Regionen offenbart, einen Charakter, der in anderen Proteinstrukturen selten erfasst wird. Alle oben genannten Ergebnisse werden kombiniert, um ein neues integratives Modell für die Struktur und Funktion des Oxoglutarat-Dehydrogenase-Komplexes im Kontext eines nativen, fraktionierten, zellfreien Systems vorzuschlagen.

5.2 Περίληψη

Πρόσφατα, οι πολύπλευροι ρόλοι των μεταβολιτών έχουν έρθει στο προσκήνιο όσον αφορά την κυτταρική και μοριακή κατανόηση της ζωής. Το κλασικό παράδειγμα της μεταγωγής σήματος εμπλουτίζεται με τη μεταβολική σηματοδότηση, η οποία, για παράδειγμα, επιτελείται από προϊόντα του πρωτογενούς μεταβολισμού του κυττάρου, όπως το ακετυλο-συνένζυμο Α, το α-κετογλουταρικό και το παλμιτικό οξύ. Οι συγκεκριμένοι μεταβολίτες, μέσω αλλαγών στη συγκέντρωση και στην διαμερισματοποίησή τους, μπορούν να επηρεάσουν την κυτταρική επικοινωνία, να ελέγξουν την φλεγμονική απόκριση και να εμπλακούν ακόμη και στην κακοήθεια. Μεγάλα ενζυμικά σύμπλοκα, που ονομάζονται "μεταβολώνια", είναι οι κύριοι ρυθμιστές της διαθεσιμότητας των μεταβολιτών, ελέγχοντας έτσι με τη λειτουργία τους κρίσιμες πτυχές της κυτταρικής μοίρας. Με τη σειρά τους, τα μεταβολώνια απαιτούν πολλαπλές διαφορετικές πρωτεΐνες για την εκτέλεση ποικίλων ενζυμικών

αντιδράσεων, σχηματίζοντας έτσι πρωτεϊνικές κοινότητες οι οποίες διατηρούν σε εγγύτητα όλους τους εταίρους μιας ενζυμικής αντίδρασης. Επιπλέον, προκειμένου ένα μεταβολώνιο να μεταφέρει με ακρίβεια τα ενδιάμεσα προϊόντα της αντίδρασης από και προς πολλαπλά ενεργά κέντρα, απαιτείται ευελιξία. Αυτή η ευελιξία έρχεται με τη μορφή μη-δομημένων περιοχών που, με την κίνησή τους, ελέγχουν τη διαθεσιμότητα των ενδιάμεσων προϊόντων της αντίδρασης προκειμένου να πραγματοποιηθεί η πλήρης ενζυμική αντίδραση.

Οι μεταβολίτες που αναφέρθηκαν παραπάνω, το ακετυλο-συνένζυμο Α, το α-κετογλουταρικό και το παλμιτικό οξύ, παράγονται και ρυθμίζονται από τα μεγάλα ενζυμικά σύμπλοκα του συμπλέγματος της αφυδρογονάσης του πυροσταφυλικού, του συμπλέγματος της αφυδρογονάσης του οξογλουταρικού και της συνθάσης των λιπαρών οξέων, αντίστοιχα. Και τα τρία σύμπλοκα αντιπροσωπεύουν μεγάλα μακρομοριακά συγκροτήματα, στην μεγαδαλτόνια τάξη βάρους, και περιέχουν σημαντική ποσότητα μη-δομημένων, ανοργάνωτων δομικών περιοχών που παρέχουν την απαραίτητη ευελιξία για τη μεταφορά και τη ρύθμιση των ενδιάμεσων αντιδράσεων. Για την καλύτερη δομική κατανόηση και τον χαρακτηρισμό αυτών των μεγάλων συμπλεγμάτων, υπάρχουν δύο προαπαιτήσεις: α) ένα σύστημα που να μπορεί να παρέχει πρόσβαση σε αυτά, ενώ παράλληλα να διατηρεί τον εγγενή χαρακτήρα τους, μέσω πολλαπλών διαφορετικών βιοχημικών τεχνικών και β) μια σωστή δομική μεθοδολογία για την οπτικοποίηση και την ερμηνεία τους. Η πρώτη προϋπόθεση μπορεί να ικανοποιηθεί με τη χρήση ενός συστήματος χωρίς κύτταρα που προέρχεται από κλασματοποιημένα εγγενή εκχυλίσματα κυττάρων.

Οι τεχνολογίες συστημάτων χωρίς κύτταρα έχουν χρησιμοποιηθεί τα τελευταία χρόνια για την παραγωγή διαφόρων προϊόντων βιοτεχνολογικού ενδιαφέροντος, αλλά συνήθως απέχουν πολύ από τις πραγματικές εγγενείς συνθήκες, καθώς είναι εκτενώς τροποποιημένα για τη βελτιστοποίηση της βιοτεχνολογικής απόδοσης, ενώ ταυτόχρονα χρησιμοποιούνται ως "μαύρα κουτιά", στερούμενα κάθε είδους δομικού χαρακτηρισμού. Ένα εγγενές, κλασματοποιημένο κυτταρικό εκχύλισμα μπορεί να χρησιμοποιηθεί αντίστοιχα σε αυτό το πλαίσιο, παραμένοντας προσβάσιμο από πολλαπλές διαφορετικές τεχνικές που μπορούν να αποσαφηνίσουν την πολυπλοκότητά του, παρέχοντας έτσι αυξημένη κατανόηση για τους υποκείμενους βιοχημικούς μηχανισμούς. Η δεύτερη προϋπόθεση μπορεί να ικανοποιηθεί μέσω της χρήση της κρυογονικής ηλεκτρονικής μικροσκοπίας, μιας δομικής τεχνικής που έχει

αποδειχθεί τα τελευταία χρόνια ότι μπορεί να απεικονίσει δείγματα πρωτεϊνών διαφόρων κλιμάκων, επιτυγχάνοντας αναλύσεις που εύκολα ανταγωνίζονται άλλες καθιερωμένες δομικές τεχνικές. Η κρυο-ΗΜ πλέον συνδυάζεται με κλασσικές βιοχημικές μεθόδους (π.χ. ελέγχους ενζυμικής ενεργότητας, δοκιμές ανοσοαποτύπωσης, χρωματογραφικές τεχνικές), με συνδυαστικές τεχνικές υπολογιστικής δομικής βιολογίας καθώς και με πρωτεωμικές μελέτες με βάση τη φασματομετρία μάζας. Σε συνδυασμό επίσης με αλγορίθμους που στοχεύουν στην αξιόπιστη πρόβλεψη πρωτεϊνικών μοντέλων με βάση την τεχνητή νοημοσύνη, η κρυο-ΗΜ μπορεί να προσφέρει τεράστια γνώση σχετικά με την οργάνωση μεγάλων ενζυμικών συμπλεγμάτων, καθώς των πρωτεϊνικών κοινοτήτων στις οποίες ανήκουν, οδηγώντας σε καλύτερη κατανόηση της παραγωγής και της διαθεσιμότητας μεταβολιτών.

Στην παρούσα διατριβή, χρησιμοποιείται ένα σύστημα χωρίς κύτταρα με δυνατότητα παραγωγής σουκίνυλο-συνένζυμου Α, το οποίο προέρχεται από τον θερμόφιλο, νηματοειδή μύκητα *Chaetomium thermophilum*, προκειμένου να κατανοηθεί αρχικά η δομή πολλαπλών διαφορετικών μελών πρωτεϊνικών κοινοτήτων της τάξης των μεγαδαλτονίων, και πιο συγκεκριμένα του συμπλέγματος της αφυδρογονάσης του οξογλουταρικού, του συμπλέγματος της αφυδρογονάσης του πυροσταφυλικού, της συνθάσης των λιπαρών οξέων, και της προ-60S ριβοσωμικής υπομονάδας. Στη συνέχεια αναπτύσσεται ένα σύνολο πειραματικών και υπολογιστικών τεχνικών εμπνευσμένων από αλγορίθμους τεχνητής νοημοσύνης για τον ενοποιητικό δομικό χαρακτηρισμό αυτών των μελών κοινοτήτων, *de novo*, επιτρέποντας έτσι την κατανόηση της οργάνωσής τους και του τρόπου με τον οποίο η ευελιξία στη δομή μπορεί ενδεχομένως να επηρεάσει τη λειτουργία τους. Τα μοντέλα που δημιουργούνται με χρήση τεχνητής νοημοσύνης είναι εύρωστα, καθώς είναι σε θέση να προβλέψουν, εκτός από τη δευτεροταγή δομή των πρωτεϊνών, κρίσιμα δομικά στοιχεία, όπως τις επιφάνειες αλληλεπίδρασης μεταξύ των υπομονάδων. Παράλληλα, παρατηρείται ότι τα μοντέλα ΤΝ συχνά υπερερμηνεύουν και εισάγουν στοιχεία δευτεροταγούς και τριτοταγούς δομής που απουσιάζουν από τα πειραματικά δεδομένα, γεγονός που καταδεικνύει την αναγκαιότητα επιβεβαίωσης των θεωρητικών μοντέλων μέσω της χρήσης πειραματικών δεδομένων από πολλαπλές τεχνικές.

Συνεχίζοντας, η παρούσα διατριβή εμβαθύνει περισσότερο στην οργάνωση του θερμόφιλου συμπλόκου της αφυδρογονάσης του οξογλουταρικού. Τα έως τώρα

υπάρχοντα δεδομένα περιέγραφαν το σύμπλοκο της οξογλουταρικής αφυδρογονάσης ως αποτελούμενο από τρεις διαφορετικές πρωτεΐνες, την E1o, την E2o και την E3. Η κρυογονική ηλεκτρονική μικροσκοπία υψηλής ανάλυσης επιτρέπει τη *de novo* προτυποποίηση του 24μερούς πυρήνα της E2o, αποκαλύπτοντας σημαντικά δομικά χαρακτηριστικά, όπως την αυξημένη συμπύκνωση του πυρήνα και την υψηλότερη ενεργειακή σταθερότητα σε σύγκριση με τον αντίστοιχο πυρήνα προερχόμενο από μεσόφιλο οργανισμό. Παράλληλα, πραγματοποιείται πλήρης χαρακτηρισμός ενεργότητας για όλες τις ενζυμικές αντιδράσεις που πραγματοποιούνται από το μεταβολώνιο, αποκαλύπτοντας τις κινητικές παραμέτρους και για τα τρία διαφορετικά υποστρώματα που χρησιμοποιούνται, το α-κετογλουταρικό, το νικοτιναμιδο-αδενινο-δινουκλεοτίδιο και το συνένζυμο Α. Η φασματομετρία μάζας και η πρωτεωμική με βάση τη φασματομετρία μάζας χημικής διασύζευξης αποκαλύπτουν μια νέα, ευέλικτη υπομονάδα που βοηθά στη συναρμολόγηση του συμπλόκου, την E3BPo, η οποία προσδένει την E3 κοντά στον E2o πυρήνα. Επίσης υπολογίζεται η σχετική στοιχειομετρία του συμπλόκου (<10 διμερή E1o : 24 μονομερή E2o : ~ 4 E3 : ~ 4 E3BPo) και επικυρώνεται με βιοχημικά και δομικά δεδομένα.

Η ενοποιητική προτυποποίηση, βασισμένη σε TN, παρέχει μοντέλα για όλες τις διαφορετικές πρωτεΐνες του συμπλόκου, τα οποία στη συνέχεια προσαρμόζονται συστηματικά σε ένα χάρτη κρυο-ΗΜ, χαμηλότερης ανάλυσης, του πλήρους συμπλόκου. Οι καθοδηγούμενες από διασύζευξη και κρυο-ΗΜ ενεργειακές βελτιώσεις των επιφανειών αλληλεπίδρασης μεταξύ των υπομονάδων του μεταβολωνίου αποκαλύπτουν ηλεκτροστατικές αλληλεπιδράσεις που παίζουν σημαντικό ρόλο στη σταθεροποίηση και την μετακίνηση της ευέλικτης περιοχής της πρωτεΐνης E2o. Η περιοχή αυτή, υπεύθυνη για τη μεταφορά των ενδιάμεσων προϊόντων της αντίδρασης, αποκαλύπτεται να έχει πιθανώς έναν επιπλέον δομικό ρόλο, αυτόν της συγκράτησης της E1o και της E3 στην εγγύς περιοχή της υπερδομής του πυρήνα του συμπλόκου.

Τέλος, υπολογίζονται όλες οι αποστάσεις που μπορεί να διανύσει η ευέλικτη περιοχή με βάση το ενοποιημένο μοντέλο του μεταβολωνίου, αποκαλύπτοντας τον ιδιαίτερο χαρακτήρα των συγκεκριμένων ευέλικτων περιοχών, έναν χαρακτήρα που σπάνια αποτυπώνεται σε άλλες πρωτεϊνικές δομές. Συνδυαστικά, όλα τα παραπάνω ευρήματα προτείνουν ένα πιο ολοκληρωμένο μοντέλο για τη δομή και τη λειτουργία του συμπλόκου της αφυδρογονάσης του οξογλουταρικού.

6 Literature

1 Krauss, G. *Biochemistry of signal transduction and regulation*. (John Wiley & Sons, 2006).

2 Novick, A. & Weiner, M. Enzyme Induction as an All-or-None Phenomenon. *Proc Natl Acad Sci U S A* **43**, 553-566 (1957). https://doi.org:10.1073/pnas.43.7.553

3 Honjo, T., Nishizuka, Y. & Hayaishi, O. Diphtheria toxin-dependent adenosine diphosphate ribosylation of aminoacyl transferase II and inhibition of protein synthesis. *J Biol Chem* **243**, 3553-3555 (1968).

4 Wellen, K. E. & Thompson, C. B. A two-way street: reciprocal regulation of metabolism and signalling. *Nat Rev Mol Cell Biol* **13**, 270-276 (2012). https://doi.org:10.1038/nrm3305

5 Gut, P. & Verdin, E. The nexus of chromatin regulation and intermediary metabolism. *Nature* **502**, 489-498 (2013). https://doi.org:10.1038/nature12752

6 Pietrocola, F., Galluzzi, L., Bravo-San Pedro, J. M., Madeo, F. & Kroemer, G. Acetyl coenzyme A: a central metabolite and second messenger. *Cell Metab* **21**, 805-821 (2015). https://doi.org:10.1016/j.cmet.2015.05.014

7 Shi, L. & Tu, B. P. Acetyl-CoA and the regulation of metabolism: mechanisms and consequences. *Curr Opin Cell Biol* **33**, 125-131 (2015). https://doi.org:10.1016/j.ceb.2015.02.003

8 Belew, G. D. *et al.* Transfer of glucose hydrogens via acetyl-CoA, malonyl-CoA, and NADPH to fatty acids during de novo lipogenesis. *J Lipid Res* **60**, 2050-2056 (2019). https://doi.org:10.1194/jlr.RA119000354

9 Schroeder, S. *et al.* Acetyl-coenzyme A: a metabolic master regulator of autophagy and longevity. *Autophagy* **10**, 1335-1337 (2014). https://doi.org:10.4161/auto.28919

10 Marino, G. *et al.* Regulation of autophagy by cytosolic acetyl-coenzyme A. *Mol Cell* **53**, 710-725 (2014). https://doi.org:10.1016/j.molcel.2014.01.016

11 McCoy, F. *et al.* Metabolic activation of CaMKII by coenzyme A. *Mol Cell* **52**, 325-339 (2013). https://doi.org:10.1016/j.molcel.2013.08.043

12 Skalidis, I., Tuting, C. & Kastritis, P. L. Unstructured regions of large enzymatic complexes control the availability of metabolites with signaling functions. *Cell Commun Signal* **18**, 136 (2020). https://doi.org:10.1186/s12964-020-00631-9

13 Ree, R., Varland, S. & Arnesen, T. Spotlight on protein N-terminal acetylation. *Exp Mol Med* **50**, 1-13 (2018). https://doi.org:10.1038/s12276-018-0116-z

14 Sadoul, K., Wang, J., Diagouraga, B. & Khochbin, S. The tale of protein lysine acetylation in the cytoplasm. *J Biomed Biotechnol* **2011**, 970382 (2011). https://doi.org:10.1155/2011/970382

15 Sabari, B. R., Zhang, D., Allis, C. D. & Zhao, Y. Metabolic regulation of gene expression through histone acylations. *Nat Rev Mol Cell Biol* **18**, 90-101 (2017). https://doi.org:10.1038/nrm.2016.140

16 Choudhary, C., Weinert, B. T., Nishida, Y., Verdin, E. & Mann, M. The growing landscape of lysine acetylation links metabolism and cell signalling. *Nat Rev Mol Cell Biol* **15**, 536-550 (2014). https://doi.org:10.1038/nrm3841

17 Janke, C. & Montagnac, G. Causes and Consequences of Microtubule Acetylation. *Curr Biol* **27**, R1287-R1292 (2017). https://doi.org:10.1016/j.cub.2017.10.044

18 Reed, N. A. *et al.* Microtubule acetylation promotes kinesin-1 binding and transport. *Curr Biol* **16**, 2166-2172 (2006). https://doi.org:10.1016/j.cub.2006.09.014

19 Brooks, C. L. & Gu, W. The impact of acetylation and deacetylation on the p53 pathway. *Protein Cell* **2**, 456-462 (2011). https://doi.org:10.1007/s13238-011-1063-9

20 Harrison, A. P. & Pierzynowski, S. G. Biological effects of 2-oxoglutarate with particular emphasis on the regulation of protein, mineral and lipid absorption/metabolism, muscle performance, kidney function, bone formation and cancerogenesis, all viewed from a healthy ageing perspective state of the art--review article. *J Physiol Pharmacol* **59 Suppl 1**, 91-106 (2008).

21 Muhling, J. *et al.* Effects of alpha-ketoglutarate on neutrophil intracellular amino and alpha-keto acid profiles and ROS production. *Amino Acids* **38**, 167-177 (2010). https://doi.org:10.1007/s00726-008-0224-5

22 Poitry, S., Poitry-Yamate, C., Ueberfeld, J., MacLeish, P. R. & Tsacopoulos, M. Mechanisms of glutamate metabolic signaling in retinal glial (Muller) cells. *J Neurosci* **20**, 1809-1821 (2000).

23 Kaelin, W. G., Jr. & Ratcliffe, P. J. Oxygen sensing by metazoans: the central role of the HIF hydroxylase pathway. *Mol Cell* **30**, 393-402 (2008). https://doi.org:10.1016/j.molcel.2008.04.009

24 Wang, G. L. & Semenza, G. L. Oxygen sensing and response to hypoxia by mammalian cells. *Redox Rep* **2**, 89-96 (1996). https://doi.org:10.1080/13510002.1996.11747034

25 Semenza, G. L. Hypoxia, clonal selection, and the role of HIF-1 in tumor progression. *Crit Rev Biochem Mol Biol* **35**, 71-103 (2000). https://doi.org:10.1080/10409230091169186

26 Mole, D. R., Maxwell, P. H., Pugh, C. W. & Ratcliffe, P. J. Regulation of HIF by the von Hippel-Lindau tumour suppressor: implications for cellular oxygen sensing. *IUBMB Life* **52**, 43-47 (2001). https://doi.org:10.1080/15216540252774757

27 Chin, R. M. *et al.* The metabolite alpha-ketoglutarate extends lifespan by inhibiting ATP synthase and TOR. *Nature* **510**, 397-401 (2014). https://doi.org:10.1038/nature13264

28 Vatrinet, R. *et al.* The alpha-ketoglutarate dehydrogenase complex in cancer metabolic plasticity. *Cancer Metab* **5**, 3 (2017). https://doi.org:10.1186/s40170-017-0165-0

29 Carta, G., Murru, E., Banni, S. & Manca, C. Palmitic Acid: Physiological Role, Metabolism and Nutritional Implications. *Front Physiol* **8**, 902 (2017). https://doi.org:10.3389/fphys.2017.00902

30 Fatima, S. *et al.* Palmitic acid is an intracellular signaling molecule involved in disease development. *Cell Mol Life Sci* **76**, 2547-2557 (2019). https://doi.org:10.1007/s00018-019-03092-7

31 Flaveny, C. A. *et al.* Broad Anti-tumor Activity of a Small Molecule that Selectively Targets the Warburg Effect and Lipogenesis. *Cancer Cell* **28**, 42-56 (2015). https://doi.org:10.1016/j.ccell.2015.05.007

32 Palomer, X., Pizarro-Delgado, J., Barroso, E. & Vazquez-Carrera, M. Palmitic and Oleic Acid: The Yin and Yang of Fatty Acids in Type 2 Diabetes Mellitus. *Trends Endocrinol Metab* **29**, 178-190 (2018). https://doi.org:10.1016/j.tem.2017.11.009

33 Ajuwon, K. M. & Spurlock, M. E. Palmitate activates the NF-kappaB transcription factor and induces IL-6 and TNFalpha expression in 3T3-L1 adipocytes. *J Nutr* **135**, 1841-1846 (2005). https://doi.org:10.1093/jn/135.8.1841

34 Sramek, J., Nemcova-Furstova, V. & Kovar, J. Kinase Signaling in Apoptosis Induced by Saturated Fatty Acids in Pancreatic beta-Cells. *Int J Mol Sci* **17** (2016). https://doi.org:10.3390/ijms17091400

35 Lin, L. *et al.* Functional lipidomics: Palmitic acid impairs hepatocellular carcinoma development by modulating membrane fluidity and glucose metabolism. *Hepatology* **66**, 432-448 (2017). https://doi.org:10.1002/hep.29033

36 Linder, M. E. & Deschenes, R. J. New insights into the mechanisms of protein palmitoylation. *Biochemistry* **42**, 4311-4320 (2003). https://doi.org:10.1021/bi034159a

37 Baekkeskov, S. & Kanaani, J. Palmitoylation cycles and regulation of protein function (Review). *Mol Membr Biol* **26**, 42-54 (2009). https://doi.org:10.1080/09687680802680108

38 Blaskovic, S., Blanc, M. & van der Goot, F. G. What does S-palmitoylation do to membrane proteins? *FEBS J* **280**, 2766-2774 (2013). https://doi.org:10.1111/febs.12263

39 Goodwin, J. S. *et al.* Depalmitoylated Ras traffics to and from the Golgi complex via a nonvesicular pathway. *J Cell Biol* **170**, 261-272 (2005). https://doi.org:10.1083/jcb.200502063

40 Salaun, C., Greaves, J. & Chamberlain, L. H. The intracellular dynamic of protein palmitoylation. *J Cell Biol* **191**, 1229-1238 (2010). https://doi.org:10.1083/jcb.201008160

41 Ciszak, E. M., Korotchkina, L. G., Dominiak, P. M., Sidhu, S. & Patel, M. S. Structural basis for flip-flop action of thiamin pyrophosphate-dependent enzymes revealed by human pyruvate dehydrogenase. *J Biol Chem* **278**, 21240-21246 (2003). https://doi.org:10.1074/jbc.M300339200

42 Ciszak, E. M. *et al.* How dihydrolipoamide dehydrogenase-binding protein binds dihydrolipoamide dehydrogenase in the human pyruvate dehydrogenase complex. *J Biol Chem* **281**, 648-655 (2006). https://doi.org:10.1074/jbc.M507850200

43 Devedjiev, Y., Steussy, C. N. & Vassylyev, D. G. Crystal structure of an asymmetric complex of pyruvate dehydrogenase kinase 3 with lipoyl domain 2 and its biological implications. *J Mol Biol* **370**, 407-416 (2007). https://doi.org:10.1016/j.jmb.2007.04.083

44 Jiang, J. *et al.* Atomic Structure of the E2 Inner Core of Human Pyruvate Dehydrogenase Complex. *Biochemistry* **57**, 2325-2334 (2018). https://doi.org:10.1021/acs.biochem.8b00357

45 Kalia, Y. N. *et al.* The high-resolution structure of the peripheral subunit-binding domain of dihydrolipoamide acetyltransferase from the pyruvate dehydrogenase multienzyme complex of Bacillus stearothermophilus. *J Mol Biol* **230**, 323-341 (1993). https://doi.org:10.1006/jmbi.1993.1145

46 Perham, R. N. Swinging arms and swinging domains in multifunctional enzymes: catalytic machines for multistep reactions. *Annu Rev Biochem* **69**, 961-1004 (2000). https://doi.org:10.1146/annurev.biochem.69.1.961

47 Miles, J. S., Guest, J. R., Radford, S. E. & Perham, R. N. Investigation of the mechanism of active site coupling in the pyruvate dehydrogenase multienzyme

complex of Escherichia coli by protein engineering. *J Mol Biol* **202**, 97-106 (1988). https://doi.org:10.1016/0022-2836(88)90522-0

48 Lengyel, J. S. *et al.* Extended polypeptide linkers establish the spatial architecture of a pyruvate dehydrogenase multienzyme complex. *Structure* **16**, 93-103 (2008). https://doi.org:10.1016/j.str.2007.10.017

49 Green, J. D., Perham, R. N., Ullrich, S. J. & Appella, E. Conformational studies of the interdomain linker peptides in the dihydrolipoyl acetyltransferase component of the pyruvate dehydrogenase multienzyme complex of Escherichia coli. *J Biol Chem* **267**, 23484-23488 (1992).

50 Radford, S. E., Laue, E. D., Perham, R. N., Martin, S. R. & Appella, E. Conformational flexibility and folding of synthetic peptides representing an interdomain segment of polypeptide chain in the pyruvate dehydrogenase multienzyme complex of Escherichia coli. *J Biol Chem* **264**, 767-775 (1989).

51 Hezaveh, S., Zeng, A. P. & Jandt, U. Full Enzyme Complex Simulation: Interactions in Human Pyruvate Dehydrogenase Complex. *J Chem Inf Model* **58**, 362-369 (2018). https://doi.org:10.1021/acs.jcim.7b00557

52 Harris, R. A., Bowker-Kinley, M. M., Wu, P., Jeng, J. & Popov, K. M. Dihydrolipoamide dehydrogenase-binding protein of the human pyruvate dehydrogenase complex. DNA-derived amino acid sequence, expression, and reconstitution of the pyruvate dehydrogenase complex. *J Biol Chem* **272**, 19746-19751 (1997). https://doi.org:10.1074/jbc.272.32.19746

53 Yu, X. *et al.* Structures of the human pyruvate dehydrogenase complex cores: a highly conserved catalytic center with flexible N-terminal domains. *Structure* **16**, 104-114 (2008). https://doi.org:10.1016/j.str.2007.10.024

54 Patel, M. S., Nemeria, N. S., Furey, W. & Jordan, F. The pyruvate dehydrogenase complexes: structure-based function and regulation. *J Biol Chem* **289**, 16615-16623 (2014). https://doi.org:10.1074/jbc.R114.563148

55 Saunier, E., Benelli, C. & Bortoli, S. The pyruvate dehydrogenase complex in cancer: An old metabolic gatekeeper regulated by new pathways and pharmacological agents. *Int J Cancer* **138**, 809-817 (2016). https://doi.org:10.1002/ijc.29564

56 Kato, M. *et al.* Structural basis for inactivation of the human pyruvate dehydrogenase complex by phosphorylation: role of disordered phosphorylation loops. *Structure* **16**, 1849-1859 (2008). https://doi.org:10.1016/j.str.2008.10.010

57 Whitley, M. J. *et al.* Pyruvate dehydrogenase complex deficiency is linked to regulatory loop disorder in the alphaV138M variant of human pyruvate dehydrogenase. *J Biol Chem* **293**, 13204-13213 (2018). https://doi.org:10.1074/jbc.RA118.003996

58 Fan, J. *et al.* Tyr-301 phosphorylation inhibits pyruvate dehydrogenase by blocking substrate binding and promotes the Warburg effect. *J Biol Chem* **289**, 26533-26541 (2014). https://doi.org:10.1074/jbc.M114.593970

59 Kato, M. *et al.* A synchronized substrate-gating mechanism revealed by cubic-core structure of the bovine branched-chain alpha-ketoacid dehydrogenase complex. *EMBO J* **25**, 5983-5994 (2006). https://doi.org:10.1038/sj.emboj.7601444

60 Sheu, K. F. & Blass, J. P. The alpha-ketoglutarate dehydrogenase complex. *Ann N Y Acad Sci* **893**, 61-78 (1999). https://doi.org:10.1111/j.1749-6632.1999.tb07818.x

61 Heublein, M. *et al.* The novel component Kgd4 recruits the E3 subunit to the mitochondrial alpha-ketoglutarate dehydrogenase. *Mol Biol Cell* **25**, 3342-3349 (2014). https://doi.org:10.1091/mbc.E14-07-1178

62 Zhou, J. *et al.* A multipronged approach unravels unprecedented protein-protein interactions in the human 2-oxoglutarate dehydrogenase multienzyme complex. *J Biol Chem* **293**, 19213-19227 (2018). https://doi.org:10.1074/jbc.RA118.005432

63 Zhong, Y. *et al.* Structural basis for the activity and regulation of human alpha-ketoglutarate dehydrogenase revealed by Cryo-EM. *Biochem Biophys Res Commun* **602**, 120-126 (2022). https://doi.org:10.1016/j.bbrc.2022.02.093

64 Perham, R. N. Domains, motifs, and linkers in 2-oxo acid dehydrogenase multienzyme complexes: a paradigm in the design of a multifunctional protein. *Biochemistry* **30**, 8501-8512 (1991). https://doi.org:10.1021/bi00099a001

65 Qi, F., Pradhan, R. K., Dash, R. K. & Beard, D. A. Detailed kinetics and regulation of mammalian 2-oxoglutarate dehydrogenase. *BMC Biochem* **12**, 53 (2011). https://doi.org:10.1186/1471-2091-12-53

66 Gibson, G. E. *et al.* Alpha-ketoglutarate dehydrogenase complex-dependent succinylation of proteins in neurons and neuronal cell lines. *J Neurochem* **134**, 86-96 (2015). https://doi.org:10.1111/jnc.13096

67 Wakil, S. J. Fatty acid synthase, a proficient multifunctional enzyme. *Biochemistry* **28**, 4523-4530 (1989). https://doi.org:10.1021/bi00437a001

68 Johansson, P. *et al.* Multimeric options for the auto-activation of the Saccharomyces cerevisiae FAS type I megasynthase. *Structure* **17**, 1063-1074 (2009). https://doi.org:10.1016/j.str.2009.06.014

69 Anselmi, C., Grininger, M., Gipson, P. & Faraldo-Gomez, J. D. Mechanism of substrate shuttling by the acyl-carrier protein within the fatty acid mega-synthase. *J Am Chem Soc* **132**, 12357-12364 (2010). https://doi.org:10.1021/ja103354w

70 Singh, K. *et al.* Discovery of a Regulatory Subunit of the Yeast Fatty Acid Synthase. *Cell* **180**, 1130-1143 e1120 (2020). https://doi.org:10.1016/j.cell.2020.02.034

71 Leibundgut, M., Jenni, S., Frick, C. & Ban, N. Structural basis for substrate delivery by acyl carrier protein in the yeast fatty acid synthase. *Science* **316**, 288-290 (2007). https://doi.org:10.1126/science.1138249

72 Kastritis, P. L. *et al.* Capturing protein communities by structural proteomics in a thermophilic eukaryote. *Mol Syst Biol* **13**, 936 (2017). https://doi.org:10.15252/msb.20167412

73 Gipson, P. *et al.* Direct structural insight into the substrate-shuttling mechanism of yeast fatty acid synthase by electron cryomicroscopy. *Proc Natl Acad Sci U S A* **107**, 9164-9169 (2010). https://doi.org:10.1073/pnas.0913547107

74 Maier, T., Leibundgut, M. & Ban, N. The crystal structure of a mammalian fatty acid synthase. *Science* **321**, 1315-1322 (2008). https://doi.org:10.1126/science.1161269

75 Lomakin, I. B., Xiong, Y. & Steitz, T. A. The crystal structure of yeast fatty acid synthase, a cellular machine with eight active sites working together. *Cell* **129**, 319-332 (2007). https://doi.org:10.1016/j.cell.2007.03.013

76 Gavin, A. C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631-636 (2006). https://doi.org:10.1038/nature04532

77 Kuhner, S. *et al.* Proteome organization in a genome-reduced bacterium. *Science* **326**, 1235-1240 (2009). https://doi.org:10.1126/science.1176343

78 Srere, P. A. The metabolon. *Trends in Biochemical Sciences* **10**, 109-110 (1985). https://doi.org:https://doi.org/10.1016/0968-0004(85)90266-X

79 Kyrilis, F. L. *et al.* Integrative structure of a 10-megadalton eukaryotic pyruvate dehydrogenase complex from native cell extracts. *Cell Rep* **34**, 108727 (2021). https://doi.org:10.1016/j.celrep.2021.108727

80 Tuting, C. *et al.* Cryo-EM snapshots of a native lysate provide structural insights into a metabolon-embedded transacetylase reaction. *Nat Commun* **12**, 6933 (2021). https://doi.org:10.1038/s41467-021-27287-4

81 Kyrilis, F. L., Meister, A. & Kastritis, P. L. Integrative biology of native cell extracts: a new era for structural characterization of life processes. *Biol Chem* **400**, 831-846 (2019). https://doi.org:10.1515/hsz-2018-0445

82 Leighton, F. *et al.* The large-scale separation of peroxisomes, mitochondria, and lysosomes from the livers of rats injected with triton WR-1339. Improved isolation procedures, automated analysis, biochemical and morphological properties of fractions. *J Cell Biol* **37**, 482-513 (1968). https://doi.org:10.1083/jcb.37.2.482

83 Beaufay, H. *et al.* Analytical study of microsomes and isolated subcellular membranes from rat liver. 3. Subfractionation of the microsomal fraction by isopycnic and differential centrifugation in density gradients. *J Cell Biol* **61**, 213-231 (1974). https://doi.org:10.1083/jcb.61.1.213

84 Palade, G. E. & Siekevitz, P. Liver microsomes; an integrated morphological and biochemical study. *J Biophys Biochem Cytol* **2**, 171-200 (1956). https://doi.org:10.1083/jcb.2.2.171

85 De Duve, C. & Berthet, J. in *International Review of Cytology* Vol. 3 (eds G. H. Bourne & J. F. Danielli) 225-275 (Academic Press, 1954).

86 Ehrenreich, J. H., Bergeron, J. J., Siekevitz, P. & Palade, G. E. Golgi fractions prepared from rat liver homogenates. I. Isolation procedure and morphological characterization. *J Cell Biol* **59**, 45-72 (1973). https://doi.org:10.1083/jcb.59.1.45

87 Claude, A. & Fullam, E. F. An Electron Microscope Study of Isolated Mitochondria : Method and Preliminary Results. *J Exp Med* **81**, 51-62 (1945). https://doi.org:10.1084/jem.81.1.51

88 Kühlbrandt, W. Cryo-EM enters a new era. *Elife* **3**, e03678 (2014). https://doi.org:10.7554/eLife.03678

89 Kühlbrandt, W. The Resolution Revolution. *Science* **343**, 1443-1444 (2014). https://doi.org:10.1126/science.1251652

90 Abbas, Y. M., Wu, D., Bueler, S. A., Robinson, C. V. & Rubinstein, J. L. Structure of V-ATPase from the mammalian brain. *Science* **367**, 1240-1246 (2020). https://doi.org:10.1126/science.aaz2924

91 Zhang, W. *et al.* Novel tau filament fold in corticobasal degeneration. *Nature* **580**, 283-287 (2020). https://doi.org:10.1038/s41586-020-2043-0

92 Braunger, K. *et al.* Structural basis for coupling protein transport and N-glycosylation at the mammalian endoplasmic reticulum. *Science* **360**, 215-219 (2018). https://doi.org:10.1126/science.aar7899

93 Yip, K. M., Fischer, N., Paknia, E., Chari, A. & Stark, H. Atomic-resolution protein structure determination by cryo-EM. *Nature* **587**, 157-161 (2020). https://doi.org:10.1038/s41586-020-2833-4

94 Nakane, T. *et al.* Single-particle cryo-EM at atomic resolution. *Nature* **587**, 152-156 (2020). https://doi.org:10.1038/s41586-020-2829-0

95 Passmore, L. A. & Russo, C. J. Specimen Preparation for High-Resolution Cryo-EM. *Methods Enzymol* **579**, 51-86 (2016). https://doi.org:10.1016/bs.mie.2016.04.011

96 Zivanov, J., Nakane, T. & Scheres, S. H. W. Estimation of high-order aberrations and anisotropic magnification from cryo-EM data sets in RELION-3.1. *IUCrJ* **7**, 253-267 (2020). https://doi.org:10.1107/S2052252520000081

97 de la Rosa-Trevin, J. M. *et al.* Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy. *J Struct Biol* **195**, 93-99 (2016). https://doi.org:10.1016/j.jsb.2016.04.010

98 Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods* **14**, 290-296 (2017). https://doi.org:10.1038/nmeth.4169

99 Tegunov, D. & Cramer, P. Real-time cryo-electron microscopy data preprocessing with Warp. *Nat Methods* **16**, 1146-1152 (2019). https://doi.org:10.1038/s41592-019-0580-y

100 Danev, R., Yanagisawa, H. & Kikkawa, M. Cryo-Electron Microscopy Methodology: Current Aspects and Future Directions. *Trends Biochem Sci* **44**, 837-848 (2019). https://doi.org:10.1016/j.tibs.2019.04.008

101 Saibil, H. R. Conformational changes studied by cryo-electron microscopy. *Nature Structural Biology* **7**, 711-714 (2000). https://doi.org:10.1038/78923

102 Ciccarelli, L. *et al.* Structure and conformational variability of the mycobacterium tuberculosis fatty acid synthase multienzyme complex. *Structure* **21**, 1251-1257 (2013). https://doi.org:10.1016/j.str.2013.04.023

103 Behrmann, E. *et al.* Structural snapshots of actively translating human ribosomes. *Cell* **161**, 845-857 (2015). https://doi.org:10.1016/j.cell.2015.03.052

104 Punjani, A. & Fleet, D. J. 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *J Struct Biol* **213**, 107702 (2021). https://doi.org:10.1016/j.jsb.2021.107702

105 O'Reilly, F. J. *et al.* In-cell architecture of an actively transcribing-translating expressome. *Science* **369**, 554-557 (2020). https://doi.org:10.1126/science.abb3758

106 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021). https://doi.org:10.1038/s41586-021-03819-2

107 Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871-876 (2021). https://doi.org:10.1126/science.abj8754

108 Park, S. Y., Yang, D., Ha, S. H. & Lee, S. Y. Metabolic Engineering of Microorganisms for the Production of Natural Compounds. *Advanced Biosystems* **2**, 1700190 (2018). https://doi.org:https://doi.org/10.1002/adbi.201700190

109 Lee, S. Y. & Kim, H. U. Systems strategies for developing industrial microbial strains. *Nat Biotechnol* **33**, 1061-1072 (2015). https://doi.org:10.1038/nbt.3365

110 Liu, Y., Liu, Q., Krivoruchko, A., Khoomrung, S. & Nielsen, J. Engineering yeast phospholipid metabolism for de novo oleoylethanolamide production. *Nat Chem Biol* **16**, 197-205 (2020). https://doi.org:10.1038/s41589-019-0431-2

111 Lee, S. Y. & Park, J. H. Integration of systems biology with bioprocess engineering: L: -threonine production by systems metabolic engineering of Escherichia coli. *Adv Biochem Eng Biotechnol* **120**, 1-19 (2010). https://doi.org:10.1007/10_2009_57

112 Mills, T. Y., Sandoval, N. R. & Gill, R. T. Cellulosic hydrolysate toxicity and tolerance mechanisms in Escherichia coli. *Biotechnol Biofuels* **2**, 26 (2009). https://doi.org:10.1186/1754-6834-2-26

113 Pei, L. & Schmidt, M. Fast-Growing Engineered Microbes: New Concerns for Gain-of-Function Research? *Front Genet* **9**, 207 (2018). https://doi.org:10.3389/fgene.2018.00207

114 Prakash, D., Verma, S., Bhatia, R. & Tiwary, B. N. Risks and Precautions of Genetically Modified Organisms. *ISRN Ecology* **2011**, 369573 (2011). https://doi.org:10.5402/2011/369573

115 Bowie, J. U. *et al.* Synthetic Biochemistry: The Bio-inspired Cell-Free Approach to Commodity Chemical Production. *Trends Biotechnol* **38**, 766-778 (2020). https://doi.org:10.1016/j.tibtech.2019.12.024

116 Korman, T. P., Opgenorth, P. H. & Bowie, J. U. A synthetic biochemistry platform for cell free production of monoterpenes from glucose. *Nat Commun* **8**, 15526 (2017). https://doi.org:10.1038/ncomms15526

117 Karim, A. S. & Jewett, M. C. A cell-free framework for rapid biosynthetic pathway prototyping and enzyme discovery. *Metab Eng* **36**, 116-126 (2016). https://doi.org:10.1016/j.ymben.2016.03.002

118 Tan, G. Y., Zhu, F., Deng, Z. & Liu, T. In vitro reconstitution guide for targeted synthetic metabolism of chemicals, nutraceuticals and drug precursors. *Synth Syst Biotechnol* **1**, 25-33 (2016). https://doi.org:10.1016/j.synbio.2016.02.003

119 Shimizu, Y. *et al.* Cell-free translation reconstituted with purified components. *Nat Biotechnol* **19**, 751-755 (2001). https://doi.org:10.1038/90802

120 Mullis, K. B. The unusual origin of the polymerase chain reaction. *Sci Am* **262**, 56-61, 64-55 (1990). https://doi.org:10.1038/scientificamerican0490-56

121 Richardson, K. N., Black, W. B. & Li, H. Aldehyde Production in Crude Lysate- and Whole Cell-Based Biotransformation Using a Noncanonical Redox Cofactor System. *ACS Catal* **10**, 8898-8903 (2020). https://doi.org:10.1021/acscatal.0c03070

122 Kay, J. E. & Jewett, M. C. Lysate of engineered Escherichia coli supports high-level conversion of glucose to 2,3-butanediol. *Metab Eng* **32**, 133-142 (2015). https://doi.org:10.1016/j.ymben.2015.09.015

123 Skalidis, I. *et al.* Cryo-EM and artificial intelligence visualize endogenous protein community members. *Structure* **30**, 575-589 e576 (2022). https://doi.org:10.1016/j.str.2022.01.001

124 Liu, X. *et al.* Biosynthetic Pathway and Metabolic Engineering of Succinic Acid. *Front Bioeng Biotechnol* **10**, 843887 (2022). https://doi.org:10.3389/fbioe.2022.843887

125 Tosaka, O., Enei, H. & Hirose, Y. The production of l-lysine by fermentation. *Trends in Biotechnology* **1**, 70-74 (1983). https://doi.org:https://doi.org/10.1016/0167-7799(83)90055-0

126 Kind, S., Becker, J. & Wittmann, C. Increased lysine production by flux coupling of the tricarboxylic acid cycle and the lysine biosynthetic pathway--metabolic engineering of the availability of succinyl-CoA in Corynebacterium glutamicum. *Metab Eng* **15**, 184-195 (2013). https://doi.org:10.1016/j.ymben.2012.07.005

127 Lin, B. *et al.* Reconstitution of TCA cycle with DAOCS to engineer Escherichia coli into an efficient whole cell catalyst of penicillin G. *Proc Natl Acad Sci U S A* **112**, 9855-9859 (2015). https://doi.org:10.1073/pnas.1502866112

128 Lee, J. S., Lin, C. J., Lee, W. C., Teng, H. Y. & Chuang, M. H. Production of succinic acid through the fermentation of Actinobacillus succinogenes on the

hydrolysate of Napier grass. *Biotechnol Biofuels Bioprod* **15**, 9 (2022). https://doi.org:10.1186/s13068-022-02106-0

129 Araujo, W. L. *et al.* On the role of the mitochondrial 2-oxoglutarate dehydrogenase complex in amino acid metabolism. *Amino Acids* **44**, 683-700 (2013). https://doi.org:10.1007/s00726-012-1392-x

130 Almaas, E., Kovacs, B., Vicsek, T., Oltvai, Z. N. & Barabasi, A. L. Global organization of metabolic fluxes in the bacterium Escherichia coli. *Nature* **427**, 839-843 (2004). https://doi.org:10.1038/nature02289

131 Bunik, V. I. & Fernie, A. R. Metabolic control exerted by the 2-oxoglutarate dehydrogenase reaction: a cross-kingdom comparison of the crossroad between energy production and nitrogen assimilation. *Biochem J* **422**, 405-421 (2009). https://doi.org:10.1042/BJ20090722

132 Matsumoto, K. *et al.* 2-Oxoglutarate downregulates expression of vascular endothelial growth factor and erythropoietin through decreasing hypoxia-inducible factor-1alpha and inhibits angiogenesis. *J Cell Physiol* **209**, 333-340 (2006). https://doi.org:10.1002/jcp.20733

133 Wang, X. W. *et al.* Taxonomy, phylogeny and identification of Chaetomiaceae with emphasis on thermophilic species. *Stud. Mycol.* (2022). https://doi.org:10.3114/sim.2022.101.03

134 Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, 2021.2010.2004.463034 (2021). https://doi.org:10.1101/2021.10.04.463034

135 Chojnowski, G., Sobolev, E., Heuser, P. & Lamzin, V. S. The accuracy of protein models automatically built into cryo-EM maps with ARP/wARP. *Acta Crystallogr D Struct Biol* **77**, 142-150 (2021). https://doi.org:10.1107/S2059798320016332

136 Spitzer, M., Wildenhain, J., Rappsilber, J. & Tyers, M. BoxPlotR: a web tool for generation of box plots. *Nat Methods* **11**, 121-122 (2014). https://doi.org:10.1038/nmeth.2811

137 Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679-682 (2022). https://doi.org:10.1038/s41592-022-01488-1

138 Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60**, 2126-2132 (2004). https://doi.org:10.1107/S0907444904019158

139 Otasek, D., Morris, J. H., Boucas, J., Pico, A. R. & Demchak, B. Cytoscape Automation: empowering workflow-based network analysis. *Genome Biol* **20**, 185 (2019). https://doi.org:10.1186/s13059-019-1758-4

140 Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat Methods* **9**, 676-682 (2012). https://doi.org:10.1038/nmeth.2019

141 Chojnowski, G. *et al.* findMySequence: a neural-network-based approach for identification of unknown proteins in X-ray crystallography and cryo-EM. *IUCrJ* **9** (2022). https://doi.org:doi:10.1107/S2052252521011088

142 Kassambara, A. & Kassambara, M. A. Package 'ggpubr'. *R package version 0.1* **6** (2020).

143 de Vries, S. J., van Dijk, M. & Bonvin, A. M. J. J. The HADDOCK web server for data-driven biomolecular docking. *Nature Protocols* **5**, 883-897 (2010). https://doi.org:10.1038/nprot.2010.32

144 Croll, T. I. ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr D Struct Biol* **74**, 519-530 (2018). https://doi.org:10.1107/S2059798318002425

145 Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191 (2009). https://doi.org:10.1093/bioinformatics/btp033

146 Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols* **11**, 2301-2319 (2016). https://doi.org:10.1038/nprot.2016.136

147 Suzuki, H., Kawabata, T. & Nakamura, H. Omokage search: shape similarity search service for biomolecular structures in both the PDB and EMDB. *Bioinformatics* **32**, 619-620 (2015). https://doi.org:10.1093/bioinformatics/btv614

148 McKinney, W. in *Proceedings of the 9th Python in Science Conference.* 51-56 (Austin, TX).

149 Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol* **75**, 861-877 (2019). https://doi.org:10.1107/S2059798319011471

150 Zivanov, J. *et al.* New tools for automated high-resolution cryo-EM structure determination in RELION-3. *Elife* **7** (2018). https://doi.org:10.7554/eLife.42166

151 Pintilie, G. & Chiu, W. Comparison of Segger and other methods for segmentation and rigid-body docking of molecular components in cryo-EM density maps. *Biopolymers* **97**, 742-760 (2012). https://doi.org:10.1002/bip.22074

152 Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* **30**, 70-82 (2021). https://doi.org:10.1002/pro.3943

153 Graham, M., Combe, C., Kolbowski, L. & Rappsilber, J. (bioRxiv, 2019).

154 Spurr, A. R. A low-viscosity epoxy resin embedding medium for electron microscopy. *Journal of ultrastructure research* **26**, 31-43 (1969).

155 Tüting, C. *et al.* Cryo-EM snapshots of a native lysate provide structural insights into a metabolon-embedded transacetylase reaction. *Nature Communications* **12**, 6933 (2021). https://doi.org:10.1038/s41467-021-27287-4

156 Ishiyama, M., Miyazono, Y., Sasamoto, K., Ohkura, Y. & Ueno, K. A highly water-soluble disulfonated tetrazolium salt as a chromogenic indicator for NADH as well as cell viability. *Talanta* **44**, 1299-1305 (1997). https://doi.org:https://doi.org/10.1016/S0039-9140(97)00017-9

157 Lineweaver, H. & Burk, D. The Determination of Enzyme Dissociation Constants. *Journal of the American Chemical Society* **56**, 658-666 (1934). https://doi.org:10.1021/ja01318a036

158 Hughes, C. S. *et al.* Ultrasensitive proteome analysis using paramagnetic bead technology. *Mol Syst Biol* **10**, 757 (2014). https://doi.org:10.15252/msb.20145625

159 Moggridge, S., Sorensen, P. H., Morin, G. B. & Hughes, C. S. Extending the Compatibility of the SP3 Paramagnetic Bead Processing Approach for Proteomics. *J Proteome Res* **17**, 1730-1740 (2018). https://doi.org:10.1021/acs.jproteome.7b00913

160 Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367-1372 (2008). https://doi.org:10.1038/nbt.1511

161 Rappsilber, J., Ishihama, Y. & Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal Chem* **75**, 663-670 (2003). https://doi.org:10.1021/ac026117i

162 Graham, M., Combe, C., Kolbowski, L. & Rappsilber, J. xiView: A common platform for the downstream analysis of Crosslinking Mass Spectrometry data. *bioRxiv*, 561829 (2019). https://doi.org:10.1101/561829

163 Punjani, A., Zhang, H. & Fleet, D. J. Non-uniform refinement: adaptive regularization improves single-particle cryo-EM reconstruction. *Nat Methods* **17**, 1214-1221 (2020). https://doi.org:10.1038/s41592-020-00990-8

164 Groothuizen, F. S. *et al.* MutS/MutL crystal structure reveals that the MutS sliding clamp loads MutL onto DNA. *Elife* **4**, e06744 (2015). https://doi.org:10.7554/eLife.06744

165 Baretic, D. *et al.* Structures of closed and open conformations of dimeric human ATM. *Sci Adv* **3**, e1700933 (2017). https://doi.org:10.1126/sciadv.1700933

166 Dick, R. A. *et al.* Structures of immature EIAV Gag lattices reveal a conserved role for IP6 in lentivirus assembly. *PLoS Pathog* **16**, e1008277 (2020). https://doi.org:10.1371/journal.ppat.1008277

167 Wang, X. *et al.* Structure of the intact ATM/Tel1 kinase. *Nat Commun* **7**, 11655 (2016). https://doi.org:10.1038/ncomms11655

168 Han, Y., Reyes, A. A., Malik, S. & He, Y. Cryo-EM structure of SWI/SNF complex bound to a nucleosome. *Nature* **579**, 452-455 (2020). https://doi.org:10.1038/s41586-020-2087-1

169 Wieczorek, M. *et al.* Asymmetric Molecular Architecture of the Human gamma-Tubulin Ring Complex. *Cell* **180**, 165-175 e116 (2020). https://doi.org:10.1016/j.cell.2019.12.007

170 Davis, J. H. *et al.* Modular Assembly of the Bacterial Large Ribosomal Subunit. *Cell* **167**, 1610-1622 e1615 (2016). https://doi.org:10.1016/j.cell.2016.11.020

171 Cavadini, S. *et al.* Cullin-RING ubiquitin E3 ligase regulation by the COP9 signalosome. *Nature* **531**, 598-603 (2016). https://doi.org:10.1038/nature17416

172 Nagy, B. *et al.* Structure of the dihydrolipoamide succinyltransferase (E2) component of the human alpha-ketoglutarate dehydrogenase complex (hKGDHc) revealed by cryo-EM and cross-linking mass spectrometry: Implications for the overall hKGDHc structure. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1865**, 129889 (2021). https://doi.org:https://doi.org/10.1016/j.bbagen.2021.129889

173 Chojnowski, G., Pereira, J. & Lamzin, V. S. Sequence assignment for low-resolution modelling of protein crystal structures. *Acta Crystallogr D Struct Biol* **75**, 753-763 (2019). https://doi.org:10.1107/S2059798319009392

174 Liang, X. *et al.* Structural snapshots of human pre-60S ribosomal particles before and after nuclear export. *Nature Communications* **11**, 3542 (2020). https://doi.org:10.1038/s41467-020-17237-x

175 Wu, S. *et al.* Diverse roles of assembly factors revealed by structures of late nuclear pre-60S ribosomes. *Nature* **534**, 133-137 (2016). https://doi.org:10.1038/nature17942

176 Shen, P. S. *et al.* Protein synthesis. Rqc2p and 60S ribosomal subunits mediate mRNA-independent elongation of nascent chains. *Science* **347**, 75-78 (2015). https://doi.org:10.1126/science.1259724

177 Ponder, J. W. & Case, D. A. Force fields for protein simulations. *Adv Protein Chem* **66**, 27-85 (2003). https://doi.org:10.1016/s0065-3233(03)66002-x

178 Lou, J. W., Iyer, K. R., Hasan, S. M. N., Cowen, L. E. & Mazhab-Jafari, M. T. Electron cryomicroscopy observation of acyl carrier protein translocation in type I fungal fatty acid synthase. *Sci Rep* **9**, 12987 (2019). https://doi.org:10.1038/s41598-019-49261-3

179 Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A* **98**, 10037-10041 (2001). https://doi.org:10.1073/pnas.181342398

180 Kastritis, P. L. & Bonvin, A. M. Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. *J Proteome Res* **9**, 2216-2225 (2010). https://doi.org:10.1021/pr9009854

181 Kastritis, P. L., Rodrigues, J. P., Folkers, G. E., Boelens, R. & Bonvin, A. M. Proteins feel more than they see: fine-tuning of binding affinity by properties of the non-interacting surface. *J Mol Biol* **426**, 2632-2652 (2014). https://doi.org:10.1016/j.jmb.2014.04.017

182 Berendsen, H. J. C., Postma, J. P. M., Gunsteren, W. F. v., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **81**, 3684-3690 (1984). https://doi.org:10.1063/1.448118

183 Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr* **54**, 905-921 (1998). https://doi.org:10.1107/s0907444998003254

184 Jorgensen, W. L. & Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J Am Chem Soc* **110**, 1657-1666 (1988). https://doi.org:10.1021/ja00214a001

185 Marsh, J. A. & Forman-Kay, J. D. Sequence determinants of compaction in intrinsically disordered proteins. *Biophys J* **98**, 2383-2390 (2010). https://doi.org:10.1016/j.bpj.2010.02.006

186 Wilkins, D. K. *et al.* Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques. *Biochemistry* **38**, 16424-16431 (1999). https://doi.org:10.1021/bi991765q

187 George, R. A. & Heringa, J. An analysis of protein domain linkers: their classification and role in protein folding. *Protein Eng* **15**, 871-879 (2002). https://doi.org:10.1093/protein/15.11.871

188 Jensen, L. J. *et al.* STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* **37**, D412-416 (2009). https://doi.org:10.1093/nar/gkn760

189 Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* **49**, D605-D612 (2021). https://doi.org:10.1093/nar/gkaa1074

190 Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457-462 (2016). https://doi.org:10.1093/nar/gkv1070

191 Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539 (2011). https://doi.org:10.1038/msb.2011.75

192 Panek, T., Elias, M., Vancova, M., Lukes, J. & Hashimi, H. Returning to the Fold for Lessons in Mitochondrial Crista Diversity and Evolution. *Curr Biol* **30**, R575-R588 (2020). https://doi.org:10.1016/j.cub.2020.02.053

193 Prasad, A. R., Kurup, C. K. & Maheshwari, R. Effect of temperature on respiration of a mesophilic and a thermophilic fungus. *Plant Physiol* **64**, 347-348 (1979). https://doi.org:10.1104/pp.64.2.347

194 Skalidis, I. *et al.* AI-guided cryo-EM probes a thermophilic cell-free system with succinyl-CoA manufacturing capability. *bioRxiv*, 2022.2010.2008.511438 (2022). https://doi.org:10.1101/2022.10.08.511438

195 Bepler, T., Kelley, K., Noble, A. J. & Berger, B. Topaz-Denoise: general deep denoising models for cryoEM and cryoET. *Nat Commun* **11**, 5208 (2020). https://doi.org:10.1038/s41467-020-18952-1

196 Wagner, G. R. *et al.* A Class of Reactive Acyl-CoA Species Reveals the Non-enzymatic Origins of Protein Acylation. *Cell Metab* **25**, 823-837 e828 (2017). https://doi.org:10.1016/j.cmet.2017.03.006

197 Nemeria, N. S. *et al.* Human 2-oxoglutarate dehydrogenase complex E1 component forms a thiamin-derived radical by aerobic oxidation of the enamine intermediate. *J Biol Chem* **289**, 29859-29873 (2014). https://doi.org:10.1074/jbc.M114.591073

198 Pawelczyk, T. & Angielski, S. Effect of ionic strength on the regulatory properties of 2-oxoglutarate dehydrogenase complex. *Biochimie* **74**, 171-176 (1992). https://doi.org:10.1016/0300-9084(92)90042-d

199 Mareck, A., Bessam, H., Delattre, P. & Foucher, B. Purification of the 2-oxoglutarate dehydrogenase and pyruvate dehydrogenase complexes of Neurospora crassa mitochondria. *Biochimie* **68**, 1175-1180 (1986). https://doi.org:10.1016/s0300-9084(86)80061-x

200 Singh, K. *et al.* Discovery of a Regulatory Subunit of the Yeast Fatty Acid Synthase. *Cell* **180**, 1130-1143.e1120 (2020). https://doi.org:10.1016/j.cell.2020.02.034

201 UniProt, C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**, D480-D489 (2021). https://doi.org:10.1093/nar/gkaa1100

202 Forsberg, B. O., Aibara, S., Howard, R. J., Mortezaei, N. & Lindahl, E. Arrangement and symmetry of the fungal E3BP-containing core of the pyruvate dehydrogenase complex. *Nature Communications* **11**, 4667 (2020). https://doi.org:10.1038/s41467-020-18401-z

203 Wu, S. *et al.* Diverse roles of assembly factors revealed by structures of late nuclear pre-60S ribosomes. *Nature* **534**, 133-137 (2016). https://doi.org:10.1038/nature17942

204 Kater, L. *et al.* Visualizing the Assembly Pathway of Nucleolar Pre-60S Ribosomes. *Cell* **171**, 1599-1610 e1514 (2017). https://doi.org:10.1016/j.cell.2017.11.039

205 Ban, N. *et al.* A new system for naming ribosomal proteins. *Curr Opin Struct Biol* **24**, 165-169 (2014). https://doi.org:10.1016/j.sbi.2014.01.002

206 Mirdita, M. *et al.* ColabFold - Making protein folding accessible to all. *bioRxiv*, 2021.2008.2015.456425 (2021). https://doi.org:10.1101/2021.08.15.456425

207 Nguyen, M. C. *et al.* Conformational flexibility of coenzyme A and its impact on the post-translational modification of acyl carrier proteins by 4'-phosphopantetheinyl transferases. *FEBS J* **287**, 4729-4746 (2020). https://doi.org:10.1111/febs.15273

208 Nemeria, N. S. *et al.* Toward an Understanding of the Structural and Mechanistic Aspects of Protein-Protein Interactions in 2-Oxoacid Dehydrogenase Complexes. *Life (Basel)* **11** (2021). https://doi.org:10.3390/life11050407

209 Skerlova, J., Berndtsson, J., Nolte, H., Ott, M. & Stenmark, P. Structure of the native pyruvate dehydrogenase complex reveals the mechanism of substrate insertion. *Nat Commun* **12**, 5277 (2021). https://doi.org:10.1038/s41467-021-25570-y

210 Nagy, B. *et al.* Structure of the dihydrolipoamide succinyltransferase (E2) component of the human alpha-ketoglutarate dehydrogenase complex (hKGDHc) revealed by cryo-EM and cross-linking mass spectrometry: Implications for the overall hKGDHc structure. *Biochim Biophys Acta Gen Subj* **1865**, 129889 (2021). https://doi.org:10.1016/j.bbagen.2021.129889

211 Culka, M. & Rulisek, L. Factors Stabilizing beta-Sheets in Protein Structures from a Quantum-Chemical Perspective. *J Phys Chem B* **123**, 6453-6461 (2019). https://doi.org:10.1021/acs.jpcb.9b04866

212 Mande, S. S., Sarfaty, S., Allen, M. D., Perham, R. N. & Hol, W. G. J. Protein–protein interactions in the pyruvate dehydrogenase multienzyme complex: dihydrolipoamide dehydrogenase complexed with the binding domain of dihydrolipoamide acetyltransferase. *Structure* **4**, 277-286 (1996). https://doi.org:https://doi.org/10.1016/S0969-2126(96)00032-9

213 Argyrou, A., Blanchard, J. S. & Palfey, B. A. The lipoamide dehydrogenase from Mycobacterium tuberculosis permits the direct observation of flavin intermediates in catalysis. *Biochemistry* **41**, 14580-14590 (2002). https://doi.org:10.1021/bi020376k

214 Benen, J. *et al.* Lipoamide dehydrogenase from Azotobacter vinelandii: site-directed mutagenesis of the His450-Glu455 diad. Spectral properties of wild type and mutated enzymes. *Eur J Biochem* **202**, 863-872 (1991). https://doi.org:10.1111/j.1432-1033.1991.tb16444.x

215 Billgren, E. S., Cicchillo, R. M., Nesbitt, N. M. & Booker, S. J. in *Comprehensive Natural Products II* (eds Hung-Wen Liu & Lew Mander) 181-212 (Elsevier, 2010).

216 Schreiber, G. in *Protein–Protein Interaction Regulators* 1-24 (The Royal Society of Chemistry, 2021).

217 De La Cruz, E. M., Wells, A. L., Rosenfeld, S. S., Ostap, E. M. & Sweeney, H. L. The kinetic mechanism of myosin V. *Proc Natl Acad Sci U S A* **96**, 13726-13731 (1999). https://doi.org:10.1073/pnas.96.24.13726

218 Rabut, G., Doye, V. & Ellenberg, J. Mapping the dynamic organization of the nuclear pore complex inside single living cells. *Nat Cell Biol* **6**, 1114-1121 (2004). https://doi.org:10.1038/ncb1184

219 Danson, M. J., Fersht, A. R. & Perham, R. N. Rapid intramolecular coupling of active sites in the pyruvate dehydrogenase complex of Escherichia coli: mechanism for rate enhancement in a multimeric structure. *Proc Natl Acad Sci U S A* **75**, 5386-5390 (1978). https://doi.org:10.1073/pnas.75.11.5386

220 Wu, F. & Minteer, S. Krebs cycle metabolon: structural evidence of substrate channeling revealed by cross-linking and mass spectrometry. *Angew Chem Int Ed Engl* **54**, 1851-1854 (2015). https://doi.org:10.1002/anie.201409336

221 Kyrilis, F. L., Belapure, J. & Kastritis, P. L. Detecting Protein Communities in Native Cell Extracts by Machine Learning: A Structural Biologist's Perspective. *Front Mol Biosci* **8**, 660542 (2021). https://doi.org:10.3389/fmolb.2021.660542

222 Piel, R. B., 3rd, Dailey, H. A., Jr. & Medlock, A. E. The mitochondrial heme metabolon: Insights into the complex(ity) of heme synthesis and distribution. *Mol Genet Metab* **128**, 198-203 (2019). https://doi.org:10.1016/j.ymgme.2019.01.006

223 Djinovic-Carugo, K. & Carugo, O. Missing strings of residues in protein crystal structures. *Intrinsically Disord Proteins* **3**, e1095697 (2015). https://doi.org:10.1080/21690707.2015.1095697

224 Burley, S. K. *et al.* RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res* **49**, D437-D451 (2021). https://doi.org:10.1093/nar/gkaa1038

225 Rollin, J. A. *et al.* High-yield hydrogen production from biomass by in vitro metabolic engineering: Mixed sugars coutilization and kinetic modeling. *Proc Natl Acad Sci U S A* **112**, 4964-4969 (2015). https://doi.org:10.1073/pnas.1417719112

226 Kyrilis, F. L., Belapure, J. & Kastritis, P. L. Detecting Protein Communities in Native Cell Extracts by Machine Learning: A Structural Biologist's Perspective. *Frontiers in Molecular Biosciences* **8** (2021). https://doi.org:10.3389/fmolb.2021.660542

227 Metallo, C. M. & Vander Heiden, M. G. Metabolism strikes back: metabolic flux regulates cell signaling. *Genes Dev* **24**, 2717-2722 (2010). https://doi.org:10.1101/gad.2010510

228 Baracco, E. E. *et al.* alpha-Ketoglutarate inhibits autophagy. *Aging (Albany NY)* **11**, 3418-3431 (2019). https://doi.org:10.18632/aging.102001

229 Jeon, S. M. Regulation and function of AMPK in physiology and diseases. *Exp Mol Med* **48**, e245 (2016). https://doi.org:10.1038/emm.2016.81

230 Ryan, D. G. *et al.* Coupling Krebs cycle metabolites to signalling in immunity and cancer. *Nat Metab* **1**, 16-33 (2019). https://doi.org:10.1038/s42255-018-0014-7

231 Wang, Y. P. & Lei, Q. Y. Metabolite sensing and signaling in cell metabolism. *Signal Transduct Target Ther* **3**, 30 (2018). https://doi.org:10.1038/s41392-018-0024-7

232 van der Lee, R. *et al.* Classification of intrinsically disordered regions and proteins. *Chem Rev* **114**, 6589-6631 (2014). https://doi.org:10.1021/cr400525m

233 Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M. & Uversky, V. N. Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J* **272**, 5129-5148 (2005). https://doi.org:10.1111/j.1742-4658.2005.04948.x

234 Simic, Z., Weiwad, M., Schierhorn, A., Steegborn, C. & Schutkowski, M. The varepsilon-Amino Group of Protein Lysine Residues Is Highly Susceptible to Nonenzymatic Acylation by Several Physiological Acyl-CoA Thioesters. *Chembiochem* **16**, 2337-2347 (2015). https://doi.org:10.1002/cbic.201500364

235 Verbeke, E. J., Mallam, A. L., Drew, K., Marcotte, E. M. & Taylor, D. W. Classification of Single Particles from Human Cell Extract Reveals Distinct Structures. *Cell Rep* **24**, 259-268 e253 (2018). https://doi.org:10.1016/j.celrep.2018.06.022

236 Ho, C. M. *et al.* Bottom-up structural proteomics: cryoEM of protein complexes enriched from the cellular milieu. *Nat Methods* **17**, 79-85 (2020). https://doi.org:10.1038/s41592-019-0637-y

237 Su, C. C. *et al.* A 'Build and Retrieve' methodology to simultaneously solve cryo-EM structures of membrane proteins. *Nat Methods* **18**, 69-75 (2021). https://doi.org:10.1038/s41592-020-01021-2

238 Kastritis, P. L. & Gavin, A. C. Enzymatic complexes across scales. *Essays Biochem* **62**, 501-514 (2018). https://doi.org:10.1042/EBC20180008

239 Bouvette, J. *et al.* Beam image-shift accelerated data acquisition for near-atomic resolution single-particle cryo-electron tomography. *Nat Commun* **12**, 1957 (2021). https://doi.org:10.1038/s41467-021-22251-8

240 Efremov, R. G. & Stroobants, A. Coma-corrected rapid single-particle cryo-EM data collection on the CRYO ARM 300. *Acta Crystallogr D Struct Biol* **77**, 555-564 (2021). https://doi.org:10.1107/S2059798321002151

241 Knoll, M. & Ruska, E. Contribution to geometrical electron optics. *Annalen der Physik* **5**, 607661 (1932).

242 Busch, H. Berechnung der Bahn von Kathodenstrahlen im axialsymmetrischen elektromagnetischen Felde. *Annalen der Physik* **386**, 974-993 (1926). https://doi.org:https://doi.org/10.1002/andp.19263862507

243 Hawkes, P. W. Ernst ruska. *Physics Today* **43**, 84 (1990).

244 Ruska, E., Binnig, G. & Rohrer, H. Nobel lecture. *The development of the electron microscope and of electron microscopy. Nobel Lecture, December* **8** (1986).

245 Williams, D. B. & Carter, C. B. in *Transmission electron microscopy* 3-17 (Springer, 1996).

246 Zuo, J. M. & Spence, J. C. *Advanced transmission electron microscopy*. (Springer, 2017).

247 David, B. W. & Carter, C. B. *Transmission electron microscopy: A textbook for materials science*. (Springer Science+ Business Media, LLC, 1996).

248 Egerton, R., Li, P. & Malac, M. Radiation damage in the TEM and SEM. *Micron* **35**, 399-409 (2004).

249 Henderson, R. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Quarterly reviews of biophysics* **28**, 171-193 (1995).

250 Glaeser, R. M. Limitations to significant information in biological electron microscopy as a result of radiation damage. *Journal of ultrastructure research* **36**, 466-482 (1971).

251 Williams, D. B. & Carter, C. B. in *Transmission Electron Microscopy* 349-366 (Springer, 1996).

252 Reimer, L. & Ross-Messemer, M. Contrast in the electron spectroscopic imaging mode of a TEM: II. Z-ratio, structure-sensitive and phase contrast. *Journal of Microscopy* **159**, 143-160 (1990).

253 Palade, G. E. A study of fixation for electron microscopy. *The Journal of experimental medicine* **95**, 285-298 (1952).

254 Palade, G. E. An electron microscope study of the mitochondrial structure. *Journal of Histochemistry & Cytochemistry* **1**, 188-211 (1953).

255 Talmon, Y. in *Cryotechniques in biological electron microscopy* 64-84 (Springer, 1987).

256 Brenner, S. & Horne, R. A negative staining method for high resolution electron microscopy of viruses. *Biochimica et biophysica acta* **34**, 103-110 (1959).

257 De Carlo, S. & Harris, J. R. Negative staining and cryo-negative staining of macromolecules and viruses for TEM. *Micron* **42**, 117-131 (2011).

258 Dubochet, J. *et al.* Cryo-electron microscopy of vitrified specimens. *Quarterly reviews of biophysics* **21**, 129-228 (1988).

259 McIntosh, J. R. Electron microscopy of cells: a new beginning for a new century. *The Journal of cell biology* **153**, F25-F32 (2001).

260 Adrian, M., Dubochet, J., Lepault, J. & McDowall, A. W. Cryo-electron microscopy of viruses. *Nature* **308**, 32-36 (1984).

261 Parey, K. *et al.* High-resolution cryo-EM structures of respiratory complex I: Mechanism, assembly, and disease. *Science advances* **5**, eaax9484 (2019).

262 Haguenau, F. *et al.* Key events in the history of electron microscopy. *Microscopy and Microanalysis* **9**, 96-138 (2003).

263 Faruqi, A. & Henderson, R. Electronic detectors for electron microscopy. *Current opinion in structural biology* **17**, 549-555 (2007).

264 McMullan, G., Chen, S., Henderson, R. & Faruqi, A. Detective quantum efficiency of electron area detectors in electron microscopy. *Ultramicroscopy* **109**, 1126-1143 (2009).

265 Ruskin, R. S., Yu, Z. & Grigorieff, N. Quantitative characterization of electron detectors for transmission electron microscopy. *Journal of structural biology* **184**, 385-393 (2013).

266 De Rosier, D. J. & Klug, A. Reconstruction of Three Dimensional Structures from Electron Micrographs. *Nature* **217**, 130-134 (1968). https://doi.org:10.1038/217130a0

267 Van Heel, M. & Frank, J. Use of multivariates statistics in analysing the images of biological macromolecules. *Ultramicroscopy* **6**, 187-194 (1981).

268 Henderson, R. *et al.* Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *Journal of molecular biology* **213**, 899-929 (1990).

269 Sigworth, F. J. A maximum-likelihood approach to single-particle image refinement. *Journal of structural biology* **122**, 328-339 (1998).

270 Scheres, S. H. *et al.* Maximum-likelihood multi-reference refinement for electron microscopy images. *Journal of molecular biology* **348**, 139-149 (2005).

271 Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *Journal of structural biology* **180**, 519-530 (2012).

272 Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature methods* **10**, 584-590 (2013).

273 Zivanov, J., Nakane, T. & Scheres, S. H. A Bayesian approach to beam-induced motion correction in cryo-EM single-particle analysis. *IUCrJ* **6**, 5-17 (2019).

274 Zhang, K. Gctf: Real-time CTF determination and correction. *J Struct Biol* **193**, 1-12 (2016). https://doi.org:10.1016/j.jsb.2015.11.003

275 Wang, F. *et al.* DeepPicker: A deep learning approach for fully automated particle picking in cryo-EM. *Journal of structural biology* **195**, 325-336 (2016).

276 Moriya, T. *et al.* Size matters: optimal mask diameter and box size for single-particle cryogenic electron microscopy. *bioRxiv* (2020).

277 Scheres, S. H. & Chen, S. Prevention of overfitting in cryo-EM structure determination. *Nature methods* **9**, 853-854 (2012).

278 Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J Mol Biol* **333**, 721-745 (2003). https://doi.org:10.1016/j.jmb.2003.07.013

279 Schwede, T., Kopp, J., Guex, N. & Peitsch, M. C. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic acids research* **31**, 3381-3385 (2003).

280 Kosinski, J. *et al.* Molecular architecture of the inner ring scaffold of the human nuclear pore complex. *Science* **352**, 363-365 (2016).

281 Fontana, P. *et al.* Structure of cytoplasmic ring of nuclear pore complex by integrative cryo-EM and AlphaFold. *Science* **376**, eabm9326 (2022).
282 Goodfellow, I., Bengio, Y. & Courville, A. *Deep learning*. (MIT press, 2016).
283 Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706-710 (2020).
284 Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *nature* **323**, 533-536 (1986).
285 LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278-2324 (1998).
286 Elman, J. L. Finding structure in time. *Cognitive science* **14**, 179-211 (1990).
287 AlQuraishi, M. Machine learning in protein structure prediction. *Current opinion in chemical biology* **65**, 1-8 (2021).
288 Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins: Structure, Function, and Bioinformatics* **89**, 1607-1617 (2021).
289 Wicky, B. *et al.* Hallucinating symmetric protein assemblies. *Science*, eadd1964 (2022).
290 Lensink, M. F. & Wodak, S. J. Docking, scoring, and affinity prediction in CAPRI. *Proteins* **81**, 2082-2095 (2013). https://doi.org:10.1002/prot.24428
291 van Dijk, A. D., Boelens, R. & Bonvin, A. M. Data-driven docking for the study of biomolecular complexes. *FEBS J* **272**, 293-312 (2005). https://doi.org:10.1111/j.1742-4658.2004.04473.x
292 Dominguez, C., Boelens, R. & Bonvin, A. M. HADDOCK: a protein− protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society* **125**, 1731-1737 (2003).
293 Bonvin, A. M. Flexible protein-protein docking. *Curr Opin Struct Biol* **16**, 194-200 (2006). https://doi.org:10.1016/j.sbi.2006.02.002

7 Appendix

7.1 Theory of methods

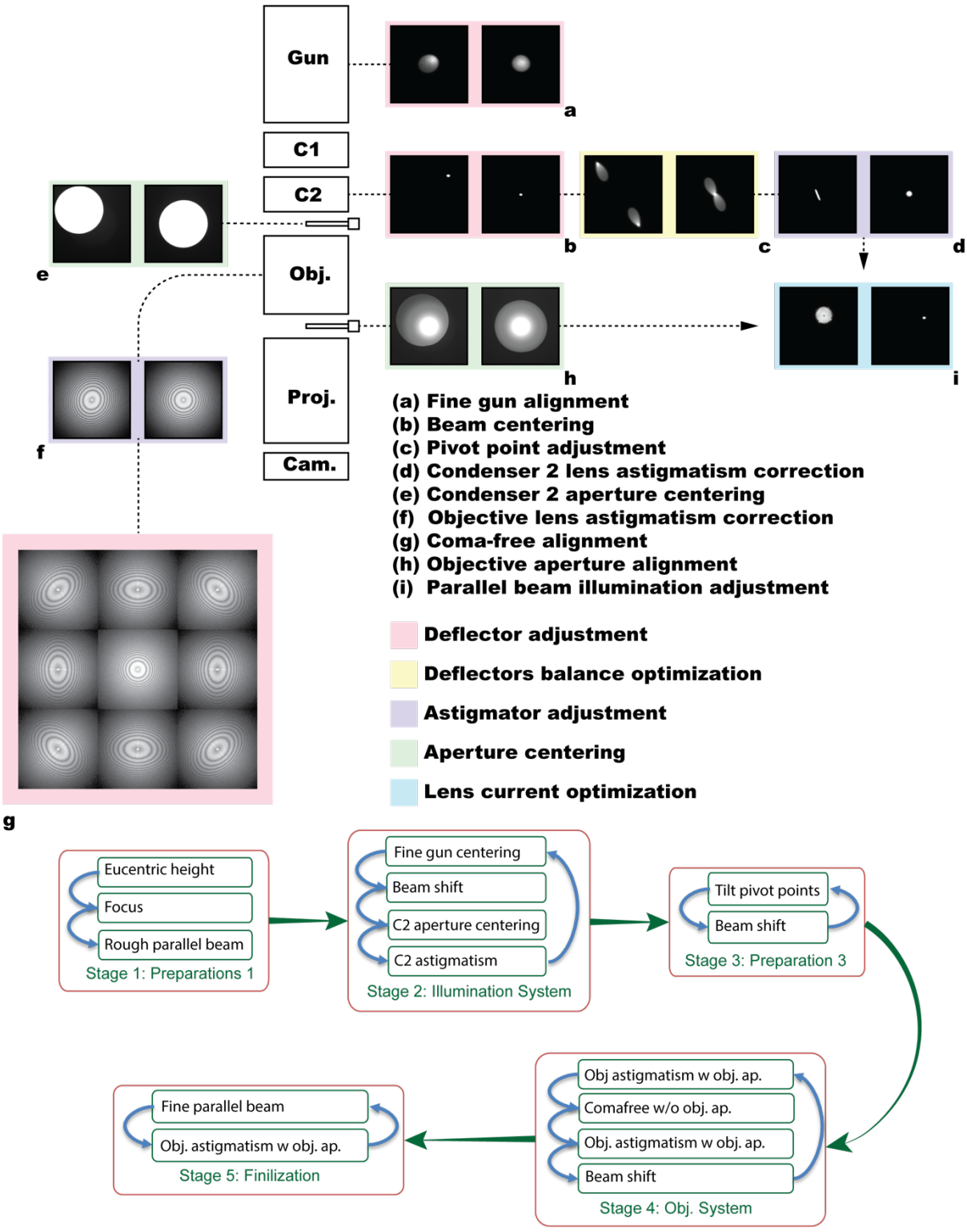
7.1.1 Cryogenic electron microscopy

7.1.1.1 Historical background

Ernst Ruska, a physicist, along with electrical engineer Max Knoll developed in 1931 the first electron microscope[241], with the prototype being able to acquire projection images at 400x magnification. Their work was based on a previous publication by Hans Busch in 1926 where he suggested that an electron beam could be directed by the application of a magnetic field, in an analogous way of how an optical lens system can refract light, and went on to prove his theory by using a cylindrical magnetic lens in order to focus an electron beam to a single point[242]. Just two years later (1933), Ruska developed the first electron microscope that could exceed any resolution reachable by light microscopy. In Ruska's instruments, the electron beam is directed to the sample and the electrons are transmitted through the specimen to create a projection image behind it, leading to the term Transmission Electron Microscopy (TEM)[243]. For his work, Ernst Ruska was awarded the 1986 Nobel Prize in Physics "for his fundamental work in electron optics, and for the design of the first electron microscope"[244].

7.1.1.2 Principles

Despite the numerous advancements in electron microscopy, the basic structure of a transmission electron microscope has mostly remained the same and follows the same principles as an optical microscope[245] (**Appendix Figure 1**).



Appendix Figure 1: General TEM layout and electron beam alignment scheme. Multiple steps have to be followed in order for the beam to have the optimal parameters for cryo-EM data collection. Figure courtesy of Dr. Farzad Hamdi, MLU.

At the top of an electron microscope's column, which is always maintained under high vacuum, lies the electron source. Depending on the type of source, the coherence of the beam changes, and by order of increasing coherence, the following electron sources can be used: (a) a tungsten filament, (b) a lanthanum hexabromide ($LaBr_6$) crystal or a field emission gun (FEG)[246]. In all cases, the source is heated, electrons acquire increased kinetic energy and are released when their kinetic energy surpasses the source's work function. Anode/cathode pairs are employed to first extract the electrons and then accelerate them to the desired voltage. An aperture and condenser magnetic lens follow, increasing the electrons' coherence to form a parallel illumination beam. The beam is then directed to and traverses the specimen under study, interacting with it. Beneath the object lies the objective lens which focuses the beam again. An aperture beneath the objective lens filters out the electrons that are scattered to the outside of the beam and the rest are magnified once again by a projector lens, before colliding with the detector at the end of the column. The detector can be a film, which then evolved to a phosphorous screen connected to a charge coupled device (CCD), and, most recently, direct electron detectors (DED).

Due to the imperfections of electron lenses, a lot of different types of aberrations are introduced to an electron beam (*e.g.,* spherical aberration (C_s), chromatic aberration, astigmatism, coma). These aberrations must be corrected and be accounted for in order to reach optimal imaging conditions during data acquisition on a TEM and obtain the required resolution that will answer a scientific question (**Appendix Figure 1**). Additionally, as mentioned above, after interaction with the specimen, the electrons can be separated into two categories. The elastically scattered electrons and the inelastically scattered electrons[247]. This distinction is critical; thus, it must be elaborated on. During an elastic scattering event, the total momentum and kinetic energy of the beam's electron and the specimen's atom interaction remain unchanged. This means that, because of the large mass difference between the electron and the nucleus of an atom, there is almost no energy transfer, thus the electron contains all the desired "high-resolution" information that is desired for image formation. On the other hand, when a transmitted electron interacts with the electrons of the atom, during collision there is going to be energy transfer and change of the kinetic energy of the transmitted electron, as the objects colliding are of the same mass. These electrons are the inelastically scattered ones and are the main

culprits of radiation damage (or "beam damage") caused to a sample[248-250]. At this point, it is also useful to explain how contrast is generally achieved in an EM image[251]. Elastically scattered electrons and the unscattered electrons have a phase difference that is the result of the different path followed by the first, in relation to the second. The problem is that as biological specimens contain weak-scattering atoms, this phase difference is very small. An additional phase shift can be introduced by interfering with the Fourier pattern in the back-focal plane of the microscope through image distortions. Image distortion can be achieved by increasing the defocus or the aberrations, thus changing the contrast transfer function (CTF). The equation that describes the CTF is:

$$K(R) = \sin\left(\frac{2\pi}{\lambda}\left(\frac{-\Delta\lambda R^2}{2} + \frac{C_s\lambda^3 R^4}{4}\right)\right)$$

R represents the reciprocal space coordinates, C_s the spherical aberration coefficient and Δ the defocus. The equation above shows that contrast is more reliant on defocus, but with increasing defocus applied for EM data collection, a lot of spatial resolution is lost. So, for TEM data collection, a balance has to always be found, between image contrast and information contained in an image[252].

TEM has been critical for many advancements in biological sciences and has been used extensively for this purpose, ever since Palade observed the first images acquired with a TEM of a sectioned cell[253,254]. The fact above is quite contradicting as the conditions for sample imaging are especially detrimental for a biological sample's integrity. More specifically, biological specimens are in principle unable to withstand the high vacuum of the EM column when in an aqueous solution. In addition, the high energy carried by the electron beam causes severe damage to a biological sample (what was described above as beam damage)[255]. This means that in order to image a biological sample, it must be protected from the adverse environmental conditions that it will encounter inside the TEM column. Up until the 1980s, the principal methodology for sample protection was to coat it with a layer of stain, comprised usually of some kind of heavy atom salt (*e.g.,* uranyl acetate)[256]. This coating, apart from protecting the sample, also serves as a way to increase the contrast of the final image formed. The heavy nuclei of the salt increase the amount of elastically scattered electrons that will carry high-resolution information to the detector, forming a negative image and coining the term "negative staining electron microscopy (NS-EM)[257]. In this case, final

resolution is limited by the size of the salt grains themselves that are coating the sample, as in reality, the heavy salt coating and not the specimen is the one being imaged. Another method of sample protection is to encase it in a layer of amorphous ice (or "vitrified" ice). In 1981, Jacques Dubochet, along with his colleague Alistair McDowall, showed, while in EMBL-Heidelberg, that when an aqueous sample solution is deposited on an EM grid and then fast-plunged in liquid ethane, the water is frozen into a form of amorphous ice, vitreous ice[258]. In vitreous ice, even though solidified, the water molecules do not assume a canonical (crystalline) structure, but retain their liquid, amorphous character. Now, with the lack of a staining layer, the biomolecules that are trapped in the vitreous ice can directly interact with the beam and be imaged, while simultaneously being protected and able to withstand radiation exposure for longer periods of time. Since now the sample must be maintained under cryogenic (liquid nitrogen) conditions, in order to preserve the vitreous ice layer, electron microscopy of biological samples evolved into what is now called cryogenic electron microscopy, or cryo-EM.

7.1.1.3 The resolution revolution.

From the early applications of electron microscopy on biological samples, investigators have observed the structures of cell sections[259], organelles[254], viral particles[260] and in the present, atomic resolution reconstructions of proteins and other biomolecular assemblies are possible[261]. The maximum attainable resolution, apart from relying on the optics of a microscope was always heavily reliant on the type of detector used to record the EM images. In the beginning, images were recorded on film, but the process was very lengthy, and in order for the images to be analyzed, the film had to be developed and digitalized before analysis. This meant that until the sample was degraded, there was only a limited time-frame that images could be acquired, limiting the amount of obtainable data. In the 90s, film was replaced with a phosphor screen that was connected to a CCD[262]. Electrons would be converted in the phosphor layer to photons and these would be in turn be detected by the CCD. This development resulted in speeding up and increasing the amount of data acquired, as the elimination of the need for film development and digitalization meant that data could be directly analyzed. Nevertheless, in terms of spatial resolution, CCD cameras

were actually quite behind film[263]. Electrons that entered the phosphor screen before detection by the camera would bounce inside the phosphor layer, losing energy. Additionally, the photon collision incidents that could be detected by a CCD camera were limited, as adjacent pixels would share their charge and increase noise propagation. The summed measure that expresses the influence of optics and electronics of a recording device to a recorded image's signal-to-noise (SNR) ratio is called detective quantum efficiency (DQE). In cryo-EM, DQE is calculated as:

$$DQE(u) = \frac{SNR_{out}^2(u)}{SNR_{in}^2(u)}$$

with (u) being the spatial frequency[264]. For the reasons mentioned above, CCD cameras were quite lacking in terms of DQE when compared to film, but their use prevailed due to their other advantages, mostly detection speed and ease of analysis.

The advent of direct electron detectors (DED) in the late 90s brought about an enormous leap forward for the field, which would be later called "the resolution revolution"[88,89]. DED did not require the conversion of electrons to photons for their detection and recording, and displayed extremely fast readout speeds[265]. The CCD problem of photon co-incidence was also eliminated and now colliding electrons can be accurately localized. This also led to lower electron doses required for sample imaging. Direct electron detectors were instrumental to the rapid advancement of the cryo-EM field, making the dream of atomic resolution reconstructions a reality.

7.1.2 Cryo-EM single particle analysis (SPA)

7.1.2.1 Historical background

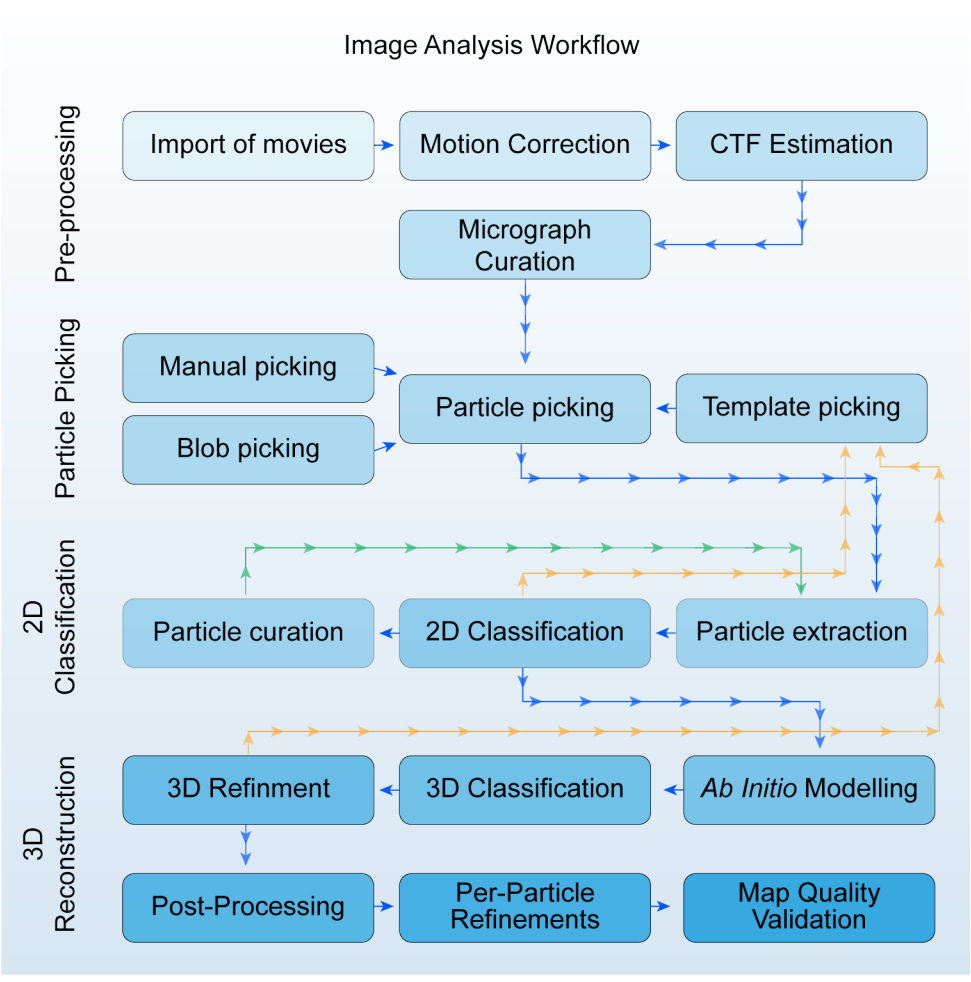
It is evident that advancements in cryo-EM instrumentation should be accompanied by analogous computational methods that would allow for efficient data analysis and interpretation. Ever since the first time that a successful 3D reconstruction of 2D EM projection images from a T4 bacteriophage was presented by DeRosier and Klug in 1968[266], image analysis algorithms have continuously evolved. The first breakthroughs came from the works of Joachim Frank, Marin van Heel and their colleagues where they used cross-correlation algorithms to group 2D

single-particle images in order to improve SNR in cryo-EM data[267]. Spurred by these developments, Richard Henderson and his colleagues finally produced in 1990 the first cryo-EM high-resolution structure of bacteriorhodopsin, showing that given the right microscopy conditions (a stable stage at cryogenic temperatures), cryo-EM image analysis can compensate for beam damage to achieve side-chain resolution for a 3D reconstruction[268]. The next big step in the advancement of cryo-EM image analysis was the introduction of maximum likelihood algorithms for 3D reconstruction of cryo-EM 2D single-particle images, where the works of Sigworth[269] and later of Scheres *et al.*[270] implemented a solution to the alignment and classification of single-particle images. Previously, principal component analysis (PCA) was mostly applied in order classify 2D single-particle images[267]. Each single-particle image could be assigned as a "single experiment" composed by *N* observations, with a PCA analysis trying to explain the differences between them. PCA could perform a classification of the single-particle images by minimizing the dimensions of each group of parameters and then cluster them in the "reduced" space. The maximum likelihood algorithms overcome the problem by aligning the particles and grouping them through an iterative calculation of model parameters that would describe the "single experiment" observations, *i.e.,* the 2D particle images, and then an expectation-maximization algorithm would optimize the overall probability of the model[271]. This way, both 2D classification of single-particles, as well as a 3D reconstruction of projected 2D images can be optimally achieved.

7.1.2.2 Single-particle analysis workflow

The advancements in cryo-EM image processing have nowadays led to a more-or-less standardized workflow that will be performed in order to reach a 3D reconstruction that will accurately recapitulate the raw, single-particle data. The pipeline has come to be known as "Single Particle Analysis (SPA)" and follows the generalized steps described here (**Appendix Figure 2**). The first step is correcting the acquired images (or micrographs) for beam-induced motion[272]. In order to minimize the effect of beam damage visible on the specimen during data acquisition, the electron dose is distributed amongst "frames", and these frames comprise afterwards

the final “movie”, or micrograph, with this technique made possible by the development of the DED described above. The frames capture the motion blurring caused by the interaction of the beam with the specimen under investigation, a movement which is corrected by averaging of the movie’s frames, to a final, single image file. In parallel, the images are also dose-weighted. As the frames continue to be acquired for a final movie generation, the electron dose accumulates, meaning that later frames have higher contrast, but also compounded beam damage. Dose-weighting allows for the exclusion of the most damaged parts of the image and the retaining of the most high-resolution information for all downstream processing steps[273].



Appendix Figure 2: Generalized cryo-EM single particle image analysis workflow

After motion correction and dose-weighting, the CTF parameters of each micrograph are calculated and corrected for, producing the power spectrum of each

micrograph[150]. In the power spectrum, the oscillations of the CTF can be visualized. The center of the power spectrum represents image regions with low resolution information and the edges the maximum spatial frequency that can be calculated for each image and are synonymous with the Nyquist limit[263], a parameter that is directly reliant on the pixel size (or magnification) that the original movie was recorded at (Nyquist limit = 2 x pixel size). The power spectrum is also a useful way to judge micrograph quality. A program can average the Thon rings of the CTF, resulting in a 1D plot of CTF versus spatial frequency. A theoretical model is then fit to this plot and a cross-correlation of the theoretical model versus the 1D plot assesses micrograph quality[274].

The next step is to locate and extract the single-particles from the micrographs. There are multiple strategies for particle picking, either relying on manual selection, or completely automated[275]. The selected single-particles are then cropped out from the micrographs, in a square box with user defined dimensions. The box should be big enough and contain enough of the surrounding solvent space of the particle so that there is enough contribution from the particle's signal, as well as the solvent's noise, making the distinction of downstream algorithms between signal and noise easier. A box should in general be at least 50% larger than the largest diameter of the particle under investigation, not just for SNR optimization, but also because there is a chance that there is signal information contained in the particle image beyond what is optically recognizable, based on CTF delocalization effects[276].

The stack of extracted single-particle images is then subjected to 2D classification[150]. During this process, the particles are classified based on similarity (maximum-likelihood algorithms, described above) in different 2D classes, taking into account each particle image's x and y translations as well as rotations, as well as CTFs. These 2D classes have enhanced features because they will present much higher SNR when compared to single-particle images, because of the averaging that they undergo. This is a very important first step for filtering out class averages that contain "junk" particles, *i.e.,* images of averaged single-particles that have been damaged either during sample preparation or image acquisition. 2D classification can be iteratively performed, with every new round discarding the "junk" particles and keeping only the healthy-looking 2D classes, until a clean set of single-particle images is left.

The cleaned set of single-particles is subsequently used for 3D reconstruction. Each particle image is a projection of the original 3D particle. The issue here is that the orientation of each 2D projection as related to the 3D space is not known beforehand. Sometimes reference models can be used to overcome this limitation by "projection matching", but in most cases this is not possible. The maximum likelihood and expectation-maximization iterative algorithm method described above[150] is also widely applied in the 3D reconstruction step. After each cycle, the new 3D reconstruction is used as a reference for the next cycle (of course the first cycle is always performed with randomized projection assignment) and the algorithm will always try to optimize the projection assignment. As the cycles continue, the projections are assigned with higher accuracy, leading to 3D models containing increased high-resolution information. When the algorithm reaches a point where continuous iterations do not result in significant resolution improvement of the 3D model, the algorithm "converges" and produces the final 3D reconstructed electron density map. This map can then be further optimized by taking into account per-particle parameters, like local motion, CTF, shifts, etc., as well as specific microscope parameters, such as the camera's modular transfer function (MTF) in what is called the "post-processing steps". The final electron map is validated by visual inspection as well as established criteria for cryo-EM model quality, such as the "Gold-standard Fourier Shell Correlation (GS-FSC) criterion"[277,278].

7.1.3 Macromolecular 3D modeling

7.1.3.1 Modeling across resolution scales

Developments in the field of cryo-EM have established it as one of the foremost modern methods of protein 3D structure determination. Nevertheless, no matter how time consuming it may be, obtaining the resulting Coulomb potential map from the acquired data is, as is the case for other structural methods (*i.e.,* XRD, NMR), only a stepping stone towards the end goal. The map needs to be modeled and a 3D protein reconstruction needs to be produced in order for investigators to obtain critical insights into structure and function of proteins. In the case of cryo-EM, as the technique can

be applied for very heterogenous types of samples, this step is not always straightforward. High-resolution EM maps (< 4 Å) can be directly modeled *de novo*, as the resolution is sufficient to observe, apart from secondary structural elements, such as α-helices and β-sheets, backbone placement and even side-chain conformations. This is possible with the use of software suites such as coot[138] and ISOLDE[144], well established software that are being used by the structural biology community for resolving crystallographic structures. Maps of medium to low resolution (> 4 Å and < 8 Å) most of the times are prohibiting for *de novo* modeling. In that case, homology modeling[279] can be applied (or AI-based modeling with, *e.g.,* AlphaFold2, see below), with previous experimentally resolved structures being used as a template, after sequence alignments have identified the similarity between the protein of interest and the target protein. The resulting model can then be fitted in the EM map, judged for quality and corrected. Another interesting aspect of cryo-EM is its ability to obtain low resolution reconstructions (> 8 Å) of biomolecular complexes that are of very low abundance in a heterogenous sample mixture[72]. On the one hand this makes structure determination possible and reveals insights for proteins that cannot be overexpressed and are natively scarce. Additionally, cryo-EM can be used to obtain reconstructions of very large biomolecular complexes, comprised of multiple subunits[280]. On the other hand, this type of resolution results in two significant issues: (1) it makes direct modeling very hard, as little information can be obtained from the EM map apart from overall protein shape and (2) in the case of large protein assemblies, revealing in-complex protein-protein interactions (PPI) is also not feasible.

7.1.3.2 Artificial intelligence in protein structure prediction

For issue (1) that is listed at the end of the previous paragraph **7.1.3.1**, the latest advances in protein structure prediction through the use of artificial intelligence (AI) methods have greatly contributed in determining structures of very flexible, low-abundant, large protein assemblies[281]. This has been mostly attributed to the application machine learning (ML) algorithms, a subcategory of AI, to the problem of protein folding, using as prior knowledge only the protein's primary sequence[226]. Deep learning (DL)[282], a subcategory of ML, is reliant on neural networks, *i.e.,* a network of

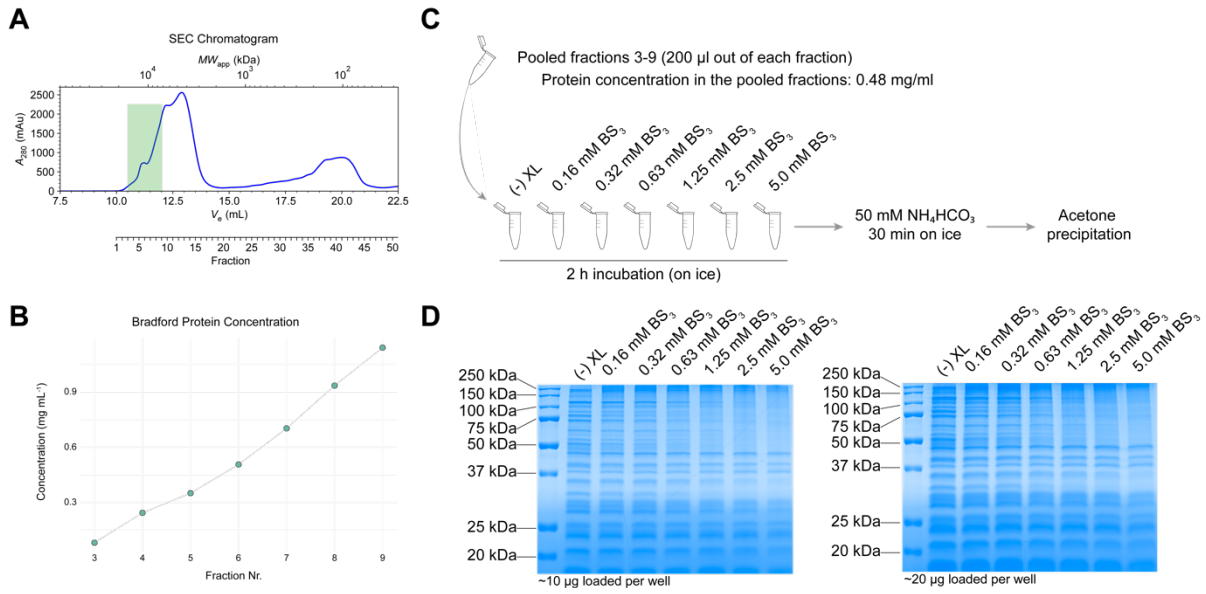
fully connected layers that converts the input sequence into multiple distinct features, that are fed to the next sequential layer, with the features every time used to predict the corresponding output[107,283]. Deep learning models use a variety of types of interconnected layer networks, such as feed-forward neural networks (FFNN)[284], convolutional neural networks (CNN)[285] or recurrent neural networks (RNN)[286]. In all cases, one of the most important parameters that define the effectiveness of a DL model is the training set that has been fed to the model as training in order to recognize and then effectively predict structural features[287]. Modern DL algorithms such as ROSETTAfold[107] and most recently AlphaFold2[106] have been trained with the entirety of the PDB and the totality of genomic data, therefore, being able to additionally infer evolutionary inter-relation across protein sequences. This has enabled such DL approaches to not only perform exceedingly well in competitions that set the standard for protein structure prediction, such as the Critical Assessment of Methods for Protein Prediction (CASP)[288], but also to even transcend single protein structure prediction, extend to multimeric structure prediction[134] and even to the "hallucinating" of novel, unfound in nature, proteins[289], completely *de novo*.

7.1.3.3 Prediction of protein-protein interactions

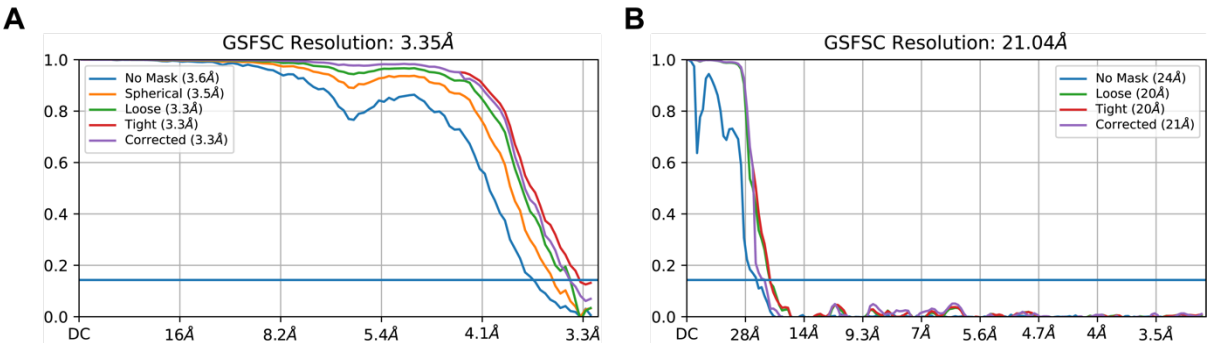
Nevertheless, issue (2) that is listed at the end of **7.1.3.1**, the problem of PPI prediction still remains. DL approaches have tried to tackle it with limited success, but still are mostly unable to outperform algorithms that exploit available experimental, biophysical and biochemical information to drive protein-protein interaction predictions, or "docking". Although, in the near future a specific implementation of DL approaches might even outperform docking approaches for predicting stable interactions, macromolecular interfaces are hypervariable and increasingly complex (*e.g.,* antibody-antigen complexes; interactions of disordered regions; transient, micromolar/millimolar interactions; diffusion-driven biomolecular binding, *etc.*). This is especially visible in the related competition, called Critical Assessment of Predicted Interactions (CAPRI), where all participants are required to predict a biomolecular complex structure, given as initial information only the 3D structure of the complex' participating proteins[290]. Docking algorithms are composed of two major algorithms:

The search function that systematically generates possible interactions of the macromolecules that interact, and the scoring function that systematically ranks the generated solutions with the purpose of ranking the closer-to-native solution at the top. Overall, docking algorithms either predict, purely computationally, the structure of a complex, or use data derived from multiple sources, such as MS, hydrogen-deuterium exchange MS (HDX-MS), XL-MS, NMR, XRD and cryo-EM to derive information about protein surface residues, especially those that are located at the protein's interface that participates in a PPI[291]. This inclusion of experimental data for macromolecular docking was pioneered by HADDOCK[292]. Specifically, in the field experimental information-driven docking of proteins, HADDOCK is an established algorithm that especially excels[143]. HADDOCK stands for High Ambiguity Driven Docking and is able to use all the available experimental information from the sources described above to drive the docking of multiple proteins by translating this information into distance restraints, while simultaneously being able to account for and allow varying degrees of flexibility during the process[293]. The ambiguity part refers mostly to the aforementioned restraints that are imposed on the protein residues participating in the interaction interface and is guided by the experimental data previously provided. Based on the fed information, residues first are recognized as "passive" (not participating in the interface) or "active" (participating in the interface) and then the experimental information is also employed as another variable to the energy function of each active residue. During sampling, HADDOCK tries to find the energy minima of the overall interface interaction (meaning that the PPI is stable). This is achieved by a 3-stage simulation, where, at each stage, increasing amounts of residue flexibility are introduced. In the first stage, residues are always considered as completely rigid, then in the second residue side-chains are allowed more flexibility and finally in the third the complete protein backbone (that participates in the interface) is allowed to move. HADDOCK finishes a run by refining all solutions for a PPI obtained in previous steps in an explicit solvent. All final solutions are then scored and ranked based on energy terms, such as van der Waals interactions, solvation energy contributions, electrostatic interactions, and violations of considered experimental restraints, if any are previously provided.

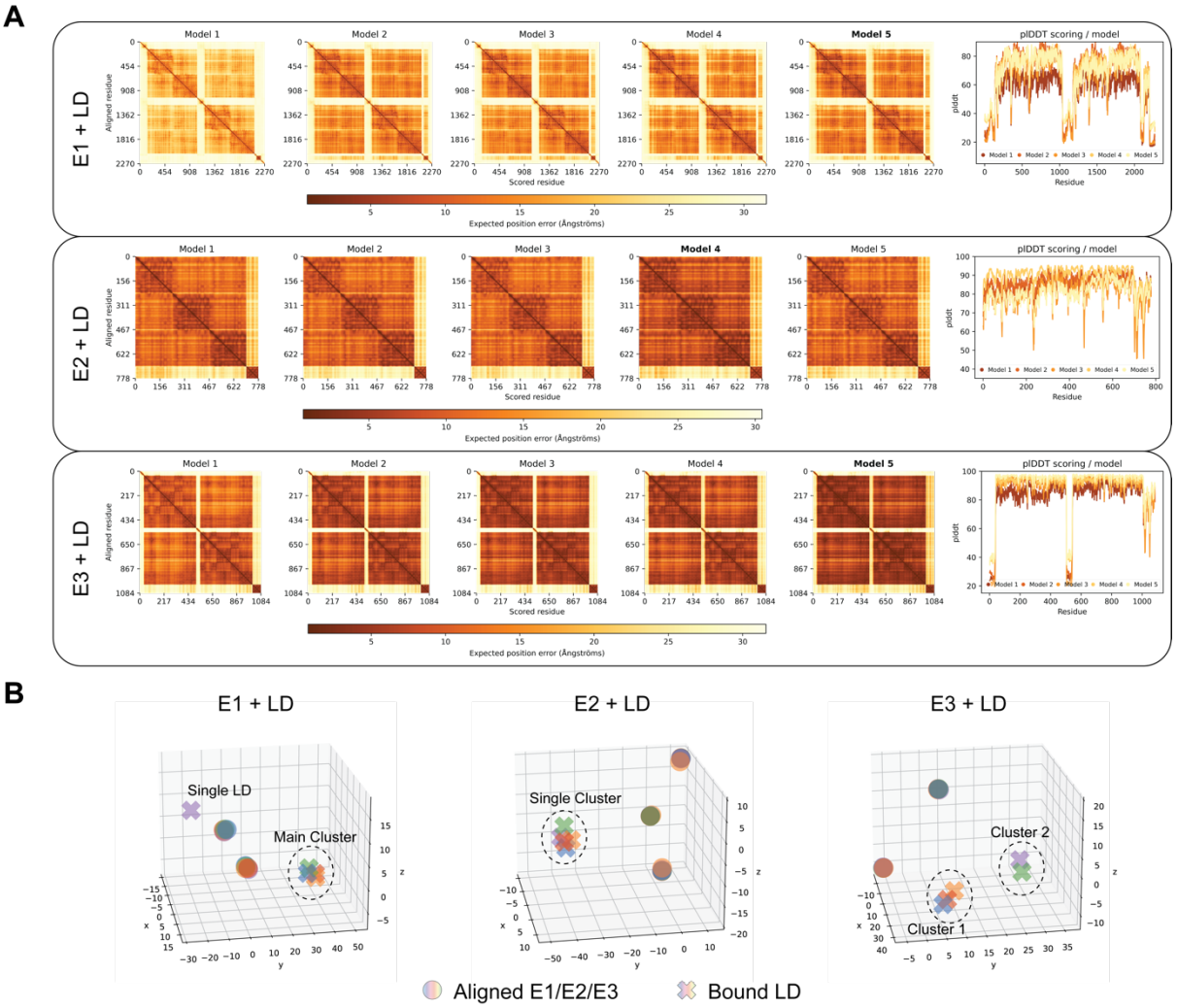
7.2 Supplementary figures



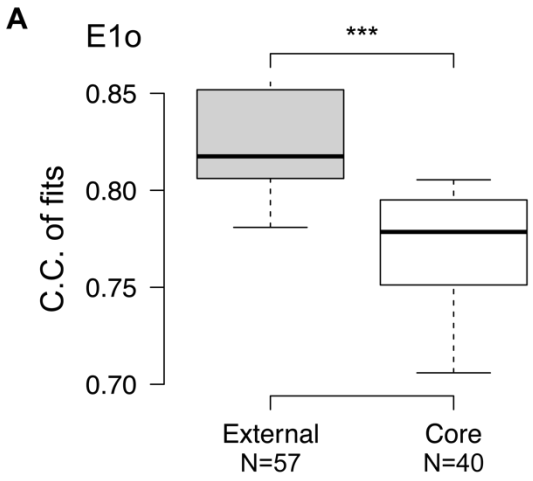
Supplementary Figure 1: Benchmarking process of the applied crosslinker concentration.
(A) The chromatogram of the SEC fracationation performed on the native *C. thermophilum* lysate. The fractions that were pooled for the benchmarking are annotated with a green box. (B) Protein concentration measurements for fractions 3 to 9. (C) Schematic representation of the crosslinking process. (D) SDS-PAGE gels as a readout for the crosslinking concentration (shown at the top). Figure reproduced from[194].



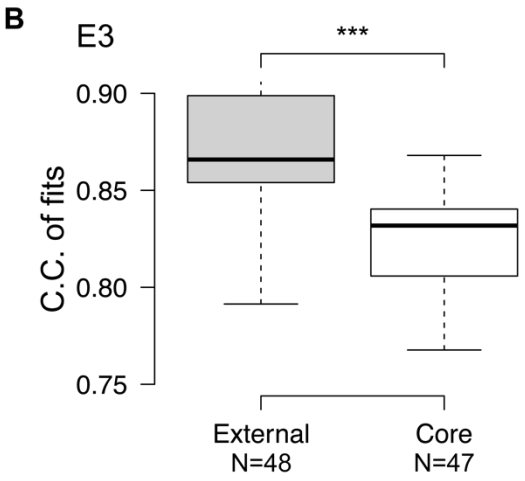
Supplementary Figure 2: Fourier Shell Correlation (FSC) plots for the final OGDHc, cryo-EM reconstructions.
(A) Gold-standard (0.143) FSC plot for the E2o core map. (B) Gold-standard (0.143) FSC plot for the OGDHc map. Figure reproduced from[194].



Supplementary Figure 3: AlphaFold2 model validation.
(A) Positional Alignment Error plots and plDDT scoring for the top 5 models of each AlphaFold-multimer prediction of E1o, E2o and E3 in complex with the E2o LD domain. (B) Spatial plots showing the position of the LD predicted for each solution returned by AlphaFold-multimer. A single cluster is observed for the predicted interaction between E1o-LD and E2o-LD, whereas 2 clusters can be seen for the E3-LD interaction. Figure reproduced from[194].

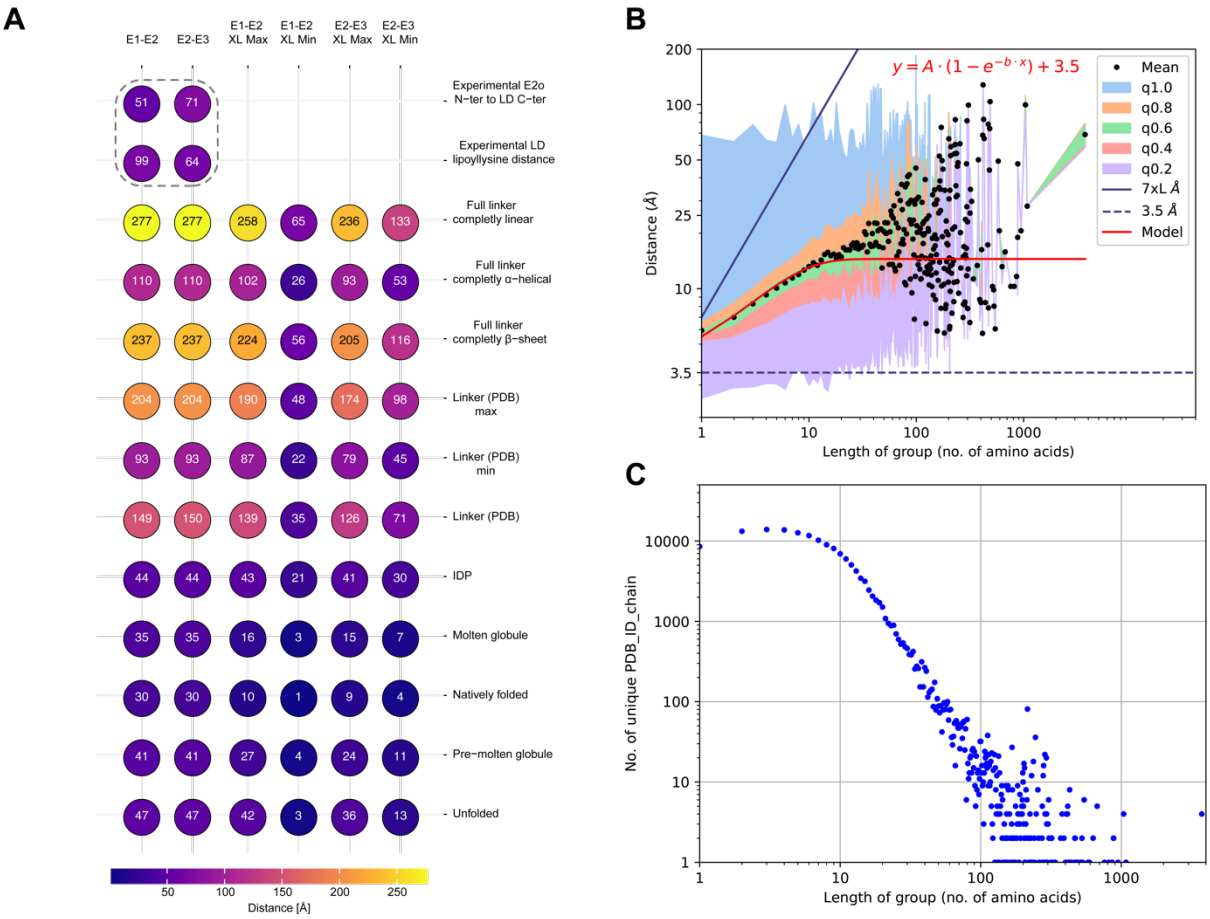


	External C.C.	Core C.C.
Upper whisker	0.88	0.81
3rd quartile	0.85	0.8
Median	0.82	0.78
1st quartile	0.81	0.75
Lower whisker	0.78	0.71

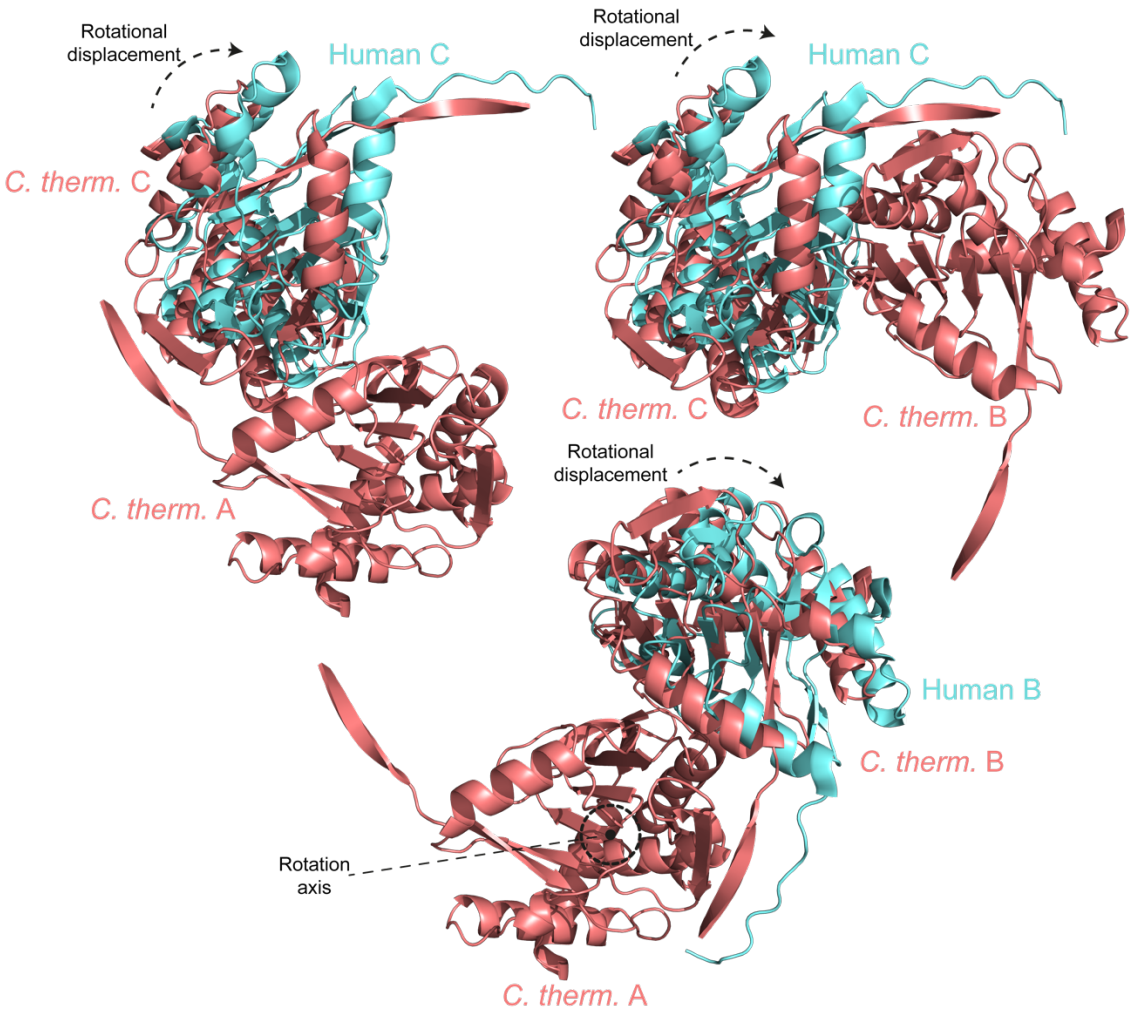


	External C.C.	Core C.C.
Upper whisker	0.92	0.87
3rd quartile	0.9	0.84
Median	0.87	0.83
1st quartile	0.85	0.81
Lower whisker	0.79	0.77

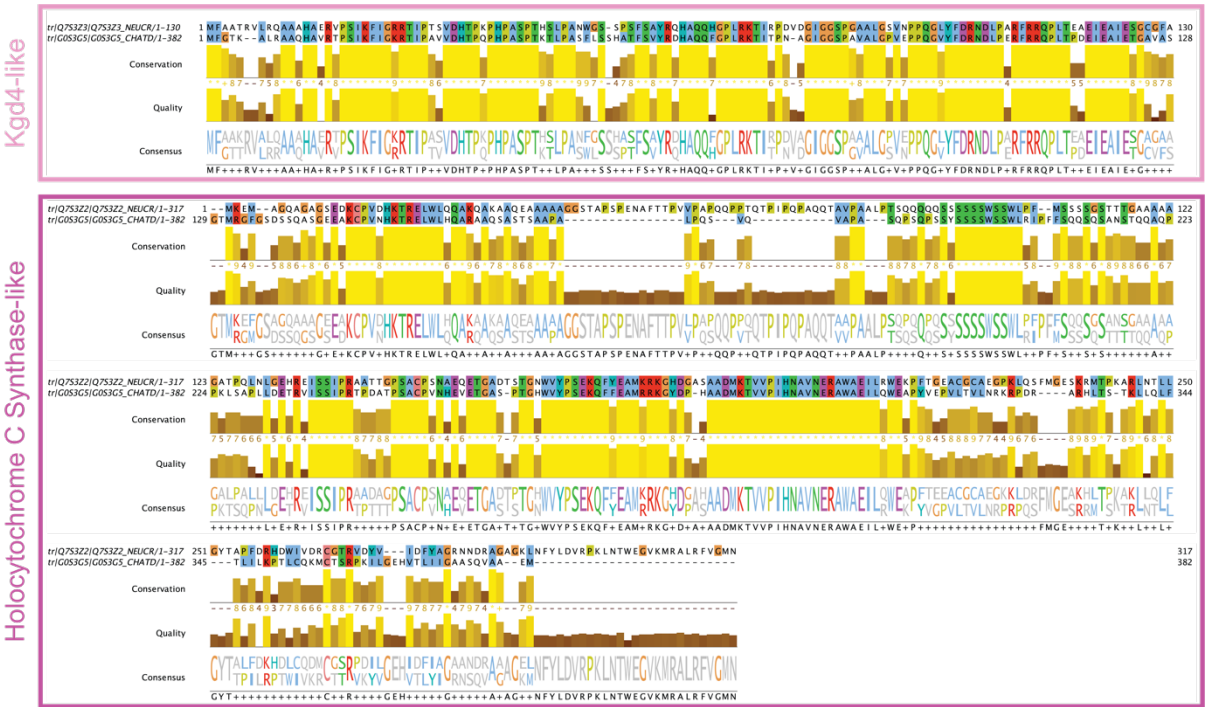
Supplementary Figure 4: Cross-correlation of E1o and E3 models on the peripheral strong densities of the OGDHc complex cryo-EM map.
(A) Boxplot of cross-correlation values of E1o fits in the OGDHc complex map. The E1o dimer fits to the external map densities with statistical significance compared to the internal densities. (B) Boxplot of cross-correlation values of E3 fits in the OGDHc complex map. The E3 dimer fits to the external map densities with statistical significance compared to the internal densities. Figure reproduced from[194].



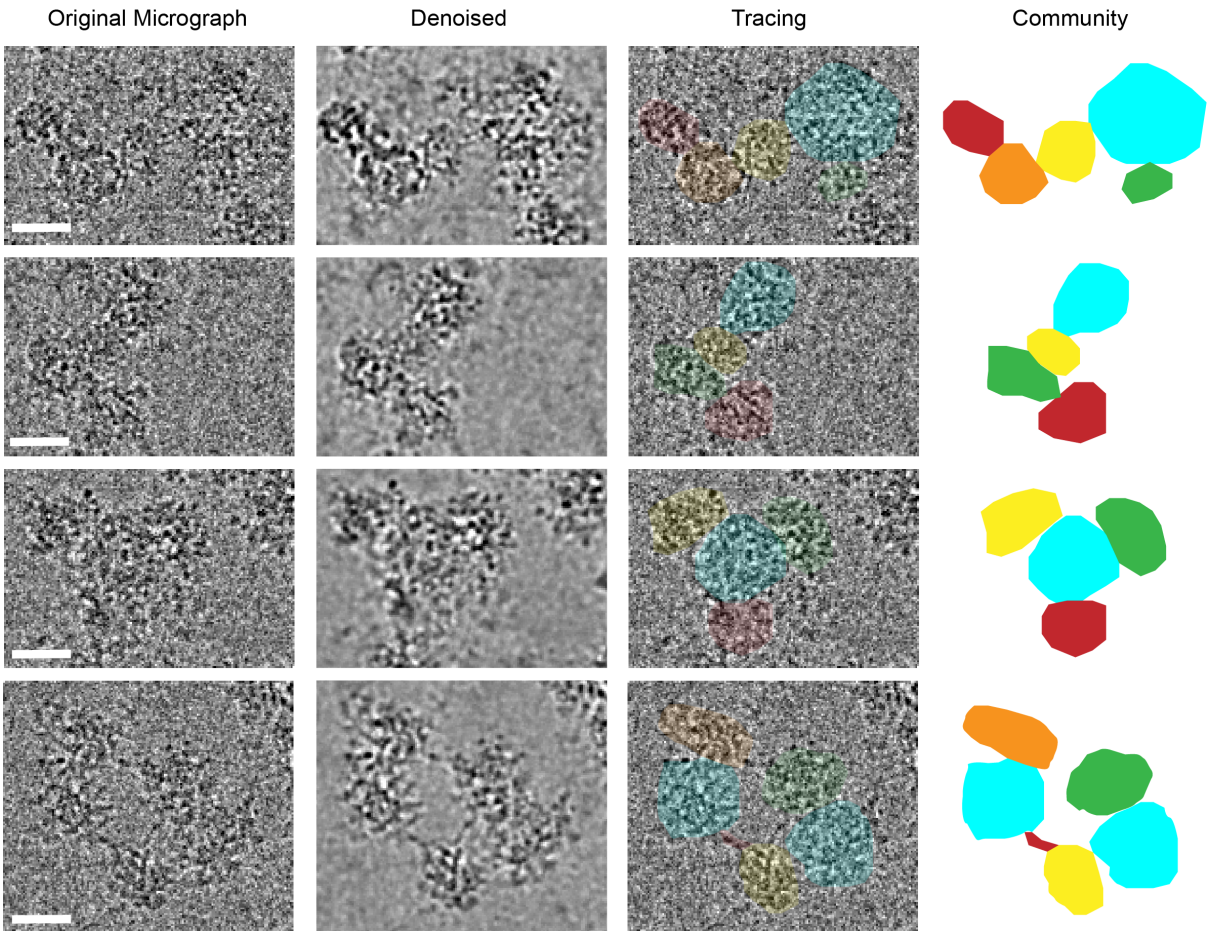
Supplementary Figure 5: Experimental and theoretical linker distances, calculated for the E2o linker region and the PDB deposited data.
(A) Bubble plot of all experimental and theoretical E2o linker distance values. XL-min and XL-max annotate the distance from the first N-ter residue of the E2o that is resolved in the core map to the closest and farthest cross-linked lysine to a peripheral subunit respectively. (B) Graph representing the mean distance values of all unresolved amino-acid sequences belonging to all protein structures deposited in the PDB. Black dots represent the mean value of a length of amino-acids group, with different colors the standard deviation after the integration of each quartile of total data, as denoted in the plot legend. The red line represents a fitted model that describes the relationship between the distance and the length of amino-acid groups. The blue line represents the 2x theoretical distance of an amino-acid sequence, while the dashed blue line represents the theoretical lower limit of any amino-acid sequence. (C) Plot representing the trend between the number of PDB structures incorporating a certain sequence length of unresolved amino-acids. Figure reproduced from[194].



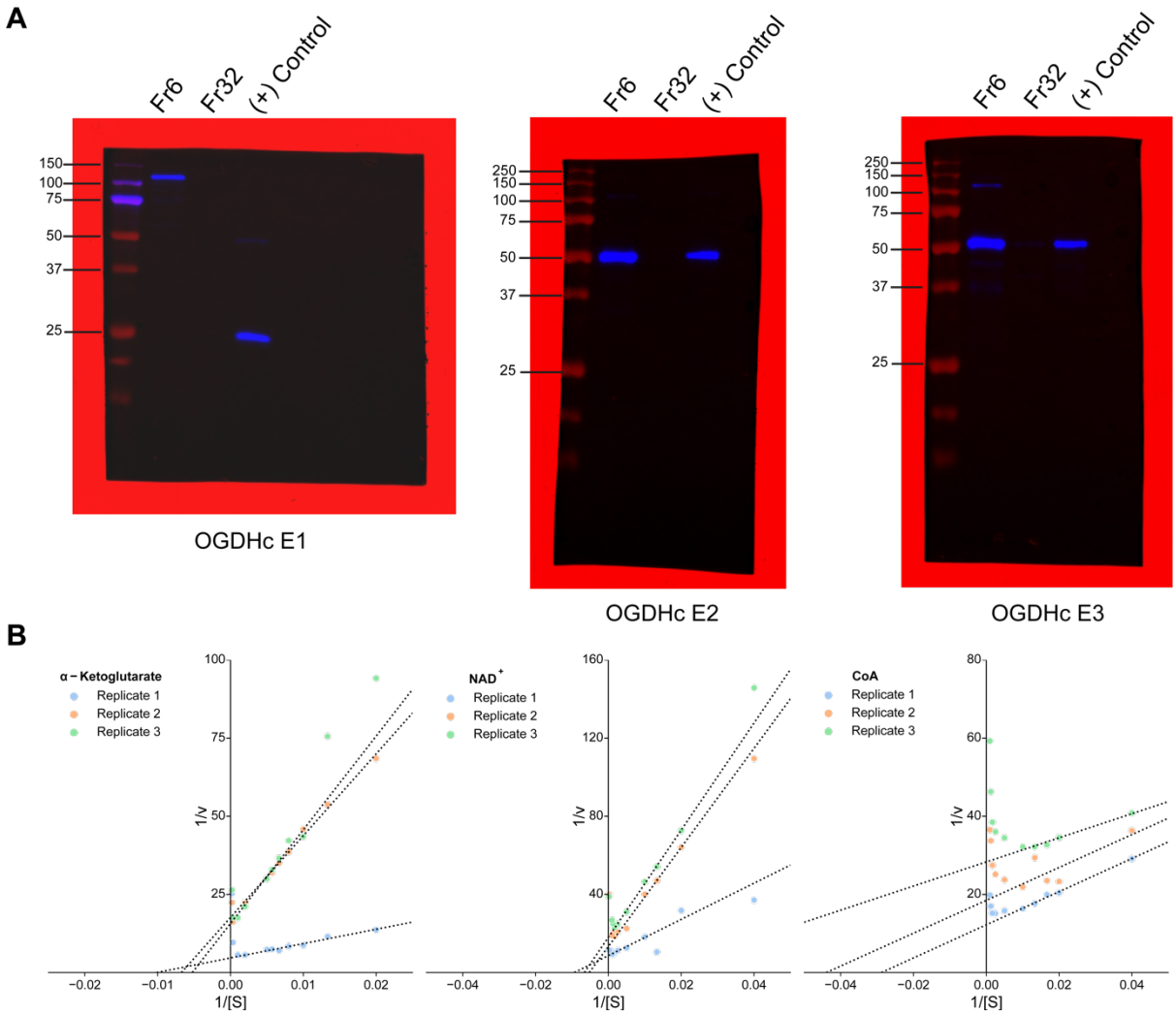
Supplementary Figure 6: Model representation for the rotational displacement of *C. thermophilum* OGDHc E2o core vertex trimer in comparison to the mesophilic human counterpart.
The rotational axis for all calculations was always centered in the previous subunit alignment, starting from the bottom and moving counter-clockwise. Figure reproduced from[194].



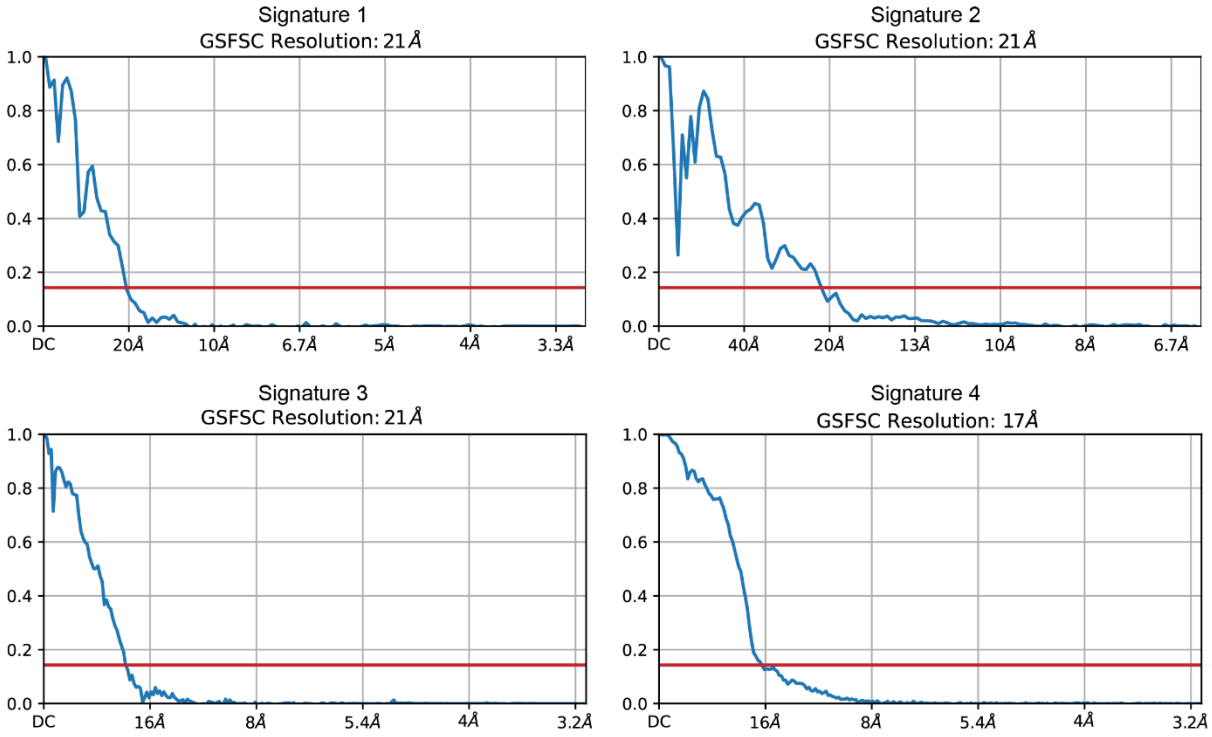
Supplementary Figure 7: Sequence alignment of *N. crassa* KGD4 and HCCS to the *C. thermophilum* putative HCCS-annotated protein sequence.
The *N. crassa* KGD4 protein sequence aligns with high confidence to the first 129 residues of the *C. thermophilum* sequence, while the *N. crassa* HCCS aligns to the rest 254 residues, strongly indicating that the *C. thermophilum* protein sequence is wrongly annotated and is in reality two separate proteins. Figure reproduced from[194].



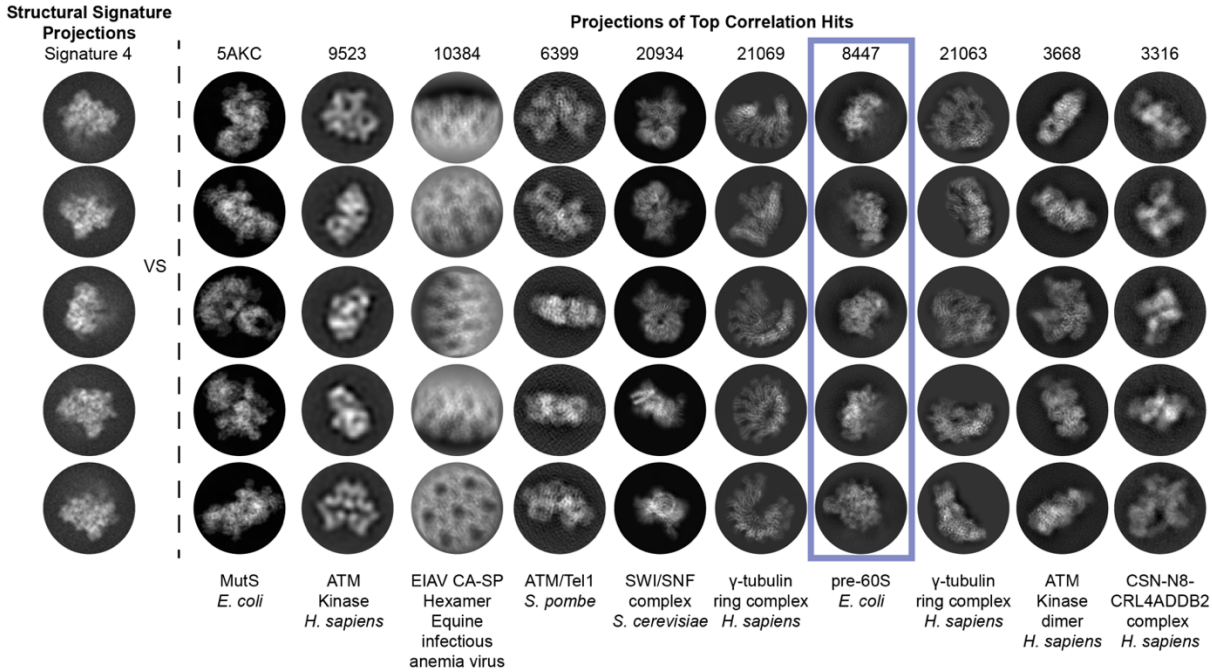
Supplementary Figure 8: Additional protein communities can be detected in-CFS. After denoising, a plethora of stable, endogenous protein communities can be detected in cryo-EM micrographs. Scale bar: 20 nm. Figure reproduced from[123].



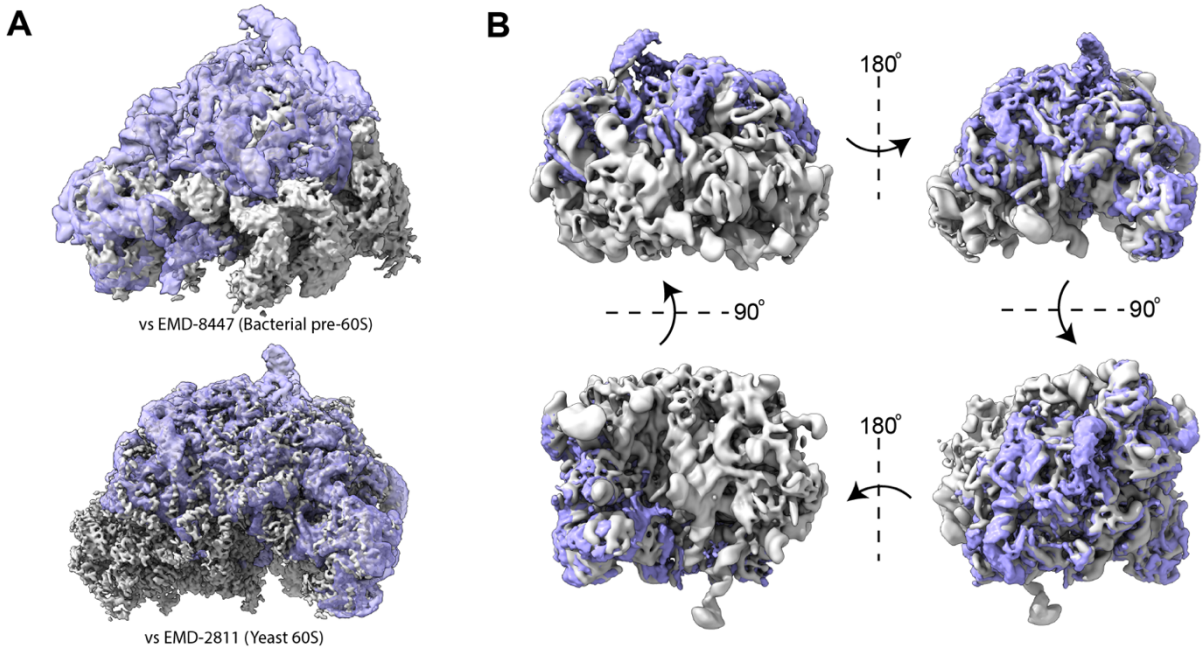
Supplementary Figure 9: Lineweaver-Burk plots and uncropped WB gels for the succinyl-producing CFS biochemical characterization.
(A) Un-cropped WB gels for the identification of each of the OGDHc components. (B) Lineweaver-Burk plots for each of the substrates that are involved in the OGDHc reaction, showing the values for the 3 biological replicates for each. Figure reproduced from[194].



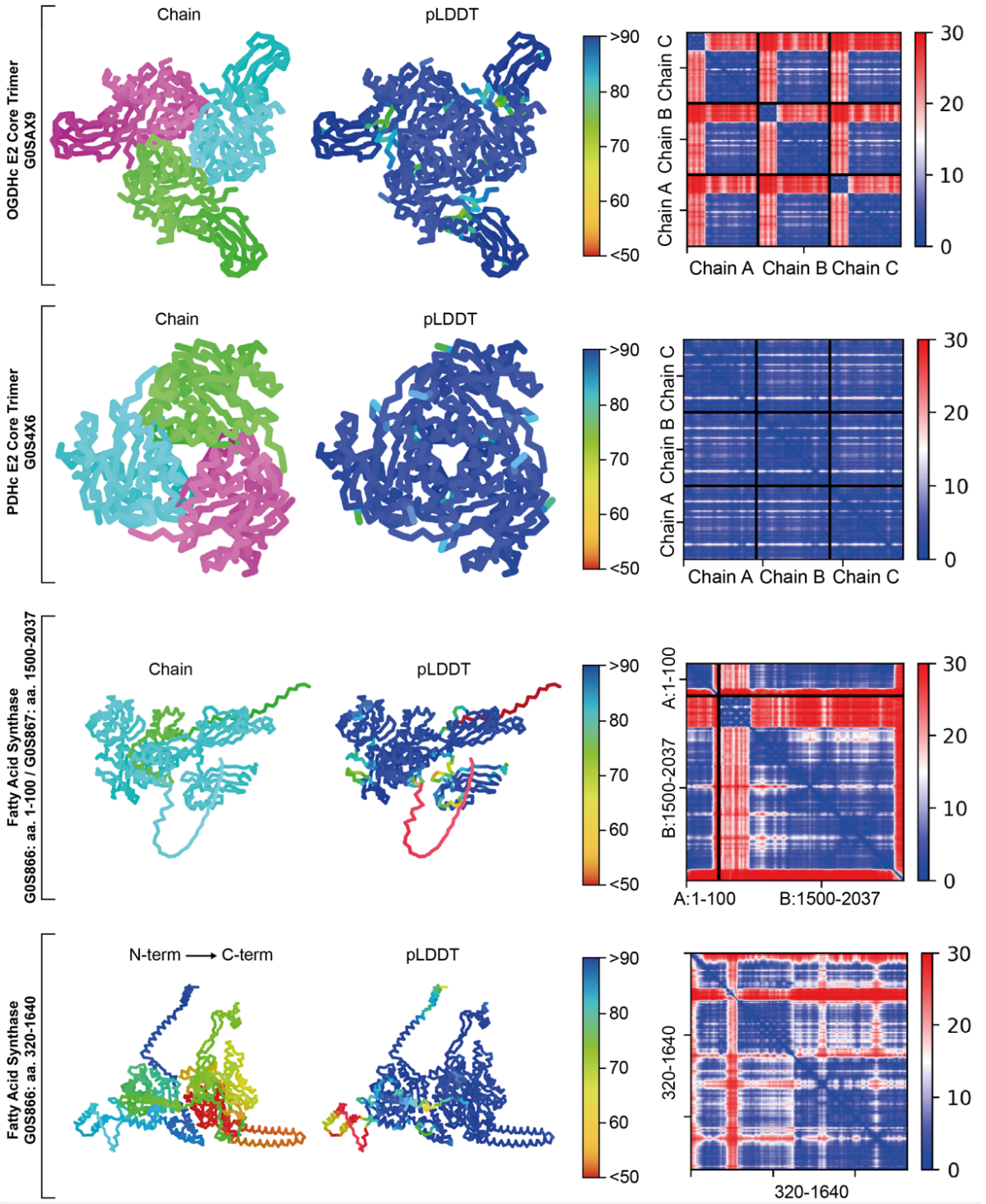
Supplementary Figure 10: FSC plots for ab-initio signature reconstructions.
Despite the overall low resolution, the reconstructed ab-initio signature maps can be used for signature identification. Resolutions reported are at FSC = 0.143. Figure reproduced from[123].



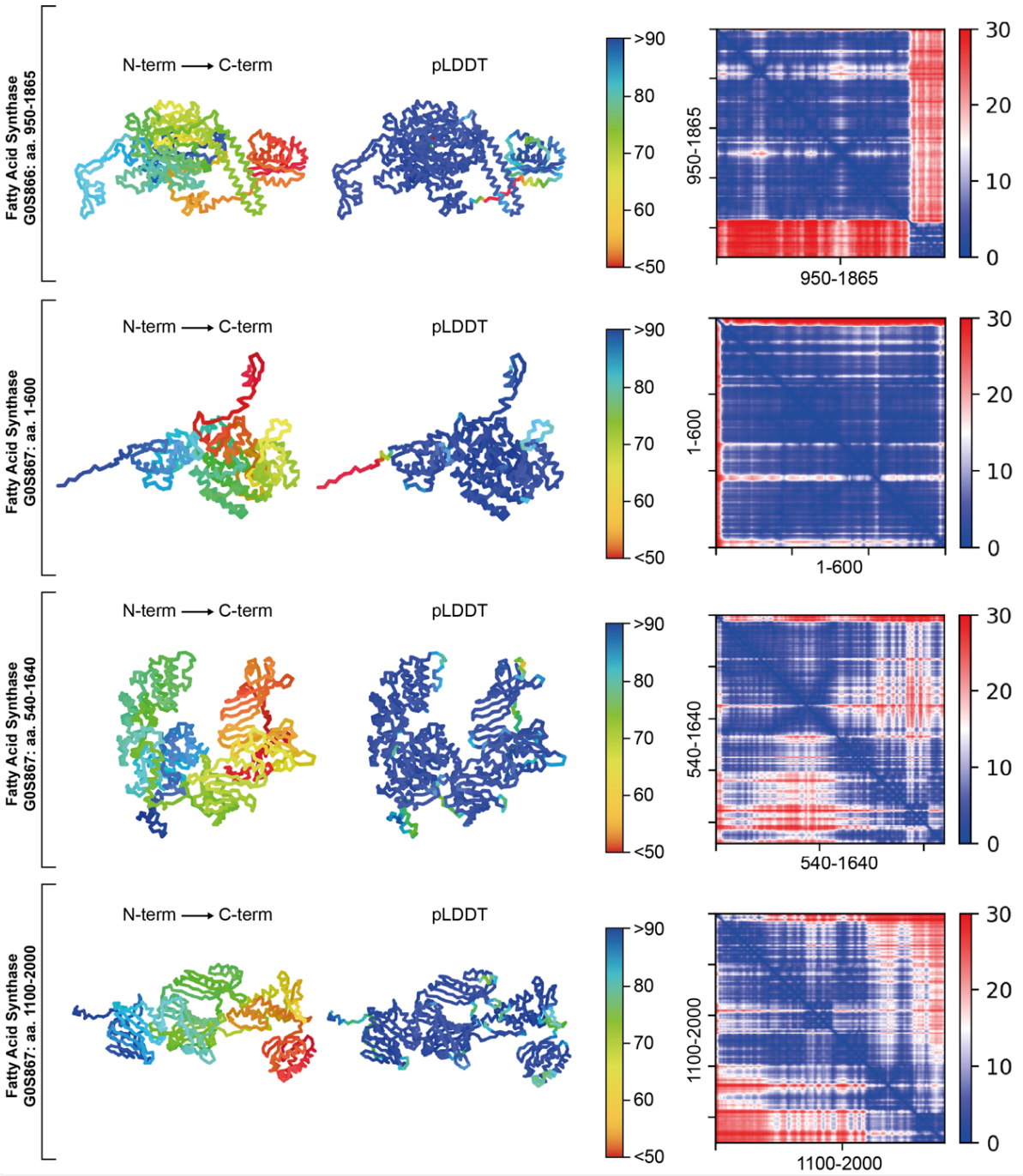
Supplementary Figure 11: Visual comparison of Omokage top-10 hits and identification of Signature 4.
After projecting the reconstructed ab-initio Signature 4 map and all top-10 hits of the Omokage search results in 2D, visual identification is possible. Figure reproduced from[123].



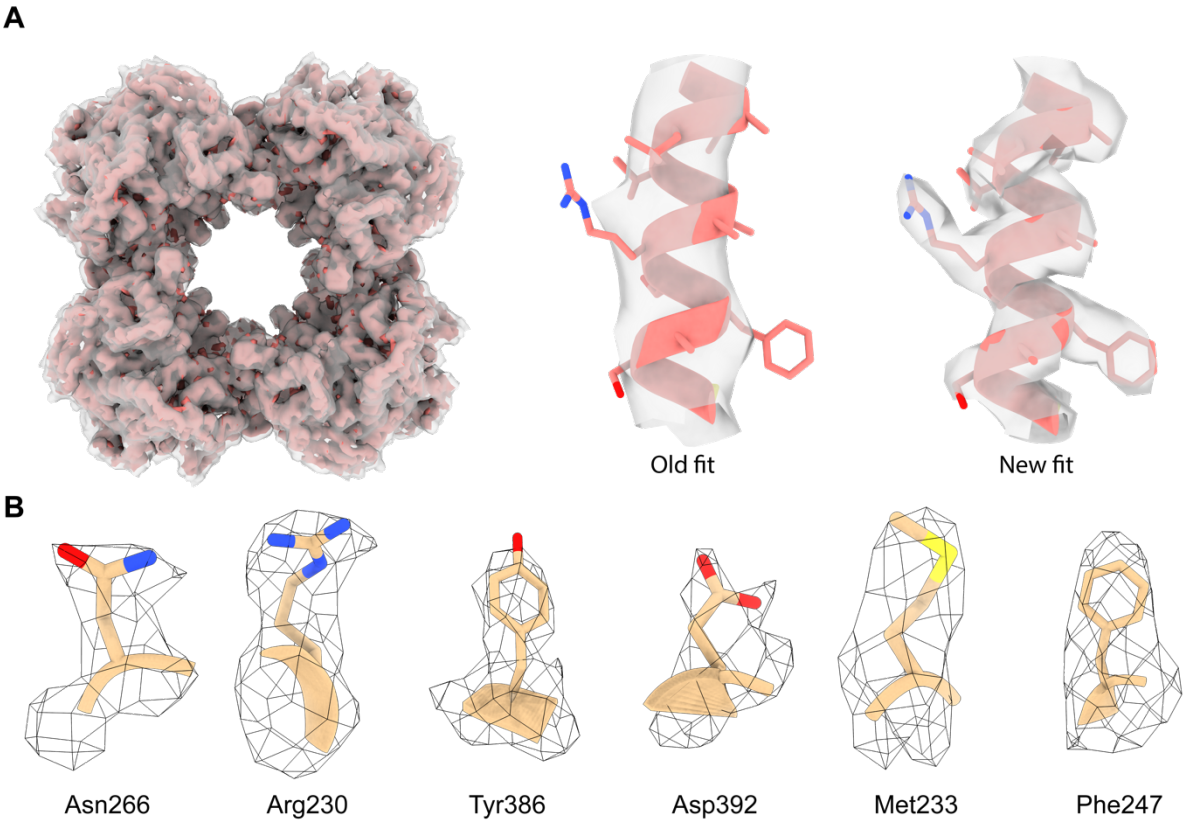
Supplementary Figure 12: Additional fits of endogenous pre-60S ribosomal subunit. (A) The final reconstructed map is fit and compared to a bacterial pre-60S and a complete yeast 60S subunit. (B) Further visual inspection of the fit between the endogenous pre-60S and the lowpass filtered yeast 60S ribosomal subunits reveals the low map coverage at the early assembly stages. Figure reproduced from[123].



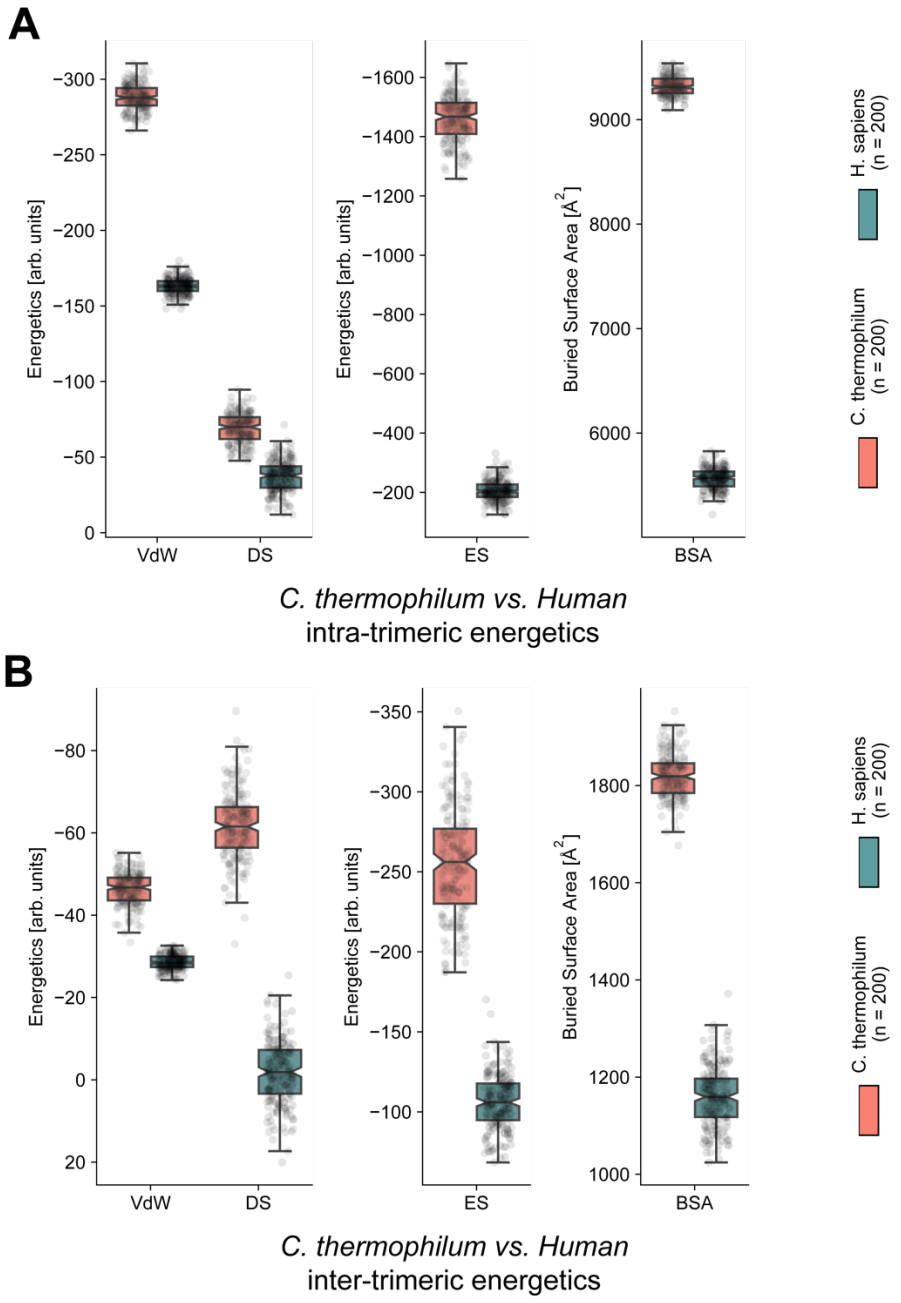
Supplementary Figure 13: Quality estimation for AlphaFold/ColabFold predicted models after exploratory protein community member analysis of the CFS.
Every row represents a different prediction and shows: an overall representation of the mainchain-predicted model (left) colored either by different chain in the case of multimers or by N-ter (blue) to C-ter (red) rainbow coloring; a model representation that is colored-coded (middle) according to the local confidence estimated with the predicted local distance difference test (pLDDT), AlphaFold2's main validation tool, where dark blue (>90) indicates high confidence for the predicted backbone and side-chain rotamers and green (>70) indicates a confident prediction only for the backbone; interchain and interdomain Predicted Aligned Error (PAE, right), which are calculated based on estimated distance error (in Å), between all different residue pairs in a complex, with blue annotating a low expected error. Figure reproduced from[123].



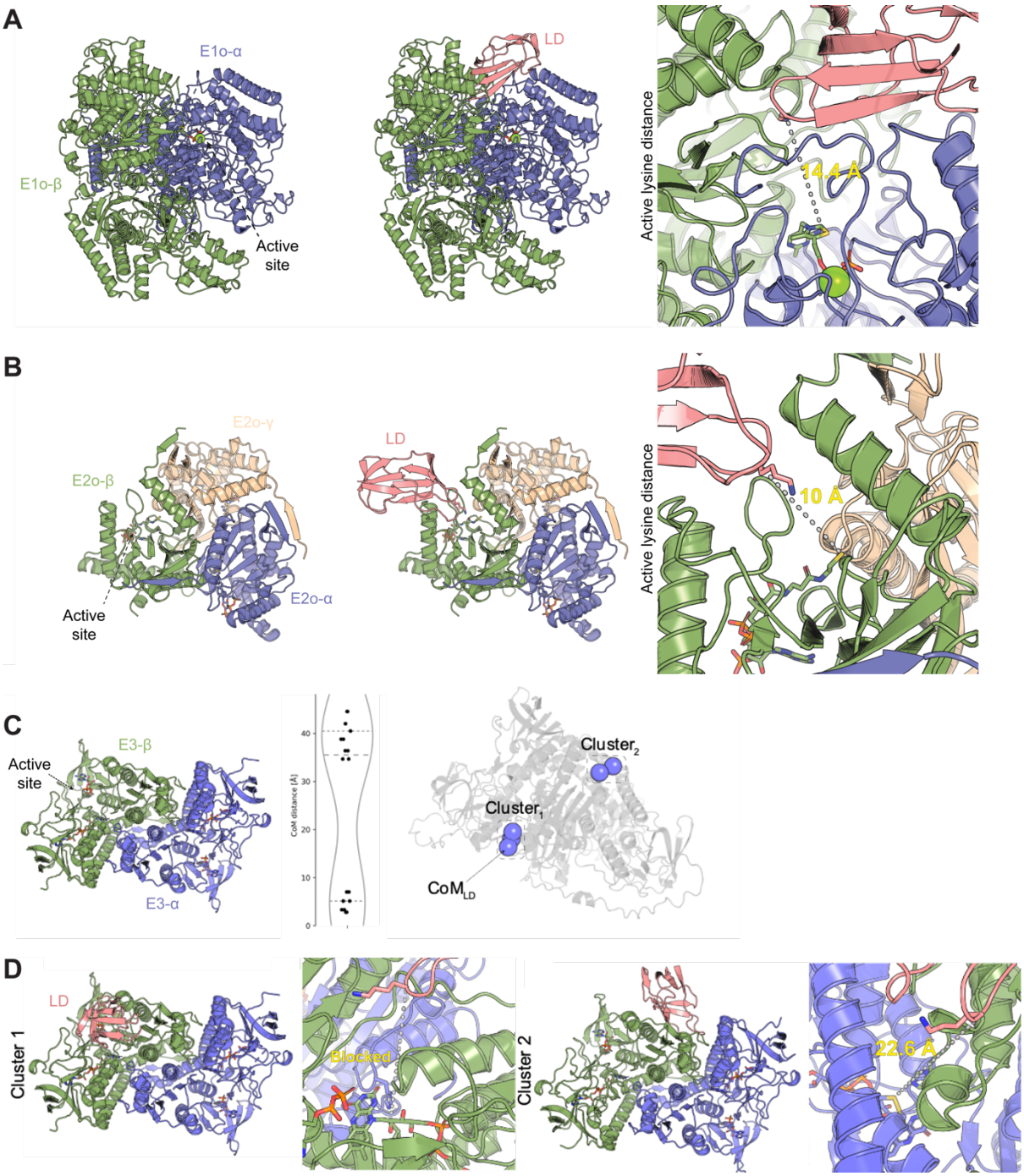
Supplementary Figure 14: Quality estimation for AlphaFold/ColabFold predicted models after exploratory protein community member analysis of the CFS (cont.). Every row represents a different prediction and shows: an overall representation of the mainchain-predicted model (left) colored either by different chain in the case of multimers or by N-ter (blue) to C-ter (red) rainbow coloring; a model representation that is colored-coded (middle) according to the local confidence estimated with the predicted local distance difference test (pLDDT), AlphaFold2's main validation tool, where dark blue (>90) indicates high confidence for the predicted backbone and side-chain rotamers and green (>70) indicates a confident prediction only for the backbone; interchain and interdomain Predicted Aligned Error (PAE, right), which are calculated based on estimated distance error (in Å), between all different residue pairs in a complex, with blue annotating a low expected error. Figure reproduced from[123].



Supplementary Figure 15: Improved resolution examples for the in-CFS cryo-EM reconstruction of the OGDHc E2o core.
(A) Comparison between the final CFS-derived and the initial, exploratory OGDHc E2o core map. On the left, the final CFS-derived E2o core map (salmon) is fit into the initially derived E2o core map (gray). On the right, the marked improvement in side-chain density coverage between the old and new maps can be observed. (B) Various examples taken from the final, cryo-EM resolved, fitted model, showing confident placement of side-chains in their corresponding densities. Figure reproduced from[194].

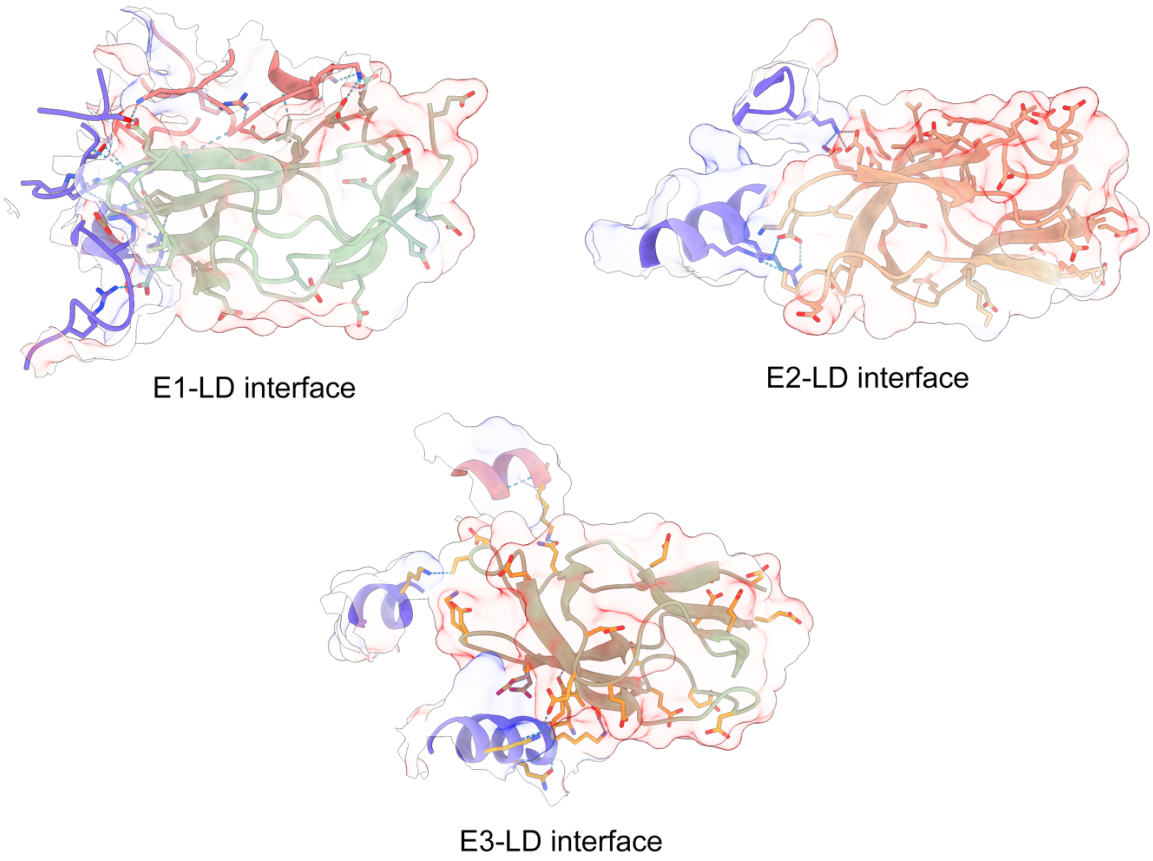


Supplementary Figure 16: Comparative energetics between *C. thermophilum* E2o interfaces and its mesophilic human counterpart.
(A) *C. thermophilum* and Human E2o core intra-trimeric energetics display stronger forces that contribute to the *C. thermophilum* E2o core compaction. (B) *C. thermophilum* and Human E2o core inter-trimeric energetics display stronger forces that contribute to the *C. thermophilum* E2o core compaction. Figure reproduced from[194].



Supplementary Figure 17: AlphaFold2 predicted models of the OGDHc subunits in complex with the E2o LD domain.
(A) AlphaFold2 predicts the LD as bound in the dimeric interface. The lipoylated lysine is in close distance (14.4 Å) to the C2 atom of the ThDP, which is succinylated during decarboxylation of α-ketoglutarate in the first step of the reaction of OGHDc. The binding cavity of the generated AlphaFold2 model is not sterically blocked, indicating a plausible docking solution. (B) In each dimeric interaction interface in these trimeric E2o building blocks, a CoA binding site is present. Clustering of all AF solutions showed all LDs within a distance range of 10 Å, indicating a single prediction solution. The LD domain is bound to a monomeric E2, meaning that theoretically 24 LD domains could be bound by the OGHDc core simultaneously. The localization of the lipoyl-¬lysine is close to the CoA binding site, with a distance between the Cα atom of the lysine and the thiol group of CoA, where the

succinate from the lipoate is transferred to, of 13.7 Å. The binding cavity is accessible, indicating, again, a plausible docking solution. (C) In the predicted structures of E3 with bound LD, there are two clear cluster (n = 3 and n = 2), with clear differences in the localization of the LD. (D) Inspecting the two different clusters of the bound LD, in the first cluster, the LD is located at a monomer near the NAD+ binding site, whereas in the second cluster, the LD is bound in the dimeric interface. Mapping the reaction path, in cluster 1, the reaction path is blocked by the FAD, whereas in the second cluster, the lipoylated lysine are in reasonable distance to the disulfide bond in the active side, with a distance of 22.6 Å. Figure reproduced from[194].



Supplementary Figure 18: Interfaces of LD with all OGDHc components.
Visualization of residues participating in the interface between the LD and E1o, E2o and E3 respectively. The high number of negatively charged side-chains on the LD surface is visible, along with the hydrogen-bond network that helps stabilize the highly electrostatically-driven interaction. Figure reproduced from[194].

7.3 Supplementary tables

Supplementary Table 1: All values plotted for the kinetic characterization of the OGDHc component reactions. Included in the accompanying CD.

Supplementary Table 2: All values related to XL-MS, MS identification, in-fraction community annotation and stoichiometric calculations. Included in the accompanying CD.

Supplementary Table 3: Signature 1 (Hybrid OGDHc/BCKDHc), Signature 2 (PDHc), Signature 3 (FAS) and Signature 4 (pre-60S ribosome) acquisition and reconstruction parameters.

	Signature 1 (OGDHc E2 core) (EMD-13844) (PDB 7Q5Q)	Signature 2 (PDHc E2 core) (EMD-13845) (PDB 7Q5R)	Signature 3 (Fatty Acid Synthase) (EMDB-13846) (PDB 7Q5S)	Signature 4 (pre-60S Ribosomal Subunit) (EMDB-13093)
Data collection and processing				
Magnification	92000X	92000X	92000X	92000X
Voltage (kV)	200	200	200	200
Microscope model	TFS Glacios	TFS Glacios	TFS Glacios	TFS Glacios
Camera model	TFS Falcon IIIEC	TFS Falcon IIIEC	TFS Falcon IIIEC	TFS Falcon IIIEC
Number of frames	13	13	13	13
Electron exposure (e^{-}/Å^2)	30	30	30	30
Per-frame exposure (e^{-}/Å^2)	2.3	2.3	2.3	2.3
Defocus range (μm)	-0.6 to -2.0	-0.6 to -2.0	-0.6 to -2.0	-0.6 to -2.0
Pixel size (Å/px)	1.568	1.568	1.568	1.568
Images acquired (no.)	2808	2808	2808	2808
Acquisition software	TFS EPU 2	TFS EPU 2	TFS EPU 2	TFS EPU 2
Symmetry imposed	O	I	D3	-
Initial particle images (no.)	276,339	276,339	276,339	276,339
Final particle images (no.)	1,819	7,825	5,231	35,773
Map resolution (Å)	4.38	3.84	4.47	4.52

FSC threshold	0.143	0.143	0.143	0.143
Map B-factor	188.3	198.2	165.9	176.1
Refinement				
Initial model used (PDB code)	-	-	-	
Map sharpening *B* factor ($Å^2$)	0 (not modified)	0 (not modified)	0 (not modified)	
Model composition				
Non-hydrogen atoms	42,336	95,520	168,630	
Protein residues	5,424	12,480	21,450	
B factors ($Å^2$)				
Protein	199	117	245	
R.m.s. deviations				
Bond lengths (Å)	0.01	0.01	0.01	
Bond angles (°)	1.43	0.93	0.90	
Validation				
MolProbity score	1.46	1.22	1.18	
Clashscore	3.0	4.4	2.5	
Poor rotamers (%)	2.1	0.6	0.7	
Ramachandran plot				
Favored (%)	97.4	98.1	97.3	
Allowed (%)	2.6	1.9	2.6	
Disallowed (%)	0.0	0.0	0.1	
Map-CC	0.89	0.83	0.79	

Supplementary Table 4: OGDHc E2o core and complex acquisition and reconstruction parameters.

	OGDHc E2o core	OGDH complex
Data collection and processing		
Magnification	92000X	92000X
Voltage (kV)	200	200
Microscope model	TFS Glacios	TFS Glacios
Camera model	TFS Falcon IIIEC	TFS Falcon IIIEC
Number of frames	13	13
Electron exposure (e^{-}/Å^2)	30	30
Per-frame exposure (e^{-}/Å^2)	2.3	2.3
Defocus range (μm)	-0.6 to -2.0	-0.6 to -2.0
Pixel size (Å/px)	1.568	1.568
Images acquired (no.)	25803	25803
Acquisition software	TFS EPU 2	TFS EPU 2
Symmetry imposed	O	C1
Initial particle images (no.)	3,596,302	2,891,518
Final particle images (no.)	52,034	5,178
Map resolution (Å)	3.35	21.04
FSC threshold	0.143	0.143
Map B-factor	176.9	-
Refinement		
Initial model used (PDB code)	7Q5Q	
Map sharpening *B* factor (Å^2)	0 (not modified)	
Model composition		
Non-hydrogen atoms	1812	
Protein residues	233	
B factors (Å^2)		
Protein (min/max/mean)	7.48/59.64/22.96	

R.m.s. deviations	
Bond lengths (Å)	0.005
Bond angles (°)	1.134
Validation	
MolProbity score	2.61
Clashscore	16.72
Poor rotamers (%)	3.05
Ramachandran plot	
Favored (%)	90.91
Allowed (%)	9.09
Disallowed (%)	0.00
Map-CC	0.79

Supplementary Table 5: All values reported in the HADDOCK scoring plots. Included in the accompanying CD.

Supplementary Table 6: All residue frequencies of residues participating in the interface between E1o, E3 and the E2o LD domain. Included in the accompanying CD.

Supplementary Table 7: All values and statistics related to OGDHc peripheral subunit fits and distance calculations. Included in the accompanying CD.

Supplementary Table 8: All values related to plots of unresolved amino-acid sequences in all PDB entries. Included in the accompanying CD.

7.4 Supplementary material

Supplementary Material 1: List of all PDB IDs that were used for the flexible linker analysis. Included in the accompanying CD.

Supplementary Material 2: Cytoscape session for the visualization of all protein community networks listed in Table S3. Included in the accompanying CD.

8 Acknowledgements

I was discussing with Kevin, a fellow PhD student at the Kastritis lab and now a dear friend, how to write the acknowledgements section. Do you keep it short and formal, like you would do in a publication? Or do you just pour your heart out until there's nothing left? In the end he said that the acknowledgements section is written at the very end, in the middle of the night, when there is not much room for more thinking. And he was right. The cold light of day is not suitable for sentimentalities and for this I have chosen to be sentimental. So, here it goes.

After two very tiring but quite interesting BSc and MSc degrees, I was looking to continue my education on structural biology, a passion instilled in me by the legendary Professor of the University of Athens Stavros Hamodrakas. I was at the office of another mentor, Professor Costas Vorgias and we were discussing where to go or what I should follow. I mentioned to him then, that I was interested in continuing my studies in what I perceived back in 2019 to be an emerging field in structural biology, cryo-EM. He surprised me by telling me that he has an old student of his that is at that moment about to begin a cryo-EM lab, somewhere in Germany, in a small city called Halle that I had never heard of before in my life. Some months later, here I was beginning a journey in a new country, in a new lab.

Pandemic notwithstanding, it was (and still is) an insane and amazing journey filled with amazing research and even more amazing people! There are so many of you that I want to thank, that helped me scientifically and personally to make it through these last years that I am afraid I don't know where to begin. I am actually lying, I know perfectly well with who to begin, I'll start with Fotis. He was afraid that if he mentioned me too much in his PhD acknowledgements somebody may accuse him of favoritism, but I have no such qualms, as he is one of the few people that truly deserves any and all praise that come his way. He is hard-working, patient, diligent, possesses a pair of "golden hands" that can make any experiment work, and it was him that mentored me in the wet lab, showed me how to work with our favorite fungus, how to get the most out of our extracts! He was always there for me whenever I needed him no matter what I asked of him and I hope that I could give back even a fraction of how much he has helped me all these years. For all these reasons, any kind of written thanks I can think of right now are truly not enough.

I will continue with my fellow PhD students! We shared an office for 3+ years, toiling away on our PCs and at the lab bench, encouraging each other, consoling each other, helping each other, and all the while drinking copious amounts of beer! Lisa, whenever I need somebody to have my back I will give you a call, just don't forget to bring your katana. Kevin, my brother in misery, nihilism and alcoholism, these years that we were together were always more bearable because you were sitting next to me, sending encouragement (or proclaiming the abject futility of it all, in equal measures). You are a true friend, a scholar and I am glad that I am able to call you so, for your selflessness and strength are a paradigm for the rest of us. Lastly there's Marija, who I may not have come to spend so much time together, but I always appreciate her intelligence and upbeat character, and Toni, who quickly snuck into my heart, an amazing guy all around, both in the lab and outside, and the keeper of the spiciest of memes.

Let's move on to the senior scientists of the group! Dmitry, wizard of image analysis, thank you for your help throughout these years, you always offered a helping hand whenever I was stuck in a project. Jaydeep, you arrived in the lab a bit later but you always amazed me with your expert knowledge on programming, physics, your humor and your impeccable fashion sense. Christian, what can I say, apart from the scientific part, where I have always benefited from your in-depth knowledge of biochemistry, molecular biology, and bioinformatics and coding, I am more grateful for your friendship. Our small breaks at the balcony, hanging out together and talking about whatever, from your kids, to music, to science in general were always a highlight of my day. We were often each other's sounding board when we were stuck and you always had some insight to offer. Whenever we collaborated, and it was often, we worked as a well-oiled machine from the get-go. I'd like to think that we will continue to do so in the future, over some metal music and beer. Farzad, you taught me much about microscopy, you took care of my data acquisitions and in you I found a kindred soul with who I could have deep discussions about the state of the world we live in, condemn together the injustices and reveal our hopes for a better tomorrow, in different conditions.

A big thank you goes to our coordinators, Ulla, Marie, Rosita, Fabienne for all their administrative help throughout the years, all my collaborators that have provided me with critical data, including Professor Stubbs, Professor Rappsilber, Professor Heilmann, Dr. O' Reilly, Dr. Hause, Dr. Chojnowski, Dr. Fratini, just to name a few! I

want to also thank the speaker of my RTG, Professor Andrea Sinz, as well as the RTG administrative coordinator Claudia Spielmann for their help and guidance while I was part of the disordered protein family! I also shouldn't forget to thank Johannes, Anna, Felix, Wiebke and Noah for their work during their internships at our lab!

From my lab, there is one person that I should truly thank above all, my professor, my mentor and, now hopefully, my friend, Jr. Professor Panagiotis Kastritis. He trusted me and brought me to his lab right at the start. He was always there when I needed him and he mentored me in all things cryo-EM and structural biology. I was always at his office, bugging him with questions, and he never dismissed me, he always made time to help me and show me something new. He is an example of how a modern, aspiring group leader should be, full of drive and brilliant ideas that will promote his science and his people to the next level. Thank you Panos for giving me the opportunity to be a part of your scientific journey and for providing me with the provisions I need to start my own! I hope I did not let you down!

We now come to the personal part. My mother, Kiki, was always supportive of my dreams. She never discouraged me from pursuing my science and it is for a big part through her sacrifices that I am able to be where I am right now! My sister, Ifigeneia, you are always a voice of reason, and your self-sacrifice and support is something that I never take for granted. You also start now your journey in the treacherous lands of the PhD and I hope I can provide some help to make it easier. I love you both very much!

I will close with the person that has stood by me the last 9 months (years, but time feels like it goes by so fast) through thick and thin, through good and bad moments, my partner, my love, Linda. Thank you for supporting me, thank you for being here, by me. You keep the darkness at bay, just my thought of you is a beacon that shines through everything! I love you to the moon and back!

Thank you all for this amazing journey!
28.11.22, 4:58 am.

P.S. If I forgot somebody important, get back to me and I'll treat you to a beer (or a co-authorship in my next manuscript, dealer's choice).

9 Curriculum Vitae

SKALIDIS IOANNIS

Education

PhD **Biochemistry** 2019 - 2023
Grade: 1,0 Summa cum laude
Kastritis Laboratory for Biomolecular Research, IWE ZIK HALOmem, Faculty of Biochemistry, Martin-Luther University Halle-Wittenberg

MSc **Systems Biology** 2017 - 2019
Grade: 9,72/10
Agricultural University of Athens, Department of Biotechnology and Food Science, Faculty of Biotechnology

BSc **Biology** 2009 - 2017
Grade: 6,1/10
National and Kapodistrian University of Athens, Faculty of Biology

Research Projects

Integrative Structural Biology of Metabolons (PhD Research Topic) 2019 - Present
Kastritis Laboratory for Biomolecular Research, IWE ZIK HALOmem, Faculty of Biochemistry, Martin-Luther Universitaet Halle-Wittenberg

Differential Gene Signature – Assisted Drug Repositioning in Malignant Melanoma. (MSc Thesis) 2018 - 2019
Biomedical Systems Laboratory, School of Mechanical Engineering, National Technical University of Athens, Athens, Greece

CK2 Protein Crystallization – X-ray Diffraction (BSc Thesis) 2014 - 2016
Structural Biology & Chemistry Laboratory, Institute of Biology, Medicinal Chemistry & Biotechnology, National Hellenic Research Foundation, Athens, Greece

CK2 Protein Expression and Homology Modelling (BSc Thesis) 2013 - 2014
Department of Cell Biology and Biophysics, Faculty of Biology, National and Kapodistrian University of Athens, Athens, Greece

Training - Workshops

MicroED Imaging Center workshop 2022
Workshop, UCLA

Academic writing: How to create good texts 2022
Workshop, IMPRS-STNS

CCP4 Study Weekend 2022 2021
Workshop, UKRI / CCP4, Online

Recent advances in structural biology of membrane proteins 2021
Workshop, EMBO, Online

Cryogenic Electron Tomography in Structural Biology 2021
Workshop, Instruct-ERIC events / eBIC – Diamond Light Source, Online

Stress Management 2021
Workshop, Martin-Luther Universitaet Halle-Wittenberg / Desiree Dickerson

Scientists in Leadership 2021
Workshop, Martin-Luther Universitaet Halle-Wittenberg / Gaby Schilling Coaching

Conflict Management in Academia 2021
Workshop, IPB, Martin-Luther Universitaet Halle-Wittenberg / Golin Wissenschaftsmanagement

Social Media and Online Science Communication 2021
Workshop, Martin-Luther Universitaet Halle-Wittenberg / NaWik, Online

Introduction to Statistics and R (Parts I & II) 2021
Workshop, iDiv, Martin-Luther Universitaet Halle-Wittenberg, IPB, Online

Good Manufacturing/Clinical/Laboratory Practice 2021
Workshop, InGrA, Martin-Luther Universitaet Halle-Wittenberg / Thomas Beer und Tobias Halfpap GbR (gmp-kurs.de), Online

Recognition of pain, suffering and distress and its application in the evaluation of severity of the procedures, species specific: mice and rats. (HERMES Project) 2020
Training, Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise, Erasmus+, Online

Innovation – Design thinking with Agile Methods 2020
Workshop, InGrA, Martin-Luther Universitaet Halle-Wittenberg / Roloff & Schumacher gmbh, Online

Good Scientific Practice 2020
Workshop, InGrA, Martin-Luther Universitaet Halle-Wittenberg / Universitaet Leipzig, Online

Statistics and Statistical thinking 2020
Workshop, Dr. Peter Paul Heym (Sum Of Square), Online

Instruct course on Image Processing for Electron Microscopy and Hybrid Modelling 2019
Training, CNB, Madrid, Spain

Master the Network. Efficient partner search and management 2019
Transnational Cooperation Activities, Erasmus+, Agencia Nacional Española, Mollina, Spain.

Sustainability Ambassadors 2017
Training Course, Erasmus+, Asociacion Accion en 3, Oviedo, Spain.

Exploring Structure Based Drug Design with Maestro 2015
High Throughput RNA Seq data and causal network analysis: From Reads to Insight
Bioinformatics Solutions for NGS applications
Introduction to Meta-Analysis
Workshops, 10th Conference of the Hellenic Society for Computational Biology and Bioinformatics - HSCBB15, Athens, Greece

Principles of Computer-Aided Drug Design 2014
Introduction to the concepts of comparative genomics focusing on bacterial genomes
Workshops, 9th Conference of the Hellenic Society for Computational Biology and Bioinformatics - HSCBB14, Athens, Greece

Conferences - Presentations

AI-guided cryo-EM probes a thermophilic cell-free system with succinyl-CoA manufacturing capability 2022
Cryo-EM and artificial intelligence visualize endogenous protein community members
Poster – 2nd HALOmem International Meeting, Halle (Saale), Germany

Cryo-EM and artificial intelligence visualize endogenous protein community members 2022
Invited Talk / Poster – Instruct Biennial Structural Biology Conference - Changes in structural biology: challenges in studying dynamics, Utrecht, Netherlands

Cryo-EM and artificial intelligence visualize endogenous protein community members 2021
Poster / Flash Talk – NWO Chains, Veldhoven, Netherlands

Disorder behind order: AlphaFold miss-shots in *de novo* modelling of cryo-EM maps from native cell extracts 2021
Presentation – GRK2467 Retreat, Wittenberg, Germany

Simultaneous analysis of endogenous protein community members by high-resolution cryo-EM 2021
Presentation – HALOmem Status Seminar, Halle (Saale), Germany

Simultaneous analysis of endogenous protein community members by high-resolution cryo-EM 2021
Poster / Flash Talk – EMBL Conference: Bringing Molecular Structure to Life: 50 Years of the PDB, Heidelberg, Germany

Cryo-EM reveals insights into the native oxoglutarate dehydrogenase complex 2021
Presentation – GRK2467 Weekly Seminars, Halle (Saale), Germany

Simultaneous analysis of endogenous protein community members by high-resolution cryo-EM 2021
Presentation – HALOmem Retreat, Wittenberg, Germany

Simultaneous analysis of endogenous protein community members by high-resolution cryo-EM 2021
Presentation – GRK2467 Student Conference, Halle (Saale), Germany

Next-generation structural systems biology of native cell extracts reveals insights into pyruvate oxidation 2021
Presentation – GRK2467 Invited Speaker, Halle (Saale), Germany

Image processing for Cryo-EM data: Principles and Applications 2020
Presentation – GRK2467 Retreat, Wittenberg, Germany

Unstructured regions of large enzymatic complexes control the availability of metabolites with signaling functions 2020
Presentation – GRK2467 Weekly Seminars, Halle (Saale), Germany

Tearing down the wall: *C. thermophilum* spheroplasts for future applications 2020
Presentation – HALOmem yearly meeting, Halle (Saale), Germany

PDHc Core / Complex Analysis 2020
Presentation – GRK2467 Weekly Seminars, Halle (Saale), Germany

Initial Disorder Analysis of main players involved in Hypoxia Signalling 2020
Presentation – GRK2467 Weekly Seminars, Halle (Saale), Germany

Computational resources for protein disorder analysis, paper formatting and ideas 2020
Presentation – GRK2467 Weekly Seminars, Halle (Saale), Germany

Structural role of disorder in pyruvate oxidation through computational and EM analysis: The case of Protein X 2019
Presentation – GRK2467 Retreat, Wittenberg, Germany

Structural insights into Chaetomium thermophilum native cell extracts with cryo-EM 2019
Poster – 1st HALOmem International Meeting, Halle (Saale), Germany

Crystallization and structure determination assays of the CK2 Protein Kinase from the insect *Ceratitis capitata* 2016
Poster – Joint Conference of the Hellenic Crystallographic Association and the Hellenic Society for Computational Biology and Bioinformatics HECRA-HSCBB16, Athens, Greece

Preliminary structural studies of the CK2 Protein Kinase from the insect *Ceratitis capitata* 2015
Poster – 66th Congress of the Hellenic Society of Biochemistry and Molecular Biology, Thessaloniki, Greece

CK2 of *Ceratitis capitata* shares central properties with human CK2
Poster – 65th Congress of the Hellenic Society of Biochemistry and Molecular Biology, Thessaloniki, Greece 2014

Software & Hardware Skills

Microsoft: Excellent Knowledge of OS (XP, Vista, 7, 8, 10) and MS Office Suite (Word, Excel, PowerPoint, Outlook)
Linux: Experience with Unix-based OS.
Statistical Analysis: Basic Knowledge of SPSS, Good Knowledge of GraphPad Prism 6.
Programming Languages: Good Knowledge of Python, R, Bash, Perl
Bioinformatics Databases: Excellent Knowledge of various Bioinfromatics Databases (UniProt, PDB, EMDB, STRING, etc)
Modelling Software: Good knowledge of MODELLER, PyMol, coot, Phenix, AlphaFold2, Cytoscape
Cryo-EM Software: RELION, Scipion, cryoSPARC, Chimera/ChimeraX
Project Management: Basic Knowledge of Trello and OpenPM²
Hardware: Basic Hardware repair and assembly skills

Languages

Greek: Native Speaker.
English: Bilingual Speaker, CPE (C2), University of Cambridge/University of Michigan.
German: Novice Speaker, Intermediate Certificate (B2), Martin-Luther Universitaet International Office Language School
French: Advanced Speaker, Sorbonne 2ème degré (C2), Université Paris-Sorbonne (Paris IV).
Spanish: Novice Speaker, Foreign Language School Certification (A2), National Kapodistrian University of Athens.

10 Publication list

First author publications (Material included in these publications has been included in the current thesis):

- **Skalidis, I.**, Kyrilis, F. L., Tueting, C., Hamdi, F., Traeger, T. K., Belapure, J., . . . Kastritis, P. L. (2022). AI-guided cryo-EM probes a thermophilic cell-free system with succinyl-CoA manufacturing capability. bioRxiv, 2022.2010.2008.511438. doi:10.1101/2022.10.08.511438
- **Skalidis, I.**, Kyrilis, F. L., Tüting, C., Hamdi, F., Chojnowski, G., & Kastritis, P. L. (2022). Cryo-EM and artificial intelligence visualize endogenous protein community members. *Structure (London, England: 1993)*, *30*(4), 575–589.e6. doi: https://doi.org/10.1016/j.str.2022.01.001
- **Skalidis, I**., Tüting, C. & Kastritis, P.L. (2020). Unstructured regions of large enzymatic complexes control the availability of metabolites with signaling functions. Cell Commun Signal 18, 136. doi: https://doi.org/10.1186/s12964-020-00631-9

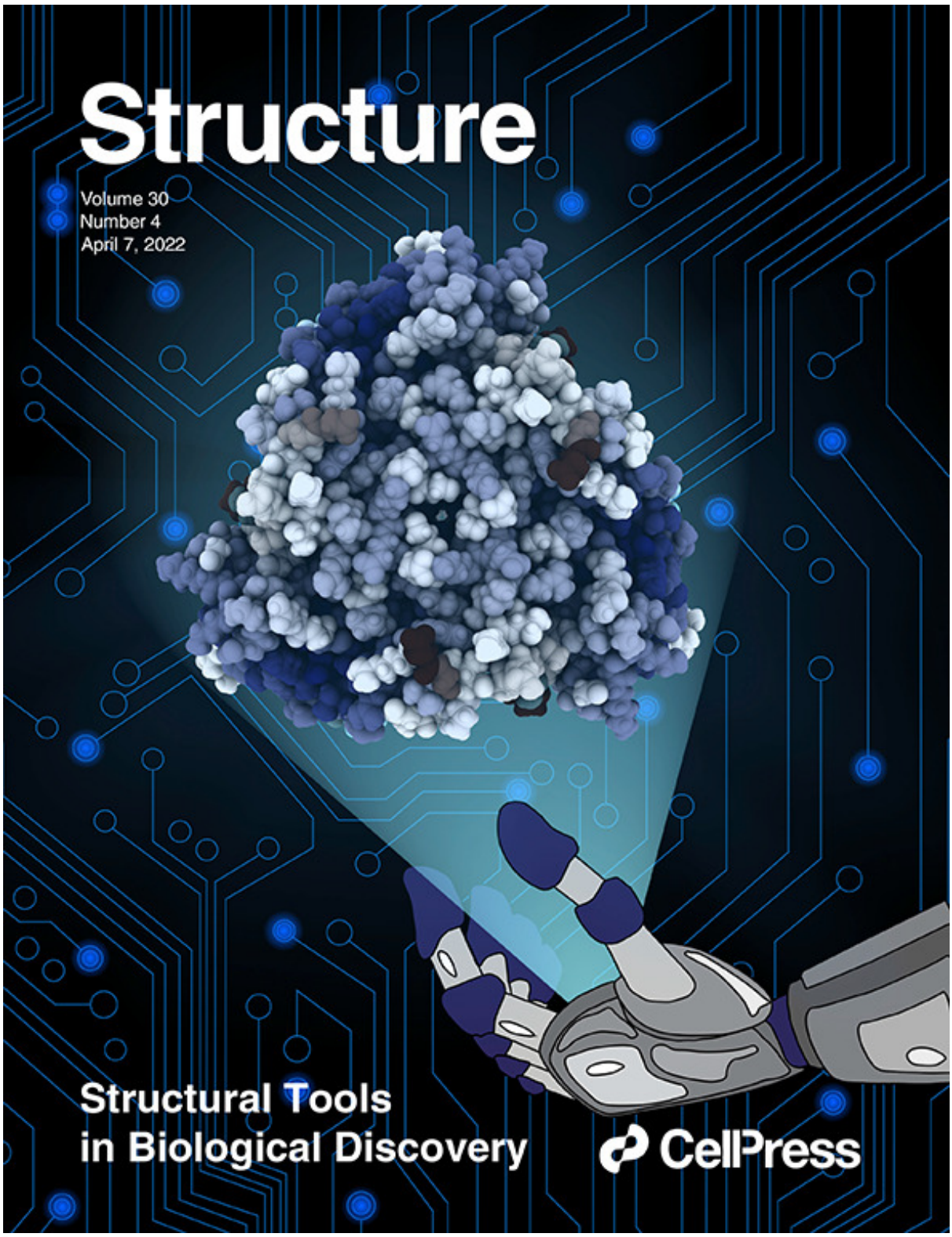
Second author publications:

- Kyrilis, F. L., Semchonok, D. A., **Skalidis, I**., Tüting, C., Hamdi, F., O'Reilly, F. J., . . . Kastritis, P. L. (2021). Integrative structure of a 10-megadalton eukaryotic pyruvate dehydrogenase complex from native cell extracts. Cell Reports, 34(6), 108727. doi: https://doi.org/10.1016/j.celrep.2021.108727

Co-authored publications (Method developments during this thesis have been used in the following publications):

- Piersimoni, L., Abd El Malek, M., Bhatia, T., Bender, J., Brankatschk, C., Calvo Sánchez, J., Dayhoff, G. W., Di Ianni, A., Figueroa Parra, J. O., Garcia-Martinez, D., Hesselbarth, J., Köppen, J., Lauth, L. M., Lippik, L., Machner, L., Sachan, S., Schmidt, L., Selle, R., **Skalidis, I.**, Sorokin, O., … Uversky, V. N. (2022). Lighting up Nobel Prize-winning studies with protein intrinsic disorder. *Cellular and molecular life sciences: CMLS*, *79*(8), 449. doi: https://doi.org/10.1007/s00018-022-04468-y
- Tüting, C., Kyrilis, F. L., Müller, J., Sorokina, M., **Skalidis, I.**, Hamdi, F., Sadian, Y., & Kastritis, P. L. (2021). Cryo-EM snapshots of a native lysate provide structural insights into a metabolon-embedded transacetylase reaction. Nature communications, 12(1), 6933. doi: https://doi.org/10.1038/s41467-021-27287-4
- Janson, K., Zierath, J., Kyrilis, F. L., Semchonok, D. A., Hamdi, F., **Skalidis, I**. et al. (2021). "Solubilization of artificial mitochondrial membranes by amphiphilic copolymers of different charge." Biochim Biophys Acta Biomembr 1863(12): 183725. doi: https://doi.org/10.1016/j.bbamem.2021.183725
- Hamdi F, Tüting C, Semchonok DA, Visscher KM, Kyrilis FL, Meister A, **Skalidis I.** et al. (2020) 2.7 Å cryo-EM structure of vitrified M. musculus H-chain apoferritin from a compact 200 keV cryo-microscope. PLoS ONE 15(5): e0232540. doi: https://doi.org/10.1371/journal.pone.0232540

Part of the work included in this thesis was published in *Structure* (Cell Press) and was featured on the cover of the journal:



]

On the cover: A robotic hand sheds an electron beam, revealing the structure of the Pyruvate Dehydrogenase Complex E2 trimer. In the background, a lit-up circuit represents the artificial intelligence that supports the structure's elucidation in all steps of the process (Skalidis *et al.*, 575–589). Image courtesy of the authors.

11 Declaration

Declaration

I hereby declare that I have written this thesis independently and without external assistance other than the mentioned sources and aids being cited in this dissertation. Therefore, any extracts of external works used literally or figuratively in the present thesis are outlined and cited accordingly. I also declare that I have not applied this thesis at any other college or university in order to obtain an academic degree.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und ohne fremde Hilfe verfasst habe. Ich habe keine anderen als die hier angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inthaltlich entnommenen Stellen als solche kenntlich gemacht habe. Ferner erkläre ich, dass ich mich mit dieser Arbeit an keiner anderen Hochschule oder Universität um die Erlangung eines akademischen Grades beworben habe.

Halle (Saale), Ioannis Skalidis