

The ultra-high affinity transport proteins of ubiquitous marine bacteria

<https://doi.org/10.1038/s41586-024-07924-w>

Ben E. Clifton¹✉, Uria Alcolombri², Gen-Ichiro Uechi¹, Colin J. Jackson^{3,4,5} & Paola Laurino^{1,6}✉

Received: 10 March 2023

Accepted: 7 August 2024

Published online: 11 September 2024

Open access

 Check for updates

SAR11 bacteria are the most abundant microorganisms in the surface ocean¹ and have global biogeochemical importance^{2–4}. To thrive in their competitive oligotrophic environment, these bacteria rely heavily on solute-binding proteins that facilitate uptake of specific substrates via membrane transporters^{5,6}. The functions and properties of these transport proteins are key factors in the assimilation of dissolved organic matter and biogeochemical cycling of nutrients in the ocean, but they have remained largely inaccessible to experimental investigation. Here we performed genome-wide experimental characterization of all solute-binding proteins in a prototypical SAR11 bacterium, revealing specific functions and general trends in their properties that contribute to the success of SAR11 bacteria in oligotrophic environments. We found that the solute-binding proteins of SAR11 bacteria have extremely high binding affinity (dissociation constant >20 pM) and high binding specificity, revealing molecular mechanisms of oligotrophic adaptation. Our functional data have uncovered new carbon sources for SAR11 bacteria and enable accurate biogeographical analysis of SAR11 substrate uptake capabilities throughout the ocean. This study provides a comprehensive view of the substrate uptake capabilities of ubiquitous marine bacteria, providing a necessary foundation for understanding their contribution to assimilation of dissolved organic matter in marine ecosystems.

The sunlit surface ocean is dominated by heterotrophic bacterioplankton, particularly those belonging to the SAR11 clade of Alphaproteobacteria (*Pelagibacterales*)². SAR11 bacteria are globally distributed and abundant, constituting 20–45% of prokaryotic cells and around 18% of the biomass in the surface ocean and having an estimated global population size of 2.4×10^{28} cells^{1,7}. Like other bacteria adapted to oligotrophic (nutrient-poor) environments, they exhibit a small size (cell volume of $0.02\text{--}0.06 \mu\text{m}^3$)⁸, extremely streamlined genome (1.2–1.4 Mb), and limited metabolic versatility^{2,9,10}. SAR11 bacteria rely largely on the uptake of dissolved organic matter (DOM) to meet their requirements for carbon, nitrogen, sulfur and phosphorus, and are highly active consumers of low molecular mass DOM, accounting for 30–60% of assimilation of amino acids, taurine, glucose and dimethylsulfoniopropionate (DMSP) in the surface ocean^{3,11–15}. Owing to their high abundance, SAR11 bacteria have major biogeochemical importance—for example, they produce climate-active gases such as methane³ and dimethylsulfide⁴, and divert carbon from the biological carbon pump through respiration of dissolved organic carbon¹⁶ (DOC). Thus, understanding the physiology and metabolic capabilities of SAR11 bacteria is critical to our understanding of marine ecosystems.

To compete for nutrients in the oligotrophic ocean environment, SAR11 bacteria rely heavily on solute-binding proteins (SBPs) to facilitate uptake of specific growth substrates. SBPs are associated with three

families of membrane transport systems: ATP-binding cassette (ABC) transporters, which represent the most abundant high-affinity substrate uptake systems in bacteria, as well as tripartite ATP-independent periplasmic (TRAP) transporters and tripartite tricarboxylate transporters (TTTs)^{17,18}. In Gram-negative bacteria, SBPs bind their substrates with high affinity in the periplasm (with dissociation constant (K_d) values typically in the nanomolar to micromolar range) and deliver them to a transmembrane protein complex that facilitates translocation of the substrate across the inner membrane against a concentration gradient, which is driven by coupling to ATP hydrolysis (ABC transporters) or an electrochemical gradient (TRAP and TTT transporters). Consistent with the physiological importance of SBP-dependent transporters for high-affinity substrate uptake, SAR11 bacteria devote a large proportion of their streamlined proteome to SBPs^{5,6}; for example, SBPs represented around 67% of SAR11-derived spectra in metaproteomic analysis of environmental samples from the Sargasso Sea⁵.

The high abundance^{1,7} and substrate uptake activity^{3,11–15} of SAR11 bacteria in the surface ocean, combined with the abundance of SBP-dependent transporters in these bacteria^{5,6}, suggests that a small number of transport proteins in SAR11 bacteria make a substantial contribution to global assimilation of key components of low molecular mass DOM in the surface ocean. However, the properties of these transporters and their specific functions (that is, the transported metabolites) are

¹Protein Engineering and Evolution Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Japan. ²Department of Plant and Environmental Sciences, Hebrew University of Jerusalem, Jerusalem, Israel. ³Research School of Chemistry, Australian National University, Canberra, Australian Capital Territory, Australia. ⁴ARC Centre of Excellence for Innovations in Peptide and Protein Science, Research School of Chemistry, Australian National University, Canberra, Australian Capital Territory, Australia. ⁵ARC Centre of Excellence in Synthetic Biology, Research School of Chemistry, Australian National University, Canberra, Australian Capital Territory, Australia. ⁶Institute for Protein Research, Osaka University, Suita, Japan. ✉e-mail: benjamin.clifton@oist.jp; paola.laurino@oist.jp

mostly unknown, which limits our knowledge of the full range of DOM that can be assimilated by SAR11 bacteria, nutrient exchange within the marine microbial community, and the molecular mechanisms for high-affinity substrate uptake. Although homology-based predictions are available, these predictions of protein function have limited accuracy, especially for functionally diverse protein superfamilies such as the ABC and TRAP transporters^{19–21}. Transport proteins can also be characterized experimentally through radioassays of substrate uptake in cultured cells, and this approach has been used to characterize the broad-specificity osmolyte transporter of the SAR11 bacterium '*Candidatus* (Ca.) Pelagibacter ubique'²². However, the difficulty of cultivating slow-growing and fastidious SAR11 bacteria makes this a challenging approach for high-throughput characterization of SBP-dependent transporters. Furthermore, owing to the genetic intractability of SAR11 bacteria, the observed transport activity cannot be linked to specific transporter genes, limiting integration of the resulting physiological data with existing multi-omics datasets to uncover the broader geochemical and ecological significance of transport activity. Given these limitations of in vivo approaches, we hypothesized that a heterologous approach, based on heterologous expression, purification and biochemical characterization of SBPs, would be an effective alternative strategy to elucidate the functions and properties of SBP-dependent transporters in SAR11 bacteria. This approach is supported by the fact that the specificity and affinity of substrate uptake by SBP-dependent transporters is mainly determined by the binding specificity and affinity of the corresponding SBPs²³, and has been proven to be a valuable method for discovery of new metabolic pathways^{24,25}.

Here we used this approach to systematically interrogate the function of all SBP-dependent transporters in the genome of the prototypical SAR11 bacterium *Ca. P. ubique* strain HTCC1062. Using high-throughput screening together with rigorous structural and biophysical characterization, we identified the function of the majority of these transporters. Revision of homology-based functional predictions enabled us to accurately interpret patterns of SAR11 transporter abundance in global ocean metagenome and metatranscriptome datasets and identify new transport capabilities and potential carbon sources for SAR11 bacteria. In particular, we identified a high-affinity, broad-specificity transporter for C₄ and C₅ dicarboxylates that is widely found among SAR11 ecotypes and abundantly distributed in metagenomic and metatranscriptomic datasets, implicating these dicarboxylates as major physiologically relevant carbon sources. Finally, we show how the identification of systematic trends in SBP properties, including their extremely high binding affinity, moderately high binding specificity and limited functional redundancy, provides insight into the evolutionary success of SAR11 bacteria in the oligotrophic ocean environment.

Functional profiling of SAR11 SBPs

We identified 18 SBPs through genomic analysis of *Ca. P. ubique* strain HTCC1062 (Methods, Supplementary Table 1); these SBPs are found widely across SAR11 bacteria, and conversely, most of the SBPs that are abundant across SAR11 bacteria are represented in this strain (Extended Data Fig. 1). Expression of most of these SBPs in cultured and/or environmental SAR11 cells has been previously demonstrated by proteomic analysis^{5,6} (Supplementary Table 2). Fourteen of the SBPs yielded soluble protein upon heterologous expression in *Escherichia coli* strains BL21(DE3) or SHuffle T7 and were successfully purified. Close homologues of two of the remaining SBPs (denoted SAR11_0271* and SAR11_1346*) from a different SAR11 strain ('*Ca. Pelagibacter*' sp. HIMB1321) could also be expressed and purified, whereas the remaining two SBPs, SAR11_0266 and SAR11_1290, could not be expressed in soluble form under any tested condition, nor refolded in vitro from insoluble material (Methods). Two of the SBPs, SAR11_1179 and SAR11_1238, were predicted to represent proteins that are found widely in bacteria and bind inorganic solutes with high specificity: phosphate and iron(III),

respectively. Thus, the functional predictions for these two proteins were directly tested and confirmed by differential scanning fluorimetry (DSF) and isothermal titration calorimetry (ITC) (SAR11_1179; Extended Data Fig. 2a–c) or UV–vis spectroscopy (SAR11_1238, Extended Data Fig. 2d–i) rather than high-throughput screening.

In the remaining cases, the tentative function of each SBP was first identified by high-throughput screening of metabolite libraries by DSF. First, the target protein was screened by DSF against a commercially available metabolite library, representing a set of around 330 unique metabolites, including many common carbon, nitrogen, phosphorus and sulfur sources (full list in Supplementary Table 3). This library was supplemented with a manually curated set of around 40 metabolites that are known to be important for *Ca. P. ubique* and other marine bacteria^{26–29} (for example, osmolytes, sulfonates, and vitamin derivatives) or that were considered to be potential ligands on the basis of the computational annotations of SBP function (for example, opines). Metabolites that resulted in an increase in melting temperature (ΔT_M) of the protein of at least 2 °C by DSF were considered to be potential ligands (Supplementary Table 4). Second, a representative subset of the resulting hits was selected, and binding of this subset of ligands to the target protein was confirmed and rank-ordered by repeating DSF with each ligand at a fixed concentration (10 mM) (Fig. 2). Finally, to provide further evidence that the observed increases in T_M were a result of specific, high-affinity protein–ligand interactions rather than non-specific protein stabilization, the DSF experiments were repeated with a range of ligand concentrations (Supplementary Fig. 1). Using this workflow, tentative functions were identified for 15 SBPs—that is, all proteins that could be expressed and purified, except SAR11_1068. We showed previously that SAR11_1068 does not have the annotated function (cyclohexadienyl dehydratase activity) and reported extensive but ultimately unsuccessful efforts to identify its function³⁰. The protein was subjected to further high-throughput screening in this work, but no potential ligands were identified.

Next, the function of each SBP was confirmed by ITC, enabling accurate quantification of binding affinity, which is important to establish the physiological relevance of the observed protein–ligand interactions. We typically performed titrations for 2 to 5 ligands for each protein, for a total of 32 protein–ligand interactions, aiming to select ligands with a range of ΔT_M values to enable estimation of binding affinity for a broader range of ligands, on the basis of the assumption that protein–ligand interactions that yield similar ΔT_M values in DSF experiments for a given protein are similar in binding affinity. Using ITC, most of the protein–ligand interactions identified by DSF could be verified (Fig. 2, Supplementary Fig. 2 and Supplementary Table 5), and at least one high-affinity ligand (with $K_d < 500$ nM) could be identified for each SBP, except SAR11_0271* and SAR11_0797 (Supplementary Note 1). The interaction between SAR11_1302 and TMAO was confirmed by ITC in another report while this work was in progress³¹. Thus, in total, based on the DSF and ITC data, 13 out of the 18 SAR11 SBPs of *Ca. P. ubique* HTCC1062 could be confidently assigned a binding function (Fig. 1 and Supplementary Table 6). Further evidence in support of the functional assignment for six SBPs could be obtained on the basis of a metabolome screening approach²⁴ using X-ray crystallography and/or gas chromatography–mass spectrometry (GC–MS) (Extended Data Fig. 3 and Supplementary Note 2). In addition, the physiological relevance of the proposed binding functions is highlighted by the identification of SBPs that bind known substrates of *Ca. P. ubique* HTCC1062 such as amino acids, D-glucose, DMSP and taurine with high affinity^{3,11–14} and the fact the measured concentrations of these substrates in the surface ocean frequently exceed the measured K_d values of the corresponding SBPs (Supplementary Note 3 and Supplementary Table 7).

The SBPs of *Ca. P. ubique* HTCC1062 showed remarkably high binding affinities, with multiple SBPs having K_d values as low as 20–30 pM (Fig. 3a,b). Seven out of the thirteen SBPs had K_d values of <5 nM, which are below the quantification limit of direct ITC experiments; thus, to

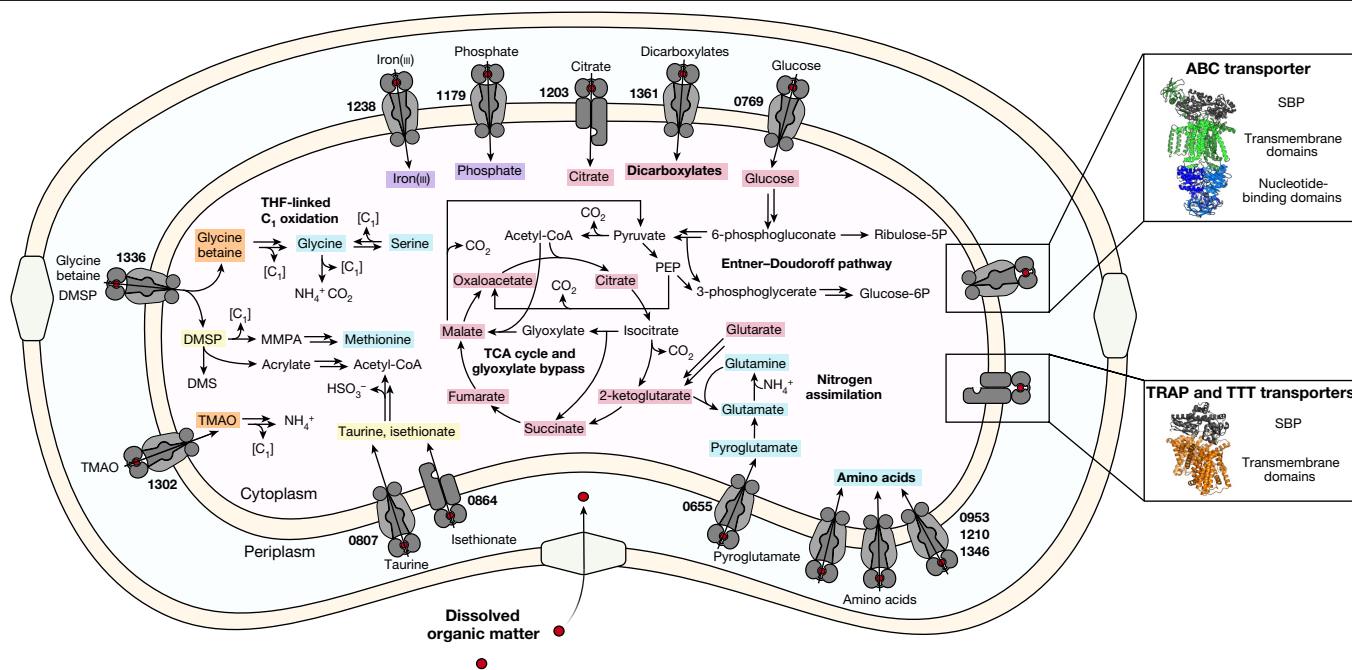


Fig. 1 | Ubiquitous SAR11 bacteria use SBP-dependent transporters to uptake DOM from the ocean and have a critical role in oceanic cycles. Schematic of a SAR11 bacterium showing the SBP-dependent transporters of *Ca. P. ubique* HTCC1062 and their relationship with key pathways of carbon, nitrogen and sulfur metabolism. Metabolites identified here as SBP ligands are highlighted in colour: blue, amino acids and their derivatives; red, carbon sources; yellow, sulfur-containing metabolites; orange, C₁ sources; purple, inorganic ions. Metabolic pathways are based on reactions proposed here (Extended Data Fig. 9) and in refs. 28,29,42,44,56,57. [C₁] refers to a C₁ unit

oxidized to CO₂ via the THF-linked oxidation pathway. The ABC transporter structure shown is an X-ray crystal structure of MalEFGK₂ from *E. coli* (Protein Data Bank (PDB) 2R6G) and the structure representative of the TRAP and TTT architecture is a model of the TRAP transporter SiaPQM from *Haemophilus influenzae*⁵⁸. DMS, dimethylsulfide; glucose-6P, glucose-6-phosphate; MMPA, methylmercaptopyruvate; PEP, phosphoenolpyruvate; ribulose-5P, ribulose-5-phosphate; THF, tetrahydrofolate; TMAO, trimethylamine-*N*-oxide. Figure partly created with BioRender.com.

obtain accurate K_d values for these interactions, we also performed competitive ITC binding experiments (Supplementary Table 5). In the case of SAR11_1210, titration with L-arginine in the presence of D-octopine as a competing ligand indicated a K_d value between 10 pM and 100 pM for L-arginine, but variable results were obtained with different concentrations of D-octopine (Supplementary Data 1). Thus, to confirm the high affinity of this interaction, we also performed a protein–protein competition experiment in which SAR11_1210 was mixed with a previously characterized arginine-binding protein, ArgT from *Salmonella enterica* (K_d = 15 nM), and then titrated with L-arginine. Fitting the resulting data to a two-sets-of-sites binding model yielded a K_d of 32 pM for the interaction between SAR11_1210 and L-arginine (Fig. 3c). We also solved the crystal structure of SAR11_1210 complexed with L-arginine, which showed an unusual binding mode involving a direct interaction between the ligand and the flexible hinge region linking the two α/β domains of the SBP, suggesting a possible structural basis for the high binding affinity (Fig. 3e, Extended Data Fig. 4 and Supplementary Note 4). Finally, in the case of SAR11_0769, titration with D-glucose reproducibly yielded a biphasic binding isotherm (Fig. 3d), which most probably reflects differential binding of the α and β anomers of D-glucose, as supported by a crystal structure of SAR11_0769 complexed with β -D-glucose (Fig. 3f, Extended Data Fig. 5 and Supplementary Note 5). Fitting the ITC data to a competitive binding model enabled estimation of the upper limit of K_d (lower limit of affinity) for the high-affinity anomer as approximately 27 pM. A systematic survey of literature data (n = 206 SBPs) revealed that the typical range of SBP K_d values for organic solutes is 10–1,000 nM, with a lower limit of 200–400 pM ($\log_{10} K_d$ values -6.76 ± 1.15 (mean $\pm$ s.d.), Fig. 3a). Together, these results provide robust evidence that some SBPs in SAR11 bacteria exceed the previously established limits of SBP affinity for organic solutes.

Reinterpreting transport gene function

Comparison of the experimentally determined functions of each SBP with the homology-based predictions indicated that the accuracy of the predictions was low (Supplementary Table 6). The binding specificities of four proteins, SAR11_0807, SAR11_1179, SAR11_1203 and SAR11_1238, were correctly predicted as taurine, phosphate, citrate and iron(III), respectively. The predictions of SAR11_0769, SAR11_0953 and SAR11_1346 as a sugar-binding protein, general amino acid-binding protein and branched-chain amino acid-binding protein were broadly correct, although experimental characterization enabled identification of specific ligands. By contrast, 7 out of the 15 testable functional annotations were incorrect; in 5 of these cases, the binding specificity could be determined experimentally. For example, SAR11_1336 (encoded by *potD*), which was annotated as a putative spermidine or putrescine-binding protein, showed broad specificity for glycine betaine, DMSP and other osmolytes. The binding specificity of this protein matches the transport activity of a broad-specificity osmolyte transporter that was previously characterized *in vivo*, which was putatively attributed to SAR11_0797 (*proX*)²². Overall, these results show that the SBP-dependent transporters of SAR11 bacteria transport a narrower range of nitrogen sources and broader range of carbon sources and exhibit less functional redundancy than predicted³².

The assignment of transport capabilities to specific genes enables integration of the functional data with existing genomic, transcriptomic, and proteomic data. For example, functional assignment of the *Ca. P. ubique* HTCC1062 SBPs enabled analysis of the geographical distribution of various transport capabilities across SAR11 and other marine bacteria using ocean metagenome and metatranscriptome data. First, using the Ocean Gene Atlas tool³³, we analysed the abundance of homologues of the characterized SBPs in metagenome and

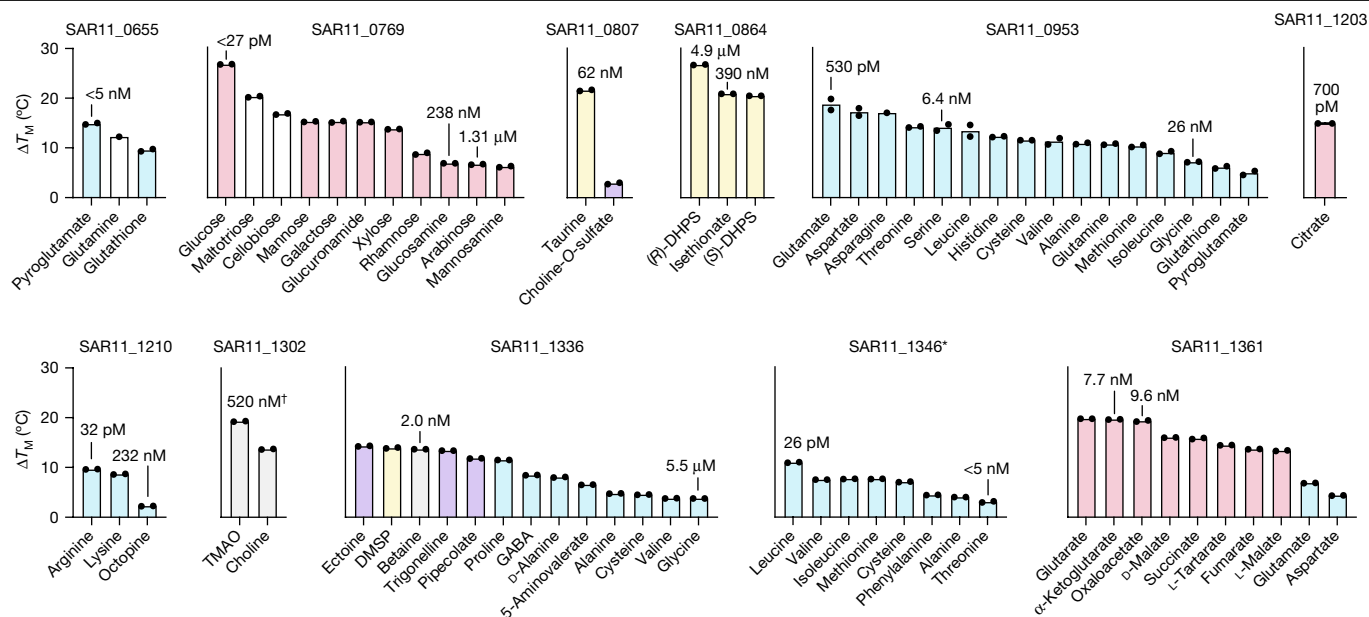


Fig. 2 | *Ca. P. ubique* HTCC1062 possesses SBPs with varying levels of binding specificity. We measured the change in denaturation temperature (ΔT_m) in the presence of 10 mM ligand by DSF. Ligands that resulted in a significant increase in T_m ($\geq 2^\circ\text{C}$) at a concentration of 10 mM are shown. Protein–ligand interactions that were verified by ITC are labelled with the measured K_d value for the interaction. Columns represent the mean of two

technical replicates shown as individual data points. Only proteins with a high-affinity ligand and confirmed by ITC are shown. The DSF results show only significant protein–ligand interactions in vitro; whether these interactions are biologically important depends on the binding affinity and the environmental concentration of the ligand. Dagger indicates that the value is from the literature³¹. DHPS, 2,3-dihydroxypropane-1-sulfonate; GABA, γ -aminobutyrate.

metatranscriptome datasets from the Tara Oceans project (Extended Data Figs. 6 and 7), which enabled us to identify abundantly transcribed transport genes that might contribute to assimilation of various components of DOM and determine whether transport gene expression correlates with known patterns of nutrient limitation and uptake. We also performed a separate metagenome analysis limited to SAR11 bacteria, estimating the percentage of SAR11 bacteria to contain a given SBP gene at each site based on the relative abundance of different SAR11 genomospecies³⁴, which yielded similar patterns of SBP abundance (Supplementary Fig. 3).

Consistent with the global abundance of SAR11 bacteria and the broad distribution of the characterized SBPs among SAR11 bacteria (Extended Data Fig. 1), and consistent with another recent analysis³⁵, homologues of most of the *Ca. P. ubique* HTCC1062 SBP genes were present at high abundance across stations in the metagenome and metatranscriptome datasets, including surface, deep chlorophyll maximum (DCM) and mesopelagic samples (Extended Data Figs. 6 and 7). In the metatranscriptome dataset, the mean abundance (fraction of mapped reads) of these SBP genes across surface stations varied from 3×10^{-5} (SAR11_0655) to 3×10^{-3} (SAR11_0953); for comparison, the mean total abundance of SBP transcripts across surface stations was estimated to be 2.7×10^{-2} , indicating that the SBPs of *Ca. P. ubique* HTCC1062 and their putatively isofunctional homologues account for a substantial proportion (around 40%) of SBP transcripts in the surface ocean. Although there is not necessarily a quantitative correlation between transcript abundance and the rate of substrate uptake, these results show qualitatively that the functional assignments of the *Ca. P. ubique* HTCC1062 SBPs are potentially significant in the broader context of global DOM assimilation. For example, the DMSP/glycine betaine transport gene SAR11_1336 showed a mean transcript abundance of 1.6×10^{-3} . The high abundance of SAR11_1336 suggests that, together with a recently described and unrelated DMSP-specific transport protein³⁶, SAR11_1336 and its homologues in other bacteria may make a significant contribution to global microbial uptake of DMSP, a metabolite that has an important role in the marine sulfur cycle and

climate regulation via its microbial conversion to the climate-active gas dimethylsulfide³⁷.

In both the metagenome and metatranscriptome analyses, transporters for sulfonates, amino acids, TMAO, glycine betaine, DMSP and dicarboxylates showed a near-universal distribution and particularly high abundance, whereas transporters for L-pyroglutamate, phosphate, iron(III) and D-glucose showed a geographically limited distribution (Extended Data Figs. 6 and 7). These results are consistent with the known contribution of SAR11 bacteria to uptake of taurine, amino acids and DMSP across different environments, compared with ecotype-specific and geographically variable uptake of D-glucose². Similar patterns of SBP gene abundance were typically observed in the metagenome and metatranscriptome datasets, consistent with high constitutive expression and limited transcriptional regulation of most SBP genes in SAR11 bacteria, with the exception of SBPs for phosphate and iron, which showed higher expression in regions of known phosphate and iron limitation^{38,39}. Notably, these interpretations of the metagenomic and metatranscriptomic data are contingent on accurate functional annotation; for example, misidentification of the SAR11 osmolyte transporter as SAR11_0797 would suggest a much more limited role for DMSP and glycine betaine uptake and broader role for polyamine uptake across SAR11 and other marine bacteria^{35,40}.

Novel functions of SAR11 SBPs

In addition to identifying transporters for known substrates, functional characterization of SBPs also enabled identification of new transport capabilities. SAR11_0655 (L-pyroglutamate) and SAR11_1361 (C_4 and C_5 dicarboxylates) represent new classes of ABC transporters and previously unknown transport capabilities of SAR11 bacteria. L-Pyroglutamate, which binds to SAR11_0655 with $K_d < 5$ nM, was an unexpected ligand, as it is a non-proteogenic amino acid that is not known to be a significant component of DOC¹⁶. Although the occurrence of SAR11_0655 is limited among SAR11 ecotypes and is

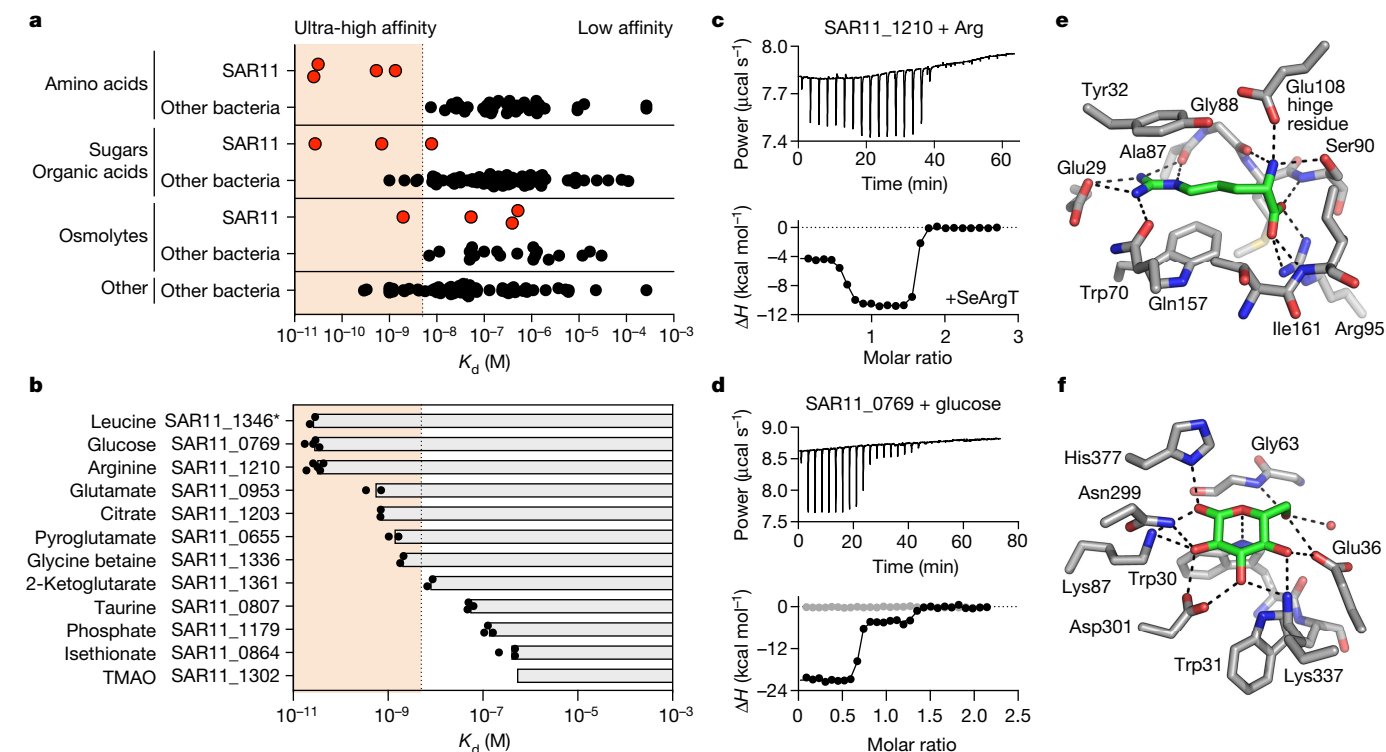


Fig. 3 | SBPs from *Ca. P. ubique* HTCC1062 exhibit ultra-high binding affinity. **a**, Comparison of K_d values for SBPs from *Ca. P. ubique* HTCC1062 with bacterial SBPs previously reported in the literature ($n = 206$; Supplementary Data 2). In cases where the SBP binds multiple ligands, data for the highest-affinity ligand is shown. SBPs with $K_d < 5$ nM are highlighted by red shading. **b**, K_d values for the highest-affinity interaction for each SBP from *Ca. P. ubique* HTCC1062. Values for SAR11_0655 and SAR11_0769 represent upper bounds on K_d (lower bounds on affinity). Bars represent mean of 2–5 technical replicates (independent titrations), shown as individual data points. Data from the literature data are shown for SAR11_1302³¹. **c, d**, Determination of binding

parameters for high-affinity interactions using competitive ITC experiments. Data fitting was performed in MicroCal PEAQ software. **c**, Simultaneous titration of SAR11_1210 and SeArgT with L-arginine. Data are representative of two replicates (separate titrations). **d**, Titration of SAR11_0769 with D-glucose, showing a biphasic binding curve. Data are representative of four replicates (separate titrations). The negative control titration (buffer plus D-glucose) is shown in grey. **e**, Binding mode of L-arginine in the crystal structure of SAR11_1210 (1.32 Å resolution). **f**, Binding mode of D-glucose in the crystal structure of SAR11_0769 (1.86 Å resolution).

restricted mainly to high latitudes (Extended Data Figs. 1, 6 and 8), other SAR11 bacteria appear to achieve L-pyroglutamate uptake using an alternative transporter (Supplementary Note 6), suggesting that L-pyroglutamate is widely utilized. Analysis of genome context suggested a putative pathway for utilization of exogenous L-pyroglutamate as a source of L-glutamate (Extended Data Fig. 9). The fact that SAR11 bacteria, despite their extremely streamlined genome, retain specific and high-affinity transporters for L-pyroglutamate indicates that this amino acid must be a widely available and useful source of carbon and/or nitrogen in the ocean. More generally, given the significant challenges of identifying environmentally important metabolites in heterogeneous, dilute and variable DOC¹⁶, this result suggests that identification of new transport capabilities from characterization of SBPs from oligotrophic marine bacteria might be a useful approach to identify new environmentally significant ocean metabolites from the DOC pool.

SAR11_1361 showed binding of a broad range of dicarboxylates that participate in the tricarboxylic acid (TCA) cycle (Figs. 1 and 2). This gene is known to be associated with carbon starvation in SAR11 bacteria; transcription and/or expression is upregulated upon carbon limitation in the dark⁴¹ (that is, energy-starved conditions) and downregulated upon nitrogen and sulfur limitation^{42,43}. Analysis of genome context also suggested a putative pathway (via SAR11_1354) for utilization of exogenous glutamate, which was subsequently confirmed to be a substrate of SAR11_1354 by ¹H-NMR and a ligand of SAR11_1361 by DSF (Extended Data Fig. 9). These results, together with the identification of a specific and high-affinity citrate-binding protein (SAR11_1203), suggest a

broad capacity of *Ca. P. ubique* HTCC1062 to assimilate dicarboxylates and TCA cycle intermediates. Genomic and biogeographical analysis indicated that the capability for dicarboxylate uptake is also widely distributed among SAR11 bacteria: the dicarboxylate transport protein SAR11_1361 shows a broader distribution among SAR11 genomes than the glucose transport protein SAR11_0769 (Fig. 4a, b and Extended Data Fig. 1), and shows a broader geographical distribution in the Tara Oceans metagenome and metatranscriptome datasets, including both coastal and open ocean samples (Fig. 4c–f), despite the fact that SAR11_1361 has a much more limited phylogenetic distribution among bacteria (Extended Data Fig. 10). In the context of uncertainty surrounding the carbon sources that are universal to SAR11 bacteria⁴⁴ (Supplementary Note 7), the identification of an SBP with high affinity ($K_d < 10$ nM) and broad specificity for C₄ and C₅ dicarboxylates that is conserved among SAR11 ecotypes despite stringent genome streamlining, and widely distributed and highly transcribed throughout the ocean, provides strong evidence that these dicarboxylates are physiologically important carbon sources in SAR11 bacteria.

Specificity and affinity of SAR11 SBPs

Systematic characterization of SBPs provided a global view of transporter specificity and affinity in *Ca. P. ubique* HTCC1062, providing broad insight into the physiology of oligotrophic bacteria. It has long been hypothesized that oligotrophic bacteria with streamlined genomes, including SAR11 bacteria, rely on broad-specificity transporters to enable transport of a broad range of substrates with a limited number

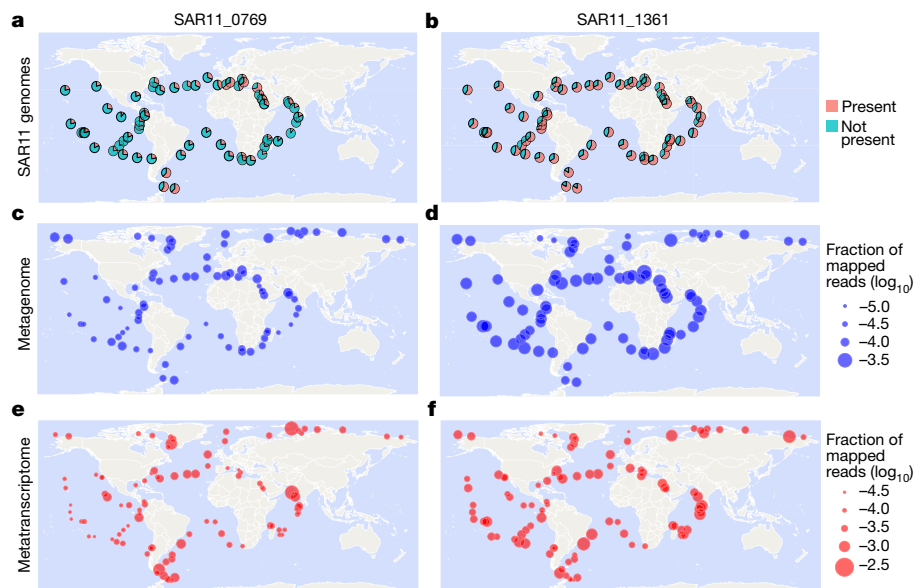


Fig. 4 | The dicarboxylate transport gene SAR11_1361 is abundantly distributed across the surface ocean. a, b. Estimated proportion of SAR11 bacteria containing homologues of SAR11_0769 (**a**) and SAR11_1361 (**b**) in surface stations from the Tara Oceans dataset, based on metagenome recruitment values of 159 SAR11 genomes at each location³⁴ and the presence or absence of an SBP homologue in the corresponding genomes (Methods). The completeness of the genomes used for this analysis was $77 \pm 15\%$ (mean $\pm$ s.d.)³⁴—thus, the given values are underestimates. For comparison, the proportion of SAR11 bacteria belonging to the *Ca. P. ubique* HTCC1062 clade and subclade of SAR11 bacteria at each location is shown in Supplementary Fig. 7. **c–f.** Abundance

of SAR11_0769 and SAR11_1361 homologues in surface metagenome and metatranscriptome datasets from the Tara Oceans project. SAR11_0769 and SAR11_1361 homologues were obtained from the Ocean Gene Atlas v2.0 and filtered using an e-value threshold of 10^{-30} and sequence identity threshold of 40%. Abundance is expressed as the fraction of mapped reads in each sample and is represented by point area on a linear scale. **c.** SAR11_0769 homologues in the Tara Oceans metagenome datasets. **d.** SAR11_1361 homologues in the Tara Oceans metagenome datasets. **e.** SAR11_0769 homologues in the Tara Oceans metatranscriptome datasets. **f.** SAR11_1361 homologues in the Tara Oceans metatranscriptome datasets.

of transporters⁹. Indeed, a broad-specificity osmolyte transporter had previously been identified²², and three more broad-specificity SBPs for amino acids and dicarboxylates were characterized in this work. However, the majority of SBPs (at least 8 out of 13) showed high binding specificity, suggesting a more nuanced view of uptake specificity in oligotrophic bacteria (Supplementary Note 8). Genome streamlining results in reduction of metabolic genes in addition to transporter genes; thus, broad-specificity transporters are associated with a risk of futile uptake of metabolites that cannot be utilized, especially given the high compositional complexity of ocean DOC. Our results show that *Ca. P. ubique* HTCC1062 is highly selective in its substrate uptake, using a small number of broad-specificity transporters mainly for metabolites that can be utilized without dedicated catabolic pathways, including amino acids and TCA cycle intermediates. The remaining transporters show high specificity and mainly cover specific gaps in the broad-specificity transporters; indeed, there is little redundancy in binding specificity between the SBPs, except for some overlap between the two broad-specificity amino acid-binding proteins. The high specificity of these transporters does not result from a negative tradeoff between specificity and affinity; for example, SAR11_0953 is estimated to have nanomolar affinity for around 15 proteinogenic amino acids (on the basis of measured ΔT_m and K_d values), with a maximum of 550 pM for L-glutamate, demonstrating that broad specificity is compatible with high affinity (Supplementary Note 9). Furthermore, three out of the four broad-specificity transporters (SAR11_0953, SAR11_1336 and SAR11_1346) appear to be widely distributed among Proteobacteria, indicating that use of broad-specificity transporters is not unique to oligotrophic bacteria (Extended Data Fig. 8 and Supplementary Fig. 4). Overall, these considerations suggest that oligotrophic bacteria probably show greater selectivity in substrate uptake than previously assumed.

Our results revealed that a systematic increase in SBP binding affinity is a major adaptation of *Ca. P. ubique* HTCC1062 to low substrate

concentrations in the oligotrophic environment. The binding affinity of the *Ca. P. ubique* HTCC1062 SBPs was remarkably high on average, and substantially exceeded the known range of SBP binding affinity in some cases (Fig. 3a). K_d values in the picomolar to low nanomolar range were observed in most cases, in concordance with the picomolar to low nanomolar concentrations of amino acids and other substrates typically observed in the surface oligotrophic ocean^{45,46} (Supplementary Note 3 and Supplementary Table 7) and picomolar to low nanomolar uptake affinities (specifically, $K_s + [S]$, the sum of the half-saturation constant (K_s) and the in situ substrate concentration) for various metabolites in environmental samples from the surface ocean^{45,47}, for which the corresponding transport proteins have generally not been identified²². Although the SBPs may have slightly different properties in vitro compared with their native cellular environment, a strong correlation is generally observed between in vitro properties of SBPs and the in vivo properties of the corresponding transporters²³. In addition, the physiological relevance of the observed binding affinity in SBPs is indicated by several considerations: (1) the observed K_d of 2.0 nM for the interaction between SAR11_1336 and glycine betaine is in excellent agreement with the previously measured K_s value for the corresponding transporter²² (0.89 nM); (2) mathematical models of ABC transporter activity indicate that uptake affinity should be greater than SBP binding affinity when SBP concentration is high⁴⁸ (as in SAR11 bacteria); and (3) the binding affinity of an SBP has physiological significance itself, because it determines the concentration at which substrates can be accumulated in the periplasm⁴⁹. Whereas previous work has shown how extreme selective pressure driven by large population size under low-nutrient conditions has driven systematic adaptation of SAR11 bacteria at the genome and cellular levels (for example, reduction of GC content⁹ and increase in periplasmic volume⁸), this work shows that systematic adaptation of the biophysical properties of SBPs is another important factor in the evolutionary success of SAR11 bacteria. We speculate that the evolutionary tradeoffs

underlying ultra-high binding affinity in SBPs may also be an important factor shaping the physiology of SAR11 and other oligotrophic bacteria (Supplementary Note 10).

The identification of SBPs with unprecedented binding affinity in the genome of *Ca. P. ubique* HTCC1062 resolves uncertainty about the discrepancy between the observed affinity of substrate uptake by microbial communities in the ocean and the binding affinity of previously characterized transporters for substrate uptake. To explain this apparent discrepancy, various alternative mechanisms for high-affinity substrate uptake in oligotrophic bacteria have been proposed. For example, a recent modelling study showed that the uptake affinity of an ABC transporter depends on both SBP concentration and binding affinity, and suggested that oligotrophic bacteria might use high SBP expression to achieve high uptake affinity without increasing SBP binding affinity⁴⁸; by contrast, our results show (without invalidating this model) that high uptake affinity can be explained without accounting for periplasmic SBP concentration. As another example, the observation that the binding affinity of known phosphate-binding proteins (around 1 μ M) is much higher than concentrations of inorganic phosphate in phosphate-depleted regions (less than 5 nM) led to the proposal of an alternative mechanism for accumulation of inorganic phosphate in the periplasm of oligotrophic bacteria^{49,50}. Of note, the phosphate-binding protein of *Ca. P. ubique* HTCC1062 does indeed have relatively low binding affinity (133 nM), which may reflect the challenge of discriminating phosphate from sulfate, which is present at a concentration of around 28 mM in the ocean; although phosphate-binding proteins from sulfate-rich environments can achieve a discrimination factor of greater than 10^5 (ref. 51), there is presumably a biophysical limit on discrimination of these two anions due to their physicochemical similarity. Consistent with the hypothesis that the binding affinity of phosphate-binding proteins may be constrained by the requirement for discrimination of phosphate and sulfate, SAR11_1179 showed a small but significant decrease in apparent binding affinity in the presence of 28 mM sulfate (6.7-fold decrease to 890 nM, $P < 0.0001$, two-tailed t -test on $\log_{10} K_d$ values (Extended Data Fig. 2c)).

Discussion

Systems-level approaches based on metatranscriptomics and related methods are highly valuable for profiling putative biological functions in complex microbial communities across different environments, providing insight into their ecological and biogeochemical functions^{52,53}. However, a limitation of these methods is that they depend on homology-based predictions of protein function, which vary markedly in accuracy between protein families and are usually not validated¹⁹. Here we have shown how targeted functional characterization of environmentally abundant proteins can be integrated with existing multi-omics and physiological data to provide insight over multiple biological scales, ranging from mechanisms of functional adaptation at the molecular level to global patterns of substrate uptake capabilities in SAR11 bacteria. We anticipate that improved computational annotation and continued experimental annotation of protein function will be essential to extract maximum value from increasingly high-resolution ocean microbiome datasets and fulfil the broader goal in microbial ecology of bridging microbial gene function and ocean ecosystems biology on a planetary scale^{54,55}.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-024-07924-w>.

- Morris, R. M. et al. SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**, 806–810 (2002).
- Giovannoni, S. J. SAR11 bacteria: the most abundant plankton in the oceans. *Ann. Rev. Mar. Sci.* **9**, 231–255 (2017).
- Carini, P., White, A. E., Campbell, E. O. & Giovannoni, S. J. Methane production by phosphate-starved SAR11 chemoheterotrophic marine bacteria. *Nat. Commun.* **5**, 4346 (2014).
- Sun, J. et al. The abundant marine bacterium *Pelagibacter* simultaneously catabolizes dimethylsulfoniopropionate to the gases dimethyl sulfide and methanethiol. *Nat. Microbiol.* **1**, 16065 (2016).
- Sowell, S. M. et al. Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. *ISME J.* **3**, 93–105 (2009).
- Sowell, S. M. et al. Proteomic analysis of stationary phase in the marine bacterium “*Candidatus Pelagibacter ubique*”. *Appl. Environ. Microbiol.* **74**, 4091–4100 (2008).
- Schattenhofer, M. et al. Latitudinal distribution of prokaryotic picoplankton populations in the Atlantic Ocean. *Environ. Microbiol.* **11**, 2078–2093 (2009).
- Zhao, X. et al. Three-dimensional structure of the ultraoligotrophic marine bacterium “*Candidatus Pelagibacter ubique*”. *Appl. Environ. Microbiol.* **83**, e02807-16 (2017).
- Giovannoni, S. J. et al. Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**, 1242–1245 (2005).
- Tripp, H. J. et al. Unique glycine-activated riboswitch linked to glycine-serine auxotrophy in SAR11. *Environ. Microbiol.* **11**, 230–238 (2009).
- Tripp, H. J. et al. SAR11 marine bacteria require exogenous reduced sulphur for growth. *Nature* **452**, 741–744 (2008).
- Malmstrom, R. R., Kiene, R. P., Cottrell, M. T. & Kirchman, D. L. Contribution of SAR11 bacteria to dissolved dimethylsulfoniopropionate and amino acid uptake in the North Atlantic Ocean. *Appl. Environ. Microbiol.* **70**, 4129–4135 (2004).
- Malmstrom, R. R., Cottrell, M. T., Elifant, H. & Kirchman, D. L. Biomass production and assimilation of dissolved organic matter by SAR11 bacteria in the Northwest Atlantic Ocean. *Appl. Environ. Microbiol.* **71**, 2979–2986 (2005).
- Clifford, E. L. et al. Taurine is a major carbon and energy source for marine prokaryotes in the North Atlantic Ocean off the Iberian Peninsula. *Microb. Ecol.* **78**, 299–312 (2019).
- Alonso, C. & Pernthaler, J. *Roseobacter* and SAR11 dominate microbial glucose uptake in coastal North Sea waters. *Environ. Microbiol.* **8**, 2022–2030 (2006).
- Moran, M. A. et al. Microbial metabolites in the marine carbon cycle. *Nat. Microbiol.* **7**, 508–523 (2022).
- Davies, J. S. et al. Selective nutrient transport in bacteria: multicomponent transporter systems reign supreme. *Front. Mol. Biosci.* **8**, 699222 (2021).
- Mulligan, C., Fischer, M. & Thomas, G. H. Tripartite ATP-independent periplasmic (TRAP) transporters in bacteria and archaea. *FEMS Microbiol. Rev.* **35**, 68–86 (2011).
- Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.* **5**, e1000605 (2009).
- Schroer, W. F. et al. Functional annotation and importance of marine bacterial transporters of plankton exometabolites. *ISME Commun.* **3**, 37 (2023).
- Ford, B. A. et al. Functional characterisation of substrate-binding proteins to address nutrient uptake in marine picocyanobacteria. *Biochem. Soc. Trans.* **49**, 2465–2481 (2021).
- Noell, S. E. & Giovannoni, S. J. SAR11 bacteria have a high affinity and multifunctional glycine betaine transporter. *Environ. Microbiol.* **21**, 2559–2575 (2019).
- Davidson, A. L., Dassa, E., Orelle, C. & Chen, J. Structure, function, and evolution of bacterial ATP-binding cassette systems. *Microbiol. Mol. Biol. Rev.* **72**, 317–364 (2008).
- Vetting, M. W. et al. Experimental strategies for functional annotation and metabolite discovery: targeted screening of solute binding proteins and unbiased panning of metabolomes. *Biochemistry* **54**, 909–931 (2015).
- Carter, M. S. et al. Functional assignment of multiple catabolic pathways for D-ribose. *Nat. Chem. Biol.* **14**, 696–705 (2018).
- Carini, P. et al. Discovery of a SAR11 growth requirement for thiamin’s pyrimidine precursor and its distribution in the Sargasso Sea. *ISME J.* **8**, 1727–1738 (2014).
- Durham, B. P. et al. Sulfonate-based networks between eukaryotic phytoplankton and heterotrophic bacteria in the surface ocean. *Nat. Microbiol.* **4**, 1706–1715 (2019).
- Carini, P., Steindler, L., Beszteri, S. & Giovannoni, S. J. Nutrient requirements for growth of the extreme oligotroph “*Candidatus Pelagibacter ubique*” HTCC1062 on a defined medium. *ISME J.* **7**, 592–602 (2013).
- Sun, J. et al. One carbon metabolism in SAR11 pelagic marine bacteria. *PLoS ONE* **6**, e23973 (2011).
- Clifton, B. E. et al. Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein. *Nat. Chem. Biol.* **14**, 542–547 (2018).
- Gao, C. et al. Characterization of the trimethylamine N-oxide transporter from *Pelagibacter* strain HTCC1062 reveals its oligotrophic niche adaption. *Front. Microbiol.* **13**, 838608 (2022).
- Jiao, N. & Zheng, Q. The microbial carbon pump: from genes to ecosystems. *Appl. Environ. Microbiol.* **77**, 7439–7444 (2011).
- Vernette, C. et al. The Ocean Gene Atlas v2.0: online exploration of the biogeography and phylogeny of plankton genes. *Nucleic Acids Res.* **50**, W516–W526 (2022).
- Haro-Moreno, J. M. et al. Ecogenomics of the SAR11 clade. *Environ. Microbiol.* **22**, 1748–1763 (2020).
- Zhao, Z., Amano, C., Reinthaler, T., Orellana, M. V. & Herndl, G. J. Substrate uptake patterns shape niche separation in marine prokaryotic microbiome. *Sci. Adv.* **10**, eadn5143 (2024).
- Li, C.-Y. et al. Ubiquitous occurrence of a dimethylsulfoniopropionate ABC transporter in abundant marine bacteria. *ISME J.* **17**, 579–587 (2023).
- Curson, A. R. J., Todd, J. D., Sullivan, M. J. & Johnston, A. W. B. Catabolism of dimethylsulfoniopropionate: microorganisms, enzymes and genes. *Nat. Rev. Microbiol.* **9**, 849–859 (2011).
- Moore, C. M. et al. Processes and patterns of oceanic nutrient limitation. *Nat. Geosci.* **6**, 701–710 (2013).

39. Ustick, L. J. et al. Metagenomic analysis reveals global-scale patterns of ocean nutrient limitation. *Science* **372**, 287–291 (2021).
40. Noell, S. E. et al. SAR11 cells rely on enzyme multifunctionality to metabolize a range of polyamine compounds. *mBio* **12**, e0109121 (2021).
41. Steindler, L., Schwalbach, M. S., Smith, D. P., Chan, F. & Giovannoni, S. J. Energy starved *Candidatus Pelagibacter ubique* substitutes light-mediated ATP production for endogenous carbon respiration. *PLoS ONE* **6**, e19725 (2011).
42. Smith, D. P. et al. Proteomic and transcriptomic analyses of “*Candidatus Pelagibacter ubique*” describe the first PII-independent response to nitrogen limitation in a free-living Alphaproteobacterium. *mBio* **4**, e00133–12 (2013).
43. Smith, D. P. et al. Proteome remodeling in response to sulfur limitation in “*Candidatus Pelagibacter ubique*”. *mSystems* **1**, e00068–16 (2016).
44. Tripp, H. J. The unique metabolism of SAR11 aquatic bacteria. *J. Microbiol.* **51**, 147–153 (2013).
45. Suttle, C. A., Chan, A. M. & Fuhrman, J. A. Dissolved free amino acids in the Sargasso Sea: uptake and respiration rates, turnover times, and concentrations. *Mar. Ecol. Prog. Ser.* **70**, 189–199 (1991).
46. Clifford, E. L. et al. Crustacean zooplankton release copious amounts of dissolved organic matter as taurine in the ocean. *Limnol. Oceanogr.* **62**, 2745–2758 (2017).
47. Kiene, R. P. & Williams, L. P. H. Glycine betaine uptake, retention, and degradation by microorganisms in seawater. *Limnol. Oceanogr.* **43**, 1592–1603 (1998).
48. Norris, N., Levine, N. M., Fernandez, V. I. & Stocker, R. Mechanistic model of nutrient uptake explains dichotomy between marine oligotrophic and copiotrophic bacteria. *PLoS Comput. Biol.* **17**, e1009023 (2021).
49. Kamennaya, N. A., Geraki, K., Scanlan, D. J. & Zubkov, M. V. Accumulation of ambient phosphate into the periplasm of marine bacteria is proton motive force dependent. *Nat. Commun.* **11**, 2642 (2020).
50. Zubkov, M. V., Martin, A. P., Hartmann, M., Grob, C. & Scanlan, D. J. Dominant oceanic bacteria secure phosphate using a large extracellular buffer. *Nat. Commun.* **6**, 7878 (2015).
51. Elias, M. et al. The molecular basis of phosphate discrimination in arsenate-rich environments. *Nature* **491**, 134–137 (2012).
52. Louca, S., Parfrey, L. W. & Doebeli, M. Decoupling function and taxonomy in the global ocean microbiome. *Science* **353**, 1272–1277 (2016).
53. Dinsdale, E. A. et al. Functional metagenomic profiling of nine biomes. *Nature* **452**, 629–632 (2008).
54. Azam, F. & Worden, A. Z. Microbes, molecules, and marine ecosystems. *Science* **303**, 1622–1624 (2004).
55. Worden, A. Z. et al. Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. *Science* **347**, 1257594 (2015).
56. Schwalbach, M. S., Tripp, H. J., Steindler, L., Smith, D. P. & Giovannoni, S. J. The presence of the glycolysis operon in SAR11 genomes is positively correlated with ocean productivity. *Environ. Microbiol.* **12**, 490–500 (2010).
57. Lidbury, I., Murrell, J. C. & Chen, Y. Trimethylamine N-oxide metabolism by abundant marine heterotrophic bacteria. *Proc. Natl Acad. Sci. USA* **111**, 2710–2715 (2014).
58. Peter, M. F. et al. Structural and mechanistic analysis of a tripartite ATP-independent periplasmic TRAP transporter. *Nat. Commun.* **13**, 4471 (2022).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024

Methods

Identification of SBP genes

Nineteen candidate SBP genes in the genome of *Ca. P. ubique* strain HTCC1062 were identified through a search of the TransportDB 2.0 database⁵⁹ (<http://membranetransport.org>; accessed 22 January 2020). One of these genes, SAR11_0371, was annotated as a 'possible trans-membrane receptor' in UniProt and showed a non-canonical predicted domain structure consisting of a short SBP-like domain (170 amino acids) followed by a coiled coil domain and unidentified C-terminal domain. Additionally, genome context analysis showed that, unlike the other ABC SBP genes in *Ca. P. ubique* HTCC1062, SAR11_0371 was not colocalized with genes encoding the membrane permease or ATP-binding cassette components of an ABC transport system. Thus, SAR11_0371 was considered not to represent the SBP component of an SBP-dependent transport system and was excluded from the analysis. We also attempted to identify additional SBP genes through a search of the UniProt database for proteins in *Ca. P. ubique* belonging to Pfam clans CLO177 (PBP; periplasmic binding protein) and CLO144 (Periplas_BP; periplasmic binding protein like); however, this search did not return any additional candidate genes.

Cloning

The protein sequence of each SBP from *Ca. P. ubique* HTCC1062 was obtained from the UniProt database. Signal sequences were predicted using the SignalP 5.0 server⁶⁰ and removed. The protein sequences were then back-translated and codon-optimized for expression in *E. coli*, and the resulting genes were obtained as synthetic DNA from Twist Bioscience or Integrated DNA Technologies. The synthetic genes were cloned into the NdeI/XhoI site of the pET-28a(+) expression vector by In-Fusion cloning using the In-Fusion HD Cloning Kit (Takara Bio), yielding expression constructs with an N-terminal hexahistidine tag and thrombin tag. Correct assembly of each expression vector was confirmed by Sanger sequencing (FASMAC). The putative *csiD* gene, SAR11_1354, and several homologues of the *Ca. P. ubique* HTCC1062 SBPs (Supplementary Table 8) were cloned similarly into the pET-28a(+) vector, except that the thrombin tag was removed from the constructs of SAR11_1354, SAR11_0266 (Fub), or SAR11_1290 (SAR324). The sequences of oligonucleotides and synthetic genes used in this study are listed in Supplementary Table 9.

Optimization of protein expression

Protein expression was initially tested in *E. coli* BL21(DE3) cells grown in Luria-Bertani (LB) and Terrific Broth (TB) media at 30 °C and 17 °C. SAR11_0655 showed optimal soluble expression in LB medium at 17 °C, SAR11_1203 showed optimal soluble expression in TB medium at 30 °C, and 7 proteins (SAR11_0797, SAR11_0807, SAR11_0864, SAR11_1068, SAR11_1179, SAR11_1210, SAR11_1238, and SAR11_1361) showed optimal soluble expression in TB medium at 17 °C. Next, the remaining proteins were tested for expression in *E. coli* SHuffle T7 cells (New England Biolabs) in TB medium at 17 °C; this strain expresses the disulfide bond isomerase DsbC, which can increase soluble recombinant expression of cytoplasmic proteins by promoting correct formation of disulfide bonds. Soluble expression of SAR11_0769, SAR11_0953, SAR11_1302, and SAR11_1336 was achieved under these conditions. Due to the lack of soluble expression for the remaining four proteins (SAR11_0266, SAR11_0271, SAR11_1290 and SAR11_1346), we also tested expression of one or two close homologues of each protein (Supplementary Table 8). The SAR11_0271 homologue from '*Ca. Pelagibacter*' sp. HIMB1321 (denoted SAR11_0271*) could be expressed in soluble form in SHuffle T7 cells in TB medium at 17 °C, while the SAR11_1346 homologue from the same species (denoted SAR11_1346*) could be expressed in soluble form in BL21(DE3) cells in TB medium at 17 °C. SAR11_0271* and SAR11_1346* share 91.4% and 88.9% sequence identity, respectively, with the corresponding proteins from *Ca. P. ubique* HTCC1062, and

the binding site residues are completely conserved (Supplementary Fig. 5), indicating that the functions and properties of the homologous SBPs are likely to be identical. Neither homologue of SAR11_0266 or SAR11_1290 could be expressed in soluble form in BL21(DE3) or SHuffle T7 cells. Expression of SAR11_0266 and SAR11_1290 without His₆ or thrombin tags also yielded insoluble protein.

Protein expression was typically evaluated by SDS-PAGE analysis as follows. Cells transformed with the relevant expression vector by electroporation were spread from a frozen glycerol stock onto an LB agar plate containing 0.2% (w/v) glucose and 25 µg ml⁻¹ kanamycin and incubated at 30 °C overnight. The cells were then scraped into a small volume of LB medium and used to inoculate 3 ml of the relevant growth medium containing 25 µg ml⁻¹ kanamycin in a 10 ml round bottom tube at a starting OD₆₀₀ of 0.05. The culture was incubated at 37 °C with shaking at 220 rpm until the OD₆₀₀ reached 0.5. One-millilitre aliquots were transferred to clean round bottom tubes and isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to a final concentration of 0.5 mM. The induced cultures were incubated with shaking at 220 rpm at 17 °C overnight or 30 °C for 3 h. A 500-µl aliquot of each culture was resuspended in lysis buffer (20 mM Tris, 0.5 M NaCl, 1% (v/v) Triton X-100, pH 8.0) and incubated at room temperature for 10 min. The cell lysate was centrifuged at 21,000g for 5 min (4 °C). The soluble fraction of the cell lysate was transferred to a tube containing 30 µl cCOMPLETE His-Tag purification Ni-NTA resin (Roche) suspended in 500 µl buffer A (8 M urea, 20 mM Tris, 0.5 M NaCl, pH 8.0), while the insoluble fraction of the cell lysate was dissolved in 500 µl buffer A, centrifuged at 21,000g for 5 min, and then transferred to a tube containing 30 µl Ni-NTA resin suspended in 500 µl buffer A. In both cases, the resin was incubated at room temperature for 10 min, washed twice with 500 µl buffer A, and then eluted by incubation with 50 µl buffer B (8 M urea, 20 mM Tris, 0.5 M NaCl, 0.5 M imidazole, pH 8.0) at room temperature for 5 min. Fifteen microliters of supernatant was mixed with 5 µl of 4× SDS-PAGE sample loading buffer and heated at 90 °C for 10 min, then loaded onto a 4–15% pre-cast SDS-PAGE gel (Bio-Rad). The gel was run at 200 V for 30 min and visualized with Coomassie Blue.

Large-scale protein expression and purification

For expression and purification of the *Ca. P. ubique* SBPs, *E. coli* BL21(DE3) or SHuffle T7 cells transformed with the relevant expression vector were spread from a frozen glycerol stock onto an LB agar plate containing 0.2% (w/v) glucose and 25 µg ml⁻¹ kanamycin, and incubated at 30 °C overnight. The cells were then scraped into 3 ml LB medium, and 500 µl of the resulting cell suspension was used to inoculate 500 ml LB or TB medium supplemented with 25 µg ml⁻¹ kanamycin in a 2 l or 3 l flask, preheated at 37 °C. The culture was incubated at 37 °C with shaking at 220 rpm until the OD₆₀₀ reached 0.5, then cooled briefly in an ice-water bath until the temperature reached -25 °C. IPTG was added to a concentration of 0.5 mM, and the culture was incubated at 17 °C with shaking at 220 rpm for a further 16 h. Cells were pelleted by centrifugation (3,300g, 15 min, 4 °C) and frozen at -20 °C until use. For protein purification, cells were thawed on ice, resuspended in 100 ml Ni binding buffer (20 mM Tris, 500 mM NaCl, 20 mM imidazole, pH 8.0), and lysed by sonication. After addition of 500 U Benzonase Nuclease (Sigma-Aldrich) to digest DNA, the cell lysate was centrifuged at 10,000g for 1 h (4 °C). The supernatant was filtered through a 0.45-µm syringe filter and then loaded onto a 1 ml HisTrap HP column (Cytiva) equilibrated with Ni wash buffer using an ÄKTA Pure FPLC system (Cytiva). For purification under native conditions, the column was washed with 10 ml Ni binding buffer followed by 10 ml Ni wash buffer (20 mM Tris, 500 mM NaCl, 44 mM imidazole, pH 8.0), and then the target protein was eluted in 10 ml Ni elution buffer (20 mM Tris, 500 mM NaCl, 500 mM imidazole, pH 8.0). For purification under denaturing conditions, the column was washed with denaturing Ni binding buffer (8 M urea, 20 mM Tris, 250 mM NaCl, 20 mM

imidazole, pH 8.0) at 1 ml min⁻¹ for 30 min after loading of the clarified cell lysate, and the target protein was eluted with 10 ml denaturing Ni elution buffer (8 M urea, 20 mM Tris, 250 mM NaCl, 250 mM imidazole, pH 8.0). Proteins purified under native conditions were concentrated to 400 µl using a 10 kDa molecular weight cut-off (MWCO) Amicon Ultra-4 centrifugal spin concentrator (Merck-Millipore) and purified by size-exclusion chromatography using a Superdex 200 Increase 10/300 column (Cytiva), eluting in DSF buffer (20 mM HEPES, 0.3 M NaCl, pH 7.50). For storage, proteins were concentrated to a volume of 0.5–2 ml and glycerol was added to a concentration of 10% (v/v). The protein was then flash-frozen in 100–200-µl aliquots in liquid nitrogen and stored at –80 °C until use. ArgT from *S. enterica* was expressed from a pETMCSIII plasmid and purified as described previously⁶¹.

Protein refolding

In most cases, protein purified under denaturing conditions was diluted to a concentration of 0.5 mg ml⁻¹ and volume of 10–30 ml in denaturing Ni binding buffer (8 M urea, 20 mM Tris, 250 mM NaCl, 20 mM imidazole, pH 8.0) and transferred to 10 kDa MWCO SnakeSkin dialysis tubing (Thermo Scientific). The protein was then dialysed against 2 l dialysis buffer (20 mM Tris, 150 mM NaCl, pH 8.0) at 4 °C with three buffer changes over a period of 24 h. The protein was collected and exchanged into DSF buffer using a 10 kDa MWCO Amicon Ultra-15 centrifugal concentrator, then concentrated to 400 µl and purified by size-exclusion chromatography as described above. For SAR11_1346*, an improved yield of monomeric protein was obtained using the rapid dilution for refolding: 2 ml of denatured protein (5 mg ml⁻¹ in denaturing Ni binding buffer) was added dropwise with stirring to 40 ml pre-chilled refolding buffer (20 mM Tris, 150 mM NaCl, 10% (v/v) glycerol, pH 8.0) and incubated at 4 °C with stirring for 20 h. The protein was then concentrated and purified by size-exclusion chromatography as above.

Differential scanning fluorimetry

DSF experiments were performed using a StepOnePlus Real-Time PCR System and StepOne software (Applied Biosystems) based on literature protocols^{62,63}. Reaction mixtures were prepared in twin.tec Real-Time PCR Plates (Eppendorf) and contained 5× SYPRO Orange (Sigma-Aldrich), 2.5 µM protein, and 2 µl 10× ligand in a total volume of 20 µl DSF buffer. The plate was sealed with optically clear sealing film and centrifuged at 2,000g for 1 min before loading into the real-time PCR instrument. The temperature was ramped at a rate of 1% (approximately 1.33 °C min⁻¹), typically over a 60 °C window centred on the melting temperature (T_m) of the target protein. Fluorescence was monitored using the ROX channel. T_m values were determined by taking the derivative of fluorescence intensity with respect to temperature and fitting the resulting data to a quadratic equation in a 6 °C window in the vicinity of the T_m in R software.

Proteins were initially screened for binding to metabolites in four Phenotype MicroArray plates, PM1 to PM4 (Biolog). The contents of each well were dissolved in 50 µl (PM1 to PM3) or 20 µl (PM4) sterile filtered water, giving a concentration of approximately 10–20 mM in each well⁶³. The plates were then sealed with aluminium sealing films and stored at –80 °C. Prior to use, the plates were thawed at room temperature and then shaken at 30 °C until the compounds had redissolved. Two microliters of each compound was added to 18 µl reaction mixture prepared as described above. A 2 °C increase in T_m compared with the median value across the plate was taken as indicative of binding^{63,64}.

For screening of individual compounds and confirmatory assays, compounds were dissolved at a concentration of 100 mM in ligand buffer (0.1 M HEPES pH 7.5), and the pH was adjusted with 1 M NaOH or 1 M HCl if necessary (specifically, if the pH of a 10 mM solution of the compound diluted in DSF buffer fell outside the range 6.5–8.0). These stock solutions were stored at –20 °C. Two microlitres of each compound was directly added to 18 µl reaction mixture, giving a final

concentration of 10 mM, or first diluted 10-fold or 100-fold in DSF buffer to give final concentrations of 1 mM or 0.1 mM in the assay. A list of chemicals used for screening, including the supplier and catalogue number, is provided in Supplementary Table 3. Sodium (*R*)- and (*S*)-2,3-dihydroxypropane-1-sulfonate were synthesized from (*R*)- and (*S*)-3-chloro-1,2-propanediol following a literature protocol⁶⁵ and verified by ¹H and ¹³C NMR.

In the case of the TRAP and TTT SBPs, SAR11_0864 and SAR11_1203, we hypothesized that a metal ion might be required for high-affinity binding, due to the biphasic melting curve observed in the presence of isethionate in Biolog screening experiments, suggesting the presence of a mixture of active and inactive protein (SAR11_0864) or due to the discord between the highly charged ligand and the largely uncharged binding site of the SBP (SAR11_1203). Therefore, we tested the effect of the addition of metal ions (Mg²⁺, Ca²⁺, K⁺, Zn²⁺, Mn²⁺, Co²⁺, Ni²⁺, Fe²⁺ and Fe³⁺) on binding of isethionate to SAR11_0864 and citrate to SAR11_1203 by DSF (Supplementary Fig. 6). DSF experiments were performed using refolded protein as described above, with the addition of 1 mM metal ion and 1 mM ligand. Based on these results, and considering the concentration of each metal ion in seawater⁶⁶, 10 mM CaCl₂ (SAR11_0864) or 53 mM MgSO₄ (SAR11_1203) were included in subsequent DSF and ITC binding experiments for these SBPs.

Isothermal titration calorimetry

ITC experiments were performed using a MicroCal PEAQ-ITC system (Malvern Panalytical). Protein samples were refolded and freshly purified (not frozen), and protein and ligand samples were prepared in the same batch of DSF buffer used for size-exclusion chromatography to minimize the heat of dilution. For SAR11_0864 and SAR11_1203, calcium chloride (final concentration 10.3 mM) or magnesium sulfate (final concentration 53 mM), respectively, was added to the protein and ligand samples. Experiments were performed at 25 °C with stirring at 700 rpm and 10 µcal s⁻¹ reference power. Titration parameters were varied depending on the protein yield, the fraction of active protein, and the affinity and enthalpy of the interaction. In a typical titration, 35 µM protein was titrated with 1× 0.4-µl and 19× 1.6-µl injections of ligand, with the ligand concentration chosen to give >1.5-fold molar excess of ligand to active protein at the end of the titration. ITC experiments were generally performed at least in duplicate.

For simple 1:1 binding interactions, the association constant (K_a), enthalpy (ΔH), and stoichiometry (n) of the interaction were determined by fitting the data to the one-set-of-sites model in MicroCal PEAQ-ITC analysis software. In the case of the SAR11_0769 + D-glucose interaction, thermodynamic parameters were estimated through Bayesian fitting to a modified competitive binding model, which incorporated an additional parameter to account for the fraction of the ligand in each anomeric form, and a two-sets-of-sites model implemented in pytc software⁶⁷; the latter model is equivalent to the two-sets-of-sites model in the MicroCal software, except without the minor correction for heat associated with the displaced volume for each injection (for consistency with the other models in pytc). Thermodynamic parameters for the SAR11_0953 + L-glutamate, SAR11_1203 + citrate, SAR11_1210 + L-arginine, SAR11_1336 + glycine betaine, and SAR11_1346* + L-leucine interactions were determined through competitive displacement experiments⁶⁸, in which L-phenylalanine, *cis*-aconitate, D-octopine, glycine, or L-serine (respectively) were included at a fixed concentration in the cell to reduce the apparent binding affinity for the ligand of interest. The data for these competitive binding experiments were analysed by Bayesian fitting to the competitive binding sites model in pytc software. To confirm the high affinity of the SAR11_1210 + L-arginine interaction, a competitive binding experiment was performed where SAR11_1210 and ArgT from *S. enterica* (which has a K_d of 15 nM for L-arginine) were included in the cell together at the same concentration (28 µM) and titrated with L-arginine. Similarly,

for the SAR11_1210(E108A) + L-arginine interaction, a mixture of SAR11_1210(E108A) and SAR11_1210 (35 μ M each) was titrated with L-arginine. For these titrations, the data was fitted to a two-sets-of-sites binding model as described above to obtain thermodynamic parameters for both protein–ligand interactions. For all analyses, the heat of dilution was assumed to be a small constant value and included as a fitted parameter in the model. The validity of this assumption was confirmed for each ligand by performing a control titration where the ligand was injected into DSF buffer.

Spectrophotometric analysis of iron(III) binding

Binding of iron(III) to SAR11_1238 was analysed using a spectrophotometric assay based on literature protocols^{69,70}. UV–vis spectra were recorded at room temperature (25 °C) in a 96-well plate from 300 nm to 630 nm with 1 nm bandwidth using a Multiskan GO spectrophotometer (Thermo Scientific). An initial protein concentration of 100 μ M and an initial volume of 200 μ l were used for all spectrophotometric assays. First, purified SAR11_1238 was thawed and exchanged into 50 mM Tris, 200 mM NaCl buffer (pH 8.0) using a centrifugal concentrator, and the spectrum of the resulting protein sample was recorded. To prepare unliganded protein for iron-binding assays, the protein was exchanged into 50 mM Tris, 200 mM NaCl, 20 mM sodium citrate buffer (pH 8.0) by three rounds of 30-fold dilution and concentration, allowing chelation and removal of the metal ligand. Citrate was then removed by four rounds of 30-fold dilution and concentration with 50 mM Tris, 200 mM NaCl buffer (pH 8.0). Binding assays were performed by titrating the unliganded protein (200 μ l of 100 μ M solution) with 8 $\times$ or 10 $\times$ 5- μ l injections of 800 μ M iron(III) solution, which was prepared from iron(III) chloride and a 2.5-fold molar excess of trisodium citrate (which ensures that the iron(III) remains soluble) in ultrapure water. To confirm that SAR11_1238 binds iron(III) rather than the iron(III)–citrate complex, the protein was also titrated under the same conditions with 800 μ M ammonium iron(II) sulfate; under the aerobic conditions of the assay, iron(II) is rapidly oxidized to iron(III)⁶⁹. UV–vis spectra were recorded 1 min (iron(II)) or 15 min (iron(III)) after each injection. Finally, a competitive binding assay with citrate was used to estimate the affinity of SAR11_1238 for iron(III). The protein was saturated with a twofold molar excess of iron(III) solution, diluted to a volume of 1 ml, and then dialysed against 500 ml of 50 mM Tris, 200 mM NaCl buffer (pH 8.0) at 4 °C overnight to remove excess iron(III) and citrate. The protein was then concentrated to 100 μ M and titrated with 5- μ l injections of 8 twofold serial dilutions of 500 mM sodium citrate (adjusted to pH 8.0 in 50 mM Tris, 200 mM NaCl buffer). The absorbance at 440 nm was recorded 5 min after each addition. The data were fitted to a hyperbolic curve, yielding an apparent K_d of 9.0 mM for citrate. Given that citrate has a K_d of $\sim 10^{-17}$ M for iron(III), this implies that SAR11_1238 has a K_d for iron(III) on the order of $\sim 10^{-19}$ M, similar to previously characterized iron(III)-binding proteins^{70,71}.

X-ray crystallography

For the SAR11_0769/D-glucose and SAR11_1210/L-arginine structures, the proteins were first expressed and purified by nickel affinity chromatography under native conditions as described above. After addition of a 20-fold molar excess of D-glucose (SAR11_0769) or L-arginine (SAR11_1210), the protein was purified further by size-exclusion chromatography on a HiLoad 26/600 Superdex 75 pg column (Cytiva), eluting in 3 $\times$ crystallization buffer (60 mM HEPES, 150 mM NaCl, pH 7.5). Fractions containing the target protein were collected, and D-glucose (SAR11_0769) or L-arginine (SAR11_1210) was added to a concentration of 30 μ M. The protein was concentrated to a volume of $\sim$ 500 μ l, diluted threefold in water to reduce the NaCl concentration to 50 mM, and then concentrated further to 12 mg ml⁻¹. For the SAR11_0769/D-galactose and SAR11_0655/L-pyroglutamate structures, the proteins were expressed and purified in the same way, except that no ligands were added. Protein crystals were obtained using the vapour diffusion method

in hanging drops at 20 °C, then cryoprotected and flash-frozen in liquid nitrogen. Crystallization and cryoprotection conditions for each protein are given in Supplementary Methods. X-ray diffraction data were collected on beamline BL32XU at the SPring-8 synchrotron (Harima, Japan), using the ZOO suite for automated data collection⁷². The data were automatically indexed, integrated, scaled and merged in XDS⁷³ using KAMO⁷⁴. The structure was solved by molecular replacement in Phaser⁷⁵ or MOLREP⁷⁶. For SAR11_1210, the structure of an opine-binding protein from *Agrobacterium fabrum* (PDB ID 5OT8) was used as a search model; in the remaining cases, an AlphaFold2 model was used⁷⁷. The structures were then refined by iterative real-space and reciprocal-space refinement in REFMAC⁷⁸, Phenix⁷⁹, and COOT⁸⁰. Data collection and refinement statistics are given in Supplementary Table 10 and Supplementary Table 11. Structures were visualized in Pymol.

Gas chromatography–mass spectrometry

SBPs purified under native conditions were exchanged into 200 mM ammonium acetate using a PD-10 desalting column (Cytiva) and concentrated to $\sim$ 1 mM. A 10-nmol aliquot of protein was mixed with 10 μ l of 300 μ M α -methylglucopyranoside (as an internal control) and 200 μ l methanol. The mixture was agitated at 1500 rpm at 24 °C for 10 min and then centrifuged at 21,000g for 20 min at 4 °C. The supernatant was evaporated to dryness using a vacuum evaporator, redissolved in 20 μ l anhydrous pyridine, and derivatized by addition of 30 μ l *N*-methyl-*N*-(trimethylsilyl)trifluoroacetamide (MSTFA) containing 1% trimethylchlorosilane (Supelco) followed by incubation at 70 °C for 1 h. In the case of SAR11_1361, the dried sample was instead dissolved in 20 μ l of 20 mg ml⁻¹ methoxyamine hydrochloride in anhydrous pyridine and incubated at 37 °C for 90 min with agitation at 750 rpm before addition of the MSTFA mixture. The derivatized samples were injected immediately onto an Agilent 7890 A GC System (Agilent Technologies) equipped with a PAL COMBI-XT autosampler (CTC Analytics) and connected to a PEGASUS 4D GC $\times$ GC TOF-MS instrument (LECO) operating in one-dimensional mode. The GC was fitted with a DB-IMS column (Agilent Technologies) with 30 m length, 0.25 mm internal diameter, and 0.25 μ m film thickness. The instrument was operated in pulsed split mode with a split ratio of 2 and injection volume of 1 μ l. The inlet temperature was 250 °C. Helium was used as the carrier gas with a flow rate of 1 ml min⁻¹. The GC oven temperature was held at 70 °C for 5 min, then raised at 12 °C min⁻¹ to 300 °C, and finally held at 300 °C for 10 min. Mass spectrometry data were collected from 50 to 500 *m/z* after a 6.5-min solvent delay. The ion source and transfer line temperatures were 250 °C and the ionization energy was 70 eV. Data analysis and spectral database searches against the NIST database were performed using ChromaTOF software (LECO). Protein-derived samples were analysed before control samples to prevent carryover.

Biogeographical analysis

Biogeographical analysis was performed using the Ocean Gene Atlas v2.0 server³³. Abundance data for each SBP gene from *Ca. P. ubique* HTCC1062 in the Tara Oceans OM-RGC_v2_metaG and OM-RGC_v2_metaT datasets was obtained through a BLAST search with a stringent *e*-value threshold of 10^{-30} . To avoid inclusion of homologous SBPs with different transport functions, hits with a sequence identity of less than 40% (for ABC SBPs) or 55% (for TRAP and TTT SBPs) compared with the corresponding HTCC1062 SBP were excluded from the analysis.

To estimate the total abundance of SBP transcripts, abundance data for each of the 38 PFAM families in CL0177 (PBP; periplasmic binding protein) and CL0144 (Periplas_BP; periplasmic binding protein like), excluding the transferrin family (PF00405) and any families that contain solely enzymes or transcription factors (PF00800, PF01379, PF01634, PF02621, PF03466, PF09084), were obtained using a hmmer

Article

search of the OM-RGC_v2_metaT dataset with an e-value threshold of 10^{-10} . Hits were obtained for 26 out of 31 PFAM families. For each PFAM family, the corresponding hidden Markov model (HMM) was obtained from the InterPro database⁸¹. The protein sequences from the hmmer search were then aligned to this HMM using hmalign and used to construct a new HMM using hmmbuild in HMMER3.4 (<http://hmmer.org>). A second hmmer search of the OM-RGC_v2_metaT dataset, with a lower e-value threshold of 10^{-5} , was then conducted using the resulting HMM. The hits from all 52 searches were combined and redundant hits were removed, resulting in a total of 211,222 unique SBP genes. The two-step search recovered 94% of the 23,879 genes identified as homologues of the *Ca. P. ubique* HTCC1062 SBPs in the BLAST analysis before application of a sequence identity threshold; the remaining 1267 genes were also added to the list of SBP genes. Finally, the total abundance of SBP genes at each site was calculated.

To estimate the percentage of SAR11 bacteria at a site containing a given SBP from *Ca. P. ubique* HTCC1062, we used the recruitment values of 159 SAR11 genomes in the Tara Ocean metagenome dataset calculated by Haro-Moreno et al.³⁴. The presence of a homologue of each SBP in each of the corresponding genomes was determined by BLAST using a 50% sequence identity and 50% coverage threshold. The relative abundance of SAR11 bacteria containing a given SBP homologue was then calculated for each station. Plots were generated using R and GraphPad Prism.

Phylogenetic analysis

Protein sequences homologous to the SBP of interest were identified via a BLAST search of the UniProtKB Reference Proteomes and Swiss-Prot databases⁸². The resulting sequences were filtered to remove a small number of unusually long sequences (>20% greater than mean length) and aligned in MUSCLE v3.8.31⁸³. The alignment was trimmed in trimAl v1.2 using the automated1 option⁸⁴ and then used to generate a maximum-likelihood phylogeny in FastTree v2.1.11, using LG + Γ_{20} as the substitution model⁸⁵. For each protein sequence in the tree, the fraction of conserved binding site residues, compared with the corresponding protein from *Ca. P. ubique* HTCC1062, was estimated. The binding site residues were obtained from the crystal structure (SAR11_0769) or estimated from an AlphaFold2 model^{86,87}. For this analysis, the following substitutions were treated as conservative: S/T, I/M, V/L, I/V, L/M, D/E, Q/N, A/V, F/Y, Y/W, F/W. Phylogenetic tree figures were generated using the ggtree package in R⁸⁸. Figures showing taxonomic distribution (Extended Data Fig. 8b) were generated using Krona⁸⁹.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Coordinates and structure factors for each crystal structure have been deposited in the Protein Data Bank under the following accession codes: SAR11_0655–L-pyroglytamate (8WCH), SAR11_0769–D-glucose (8HQQ), SAR11_0769–D-galactose (8KD0), SAR11_1210–L-arginine (8HQR). Experimental protein structures referenced in the text are available from the Protein Data Bank under accession codes 1EU8, 1GLG, 2B3B, 2B3F, 2FVY, 2PFY, 2PFZ, 2Q2A, 3OO6, 3ZKK, 4R2B, 4UAB, 4UA8, 5DVI, 5DVJ, 6WGM. Structural models are available from the AlphaFold Protein Structure Database using the following UniProt accessions: SAR11_0266, Q4FP02; SAR11_0271, Q4FNZ7; SAR11_0655, Q4FMW4; SAR11_0769, Q4FMK2; SAR11_0953, Q4FM26; SAR11_1290, Q4FL44; SAR11_1336, Q4FKZ8; SAR11_1346, Q4FKY9; SAR11_1361, Q4FKX4; SAR11_0271*, A0A1X7H7N5; SAR11_1346*, A0A1X7H1Y9; A2cp1_3084, B8JG16. Raw ITC data, DSF data for Biolog assays, GC–MS data, phylogenetic and biogeographical data, and source data for figures are available via the Open Science Framework (<https://doi.org/10.17605/OSF.IO/47TR5>).

Code availability

Scripts used for data analysis are available via the Open Science Framework (<https://doi.org/10.17605/OSF.IO/47TR5>).

59. Elbourne, L. D. H., Tetu, S. G., Hassan, K. A. & Paulsen, I. T. TransportDB 2.0: a database for exploring membrane transporters in sequenced genomes from all domains of life. *Nucleic Acids Res.* **45**, D320–D324 (2017).
60. Armenteros, J. J. A. et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019).
61. Clifton, B. E. & Jackson, C. J. Ancestral protein reconstruction yields insights into adaptive evolution of binding specificity in solute-binding proteins. *Cell Chem. Biol.* **23**, 236–245 (2016).
62. McKellar, J. L. O., Minnell, J. J. & Gerth, M. L. A high-throughput screen for ligand binding reveals the specificities of three amino acid chemoreceptors from *Pseudomonas syringae* pv. *actinidiae*. *Mol. Microbiol.* **96**, 694–707 (2015).
63. Ehrhardt, M. K. G., Warring, S. L. & Gerth, M. L. Screening chemoreceptor-ligand interactions by high-throughput thermal-shift assays. *Methods Mol. Biol.* **1729**, 281–290 (2018).
64. Fernández, M. et al. High-throughput screening to identify chemoreceptor ligands. *Methods Mol. Biol.* **1729**, 291–301 (2018).
65. Chen, X. et al. Metabolism of chiral sulfonate compound 2,3-dihydroxypropane-1-sulfonate (DHPS) by *Roseobacter* bacteria in marine environment. *Environ. Int.* **157**, 106829 (2021).
66. Pilson, M. E. Q. *An Introduction to the Chemistry of the Sea* (Cambridge Univ. Press, 2013).
67. Duvvuri, H., Wheeler, L. C. & Harms, M. J. pytc: open-source Python software for global analyses of isothermal titration calorimetry data. *Biochemistry* **57**, 2578–2583 (2018).
68. Velazquez-Campoy, A. & Freire, E. Isothermal titration calorimetry to determine association constants for high-affinity ligands. *Nat. Protoc.* **1**, 186–191 (2006).
69. Badarau, A. et al. FutA2 is a ferric binding protein from *Synechocystis* PCC 6803. *J. Biol. Chem.* **283**, 12520–12527 (2008).
70. Koropatkin, N., Randich, A. M., Bhattacharyya-Pakrasi, M., Pakrasi, H. B. & Smith, T. J. The structure of the iron-binding protein, FutA1, from *Synechocystis* 6803. *J. Biol. Chem.* **282**, 27468–27477 (2007).
71. Chen, C. Y., Berish, S. A., Morse, S. A. & Mietzner, T. A. The ferric iron-binding protein of pathogenic *Neisseria* spp. functions as a periplasmic transport protein in iron acquisition from human transferrin. *Mol. Microbiol.* **10**, 311–318 (1993).
72. Hirata, K. et al. ZOO: an automatic data-collection system for high-throughput structure analysis in protein microcrystallography. *Acta Crystallogr. D* **75**, 138–150 (2019).
73. Kabsch, W. XDS. *Acta Crystallogr. D* **66**, 125–132 (2010).
74. Yamashita, K., Hirata, K. & Yamamoto, M. KAMO: towards automated data processing for microcrystals. *Acta Crystallogr. D* **74**, 441–449 (2018).
75. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
76. Vagin, A. & Teplyakov, A. *molrep*: an automated program for molecular replacement. *J. Appl. Crystallogr.* **30**, 1022–1025 (1997).
77. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
78. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
79. Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr. D* **75**, 861–877 (2019).
80. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
81. Paysan-Lafosse, T. et al. InterPro in 2022. *Nucleic Acids Res.* **51**, D418–D427 (2023).
82. UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
83. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
84. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
85. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
86. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
87. Varadi, M. et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
88. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. Y. ggtree: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2016).
89. Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a Web browser. *ELMBioinf.* **12**, 385 (2011).
90. Niehaus, T. D., Elbadawi-Sidhu, M., de Crécy-Lagard, V., Fiehn, O. & Hanson, A. D. Discovery of a widespread prokaryotic 5-oxoprolinase that was hiding in plain sight. *J. Biol. Chem.* **292**, 16360–16367 (2017).
91. Knorr, S. et al. Widespread bacterial lysine degradation proceeding via glutarate and L-2-hydroxyglutarate. *Nat. Commun.* **9**, 5071 (2018).

Acknowledgements B.E.C. was supported by a JSPS Postdoctoral Fellowship for Overseas Researchers from the Japan Society for the Promotion of Science. P.L. gratefully acknowledges funding from the Okinawa Institute of Science and Technology. U.A. gratefully acknowledges funding from Alon Scholarships (The Council for Higher Education, Israel). The synchrotron radiation experiments were performed at BL32XU of SPring-8 with the approval of the Japan

Synchrotron Radiation Research Institute (JASRI) (proposal no. 2023A2731). We thank A. Vardi for critical reading of the initial manuscript; D. Kozome for assistance with X-ray data collection; P. Jain for assistance with synthesis and NMR; Y. Iinuma and O. Smith for assistance with GC-MS; and the Instrumental Analysis Section and Sequencing Section at OIST for providing instrument access and training.

Author contributions B.E.C. and P.L. conceived the study. B.E.C. and G.-I.U. performed cloning, protein expression and purification. B.E.C. performed the remaining experimental and computational work; B.E.C., U.A., C.J.J. and P.L. analysed the data and wrote the paper. P.L. supervised the project.

Competing interests The authors declare no competing interests.

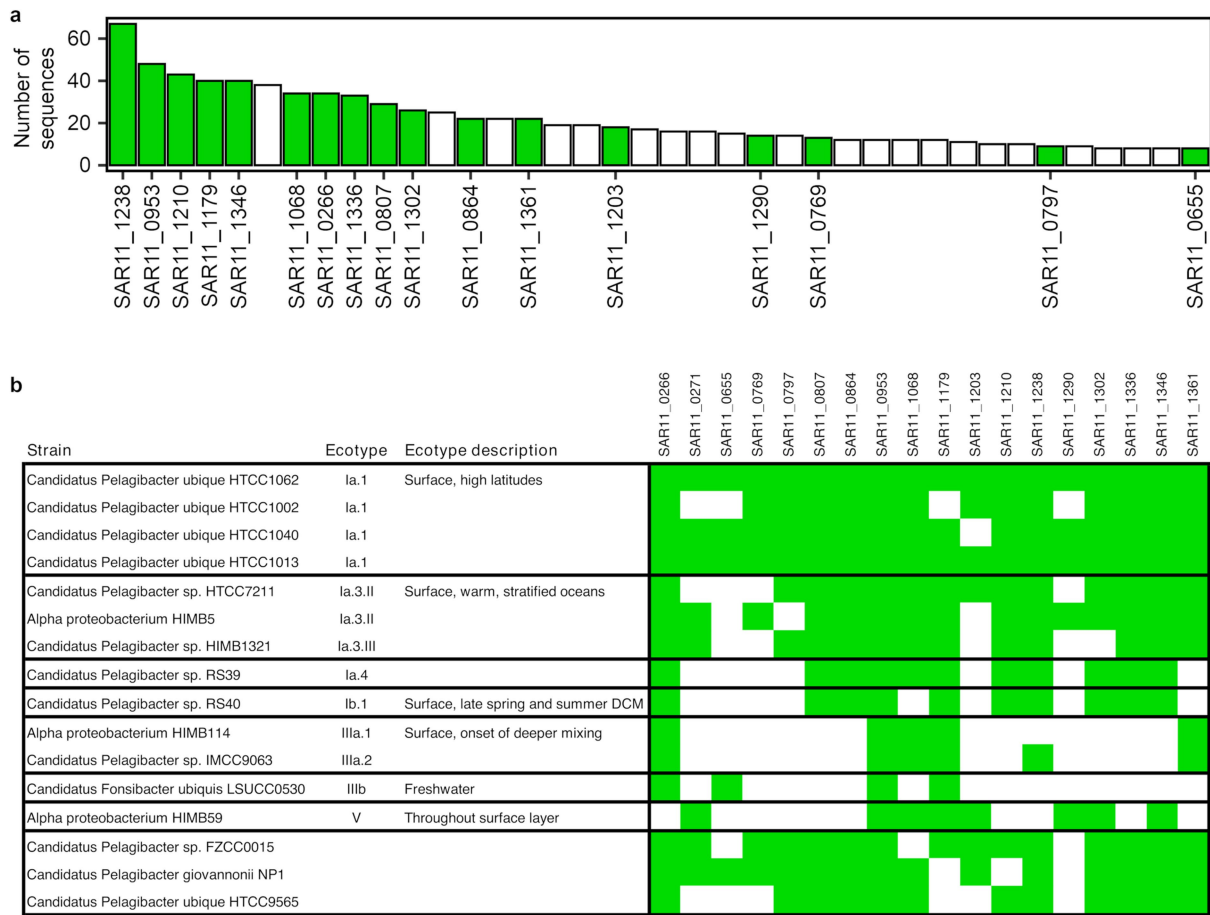
Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-024-07924-w>.

Correspondence and requests for materials should be addressed to Ben E. Clifton or Paola Laurino.

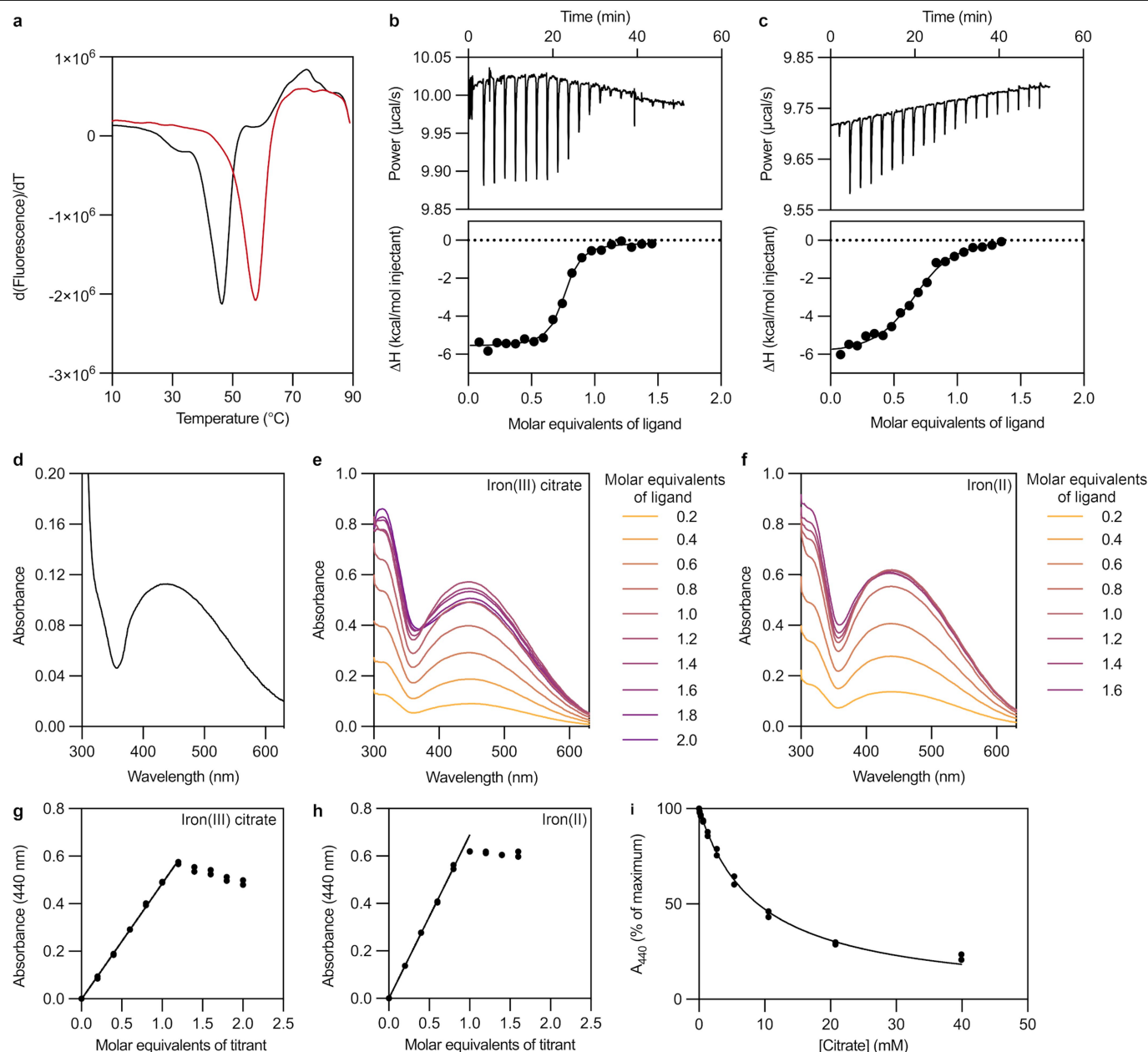
Peer review information *Nature* thanks Gavin Thomas and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



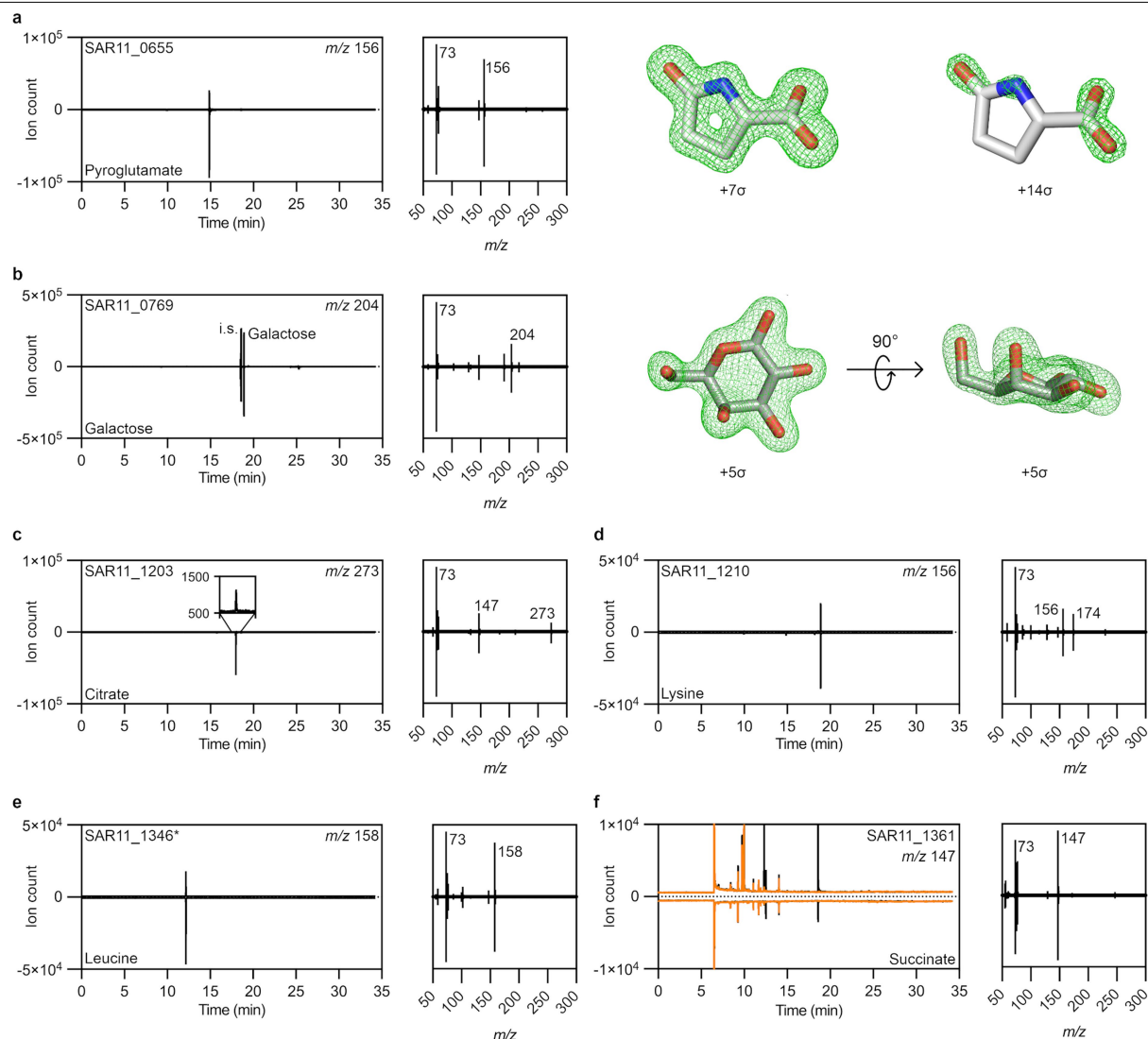
Extended Data Fig. 1 | Distribution of *Ca. P. ubique* HTCC1062 SBPs among SAR11 bacteria. (a) SBP sequences belonging to the SBP PFAM clans CL0177 and CL0144 from candidate order ‘*Pelagibacterales*’ were obtained from the InterPro database and clustered at 35% sequence identity. The graph shows the number of sequences in each SBP cluster containing ≥ 8 sequences. Clusters that are represented in *Ca. P. ubique* HTCC1062 are shown in green.

(b) Distribution of homologs of SBPs from *Ca. P. ubique* HTCC1062 in SAR11 strains with complete genome sequences. Homologs sharing $< 50\%$ sequence identity with the corresponding protein in *Ca. P. ubique* HTCC1062 are not shown. Ecotype classifications are taken from ref. 34. and descriptions are taken verbatim from ref. 2.



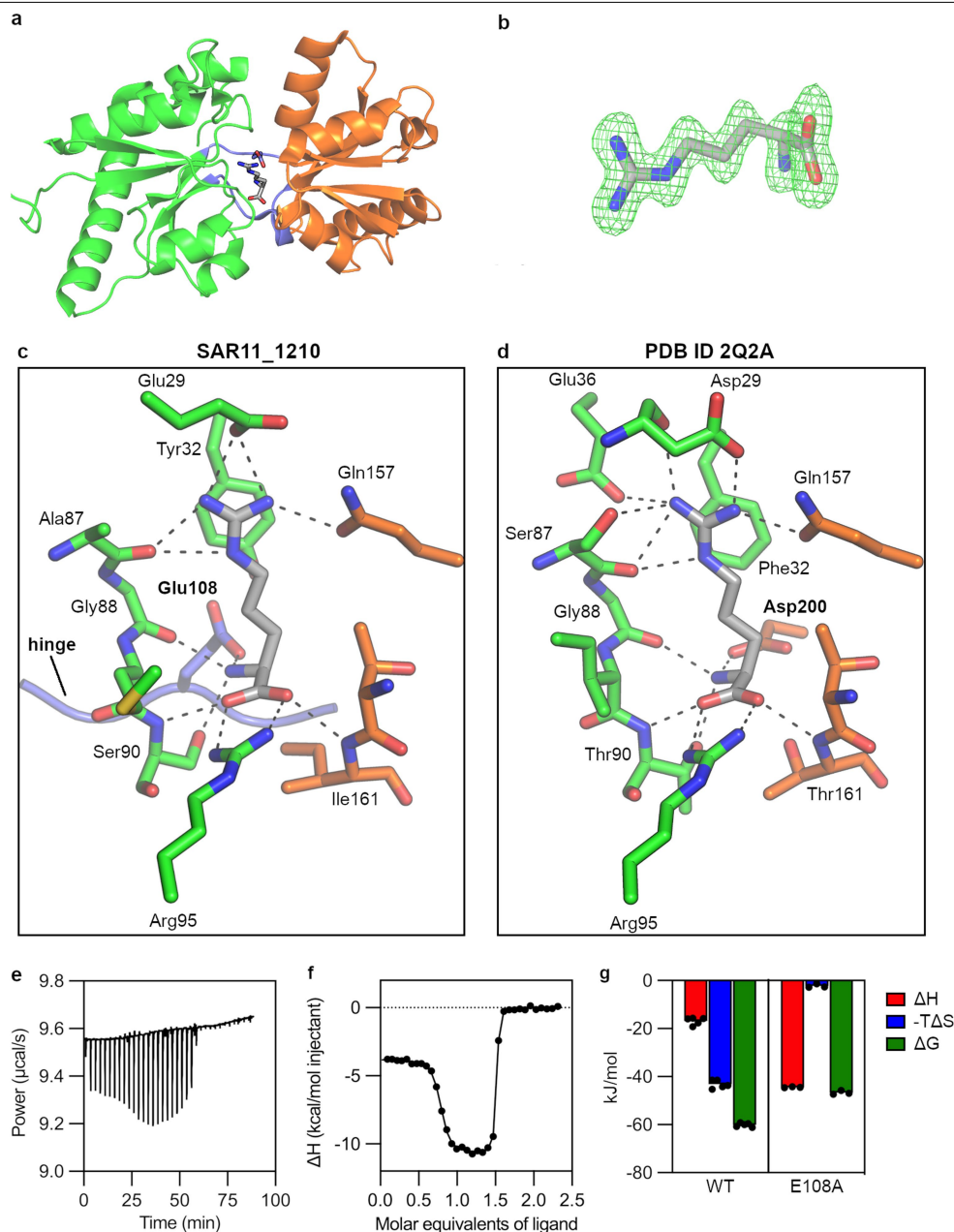
Extended Data Fig. 2 | The SBPs SAR11_1179 and SAR11_1238 are phosphate- and iron(III)-binding proteins. (a) Thermal denaturation of SAR11_1179 in the absence (black) or presence (red) of $1\text{ mM Na}_2\text{HPO}_4$ measured by DSF ($\Delta T_m = 10.9 \pm 0.4\text{ }^{\circ}\text{C}$, mean $\pm$ s.d., $n = 3$ technical replicates). (b-c) Representative ITC data for titration of $23\text{ }\mu\text{M SAR11_1179}$ with $185\text{ }\mu\text{M Na}_2\text{HPO}_4$ in the absence (b) or presence (c) of $28\text{ mM Na}_2\text{SO}_4$. Fitting the data to the one-set-of-sites model gave a K_d of $133 \pm 28\text{ nM}$ in the absence of sulfate and $892 \pm 122\text{ nM}$ in the presence of 28 mM sulfate (mean $\pm$ s.d., $n = 3$ or 4 replicate titrations; significantly different by two-tailed t -test on $\log_{10} K_d$ values; $P = 3.03 \times 10^{-5}$, $t = 14.28$, $df = 5$, difference between means = 0.831 , 95% confidence interval = 0.681 to 0.981). (d) UV-visible spectrum of SAR11_1238 purified from *E. coli*,

without addition of ligand, showing presence of endogenously bound iron(III). (e-f) UV-visible spectra of unliganded SAR11_1238 titrated with iron(III) delivered as (e) iron(III) citrate or (f) ammonium iron(II) sulfate (see Methods section for further explanation). (g-h) Titration of SAR11_1238 with iron(III) delivered as (g) iron(III) citrate or (h) ammonium iron(II) sulfate, monitored by absorbance at 440 nm . Discrete data points from four (g) or two (h) technical replicates (independent titrations) are shown. The line represents a fit to the linear portion of the titration. (i) Competitive titration of iron(III)-bound SAR11_1238 with citrate, monitored by absorbance at 440 nm . Results from two technical replicates (independent titrations) are shown as discrete data points.



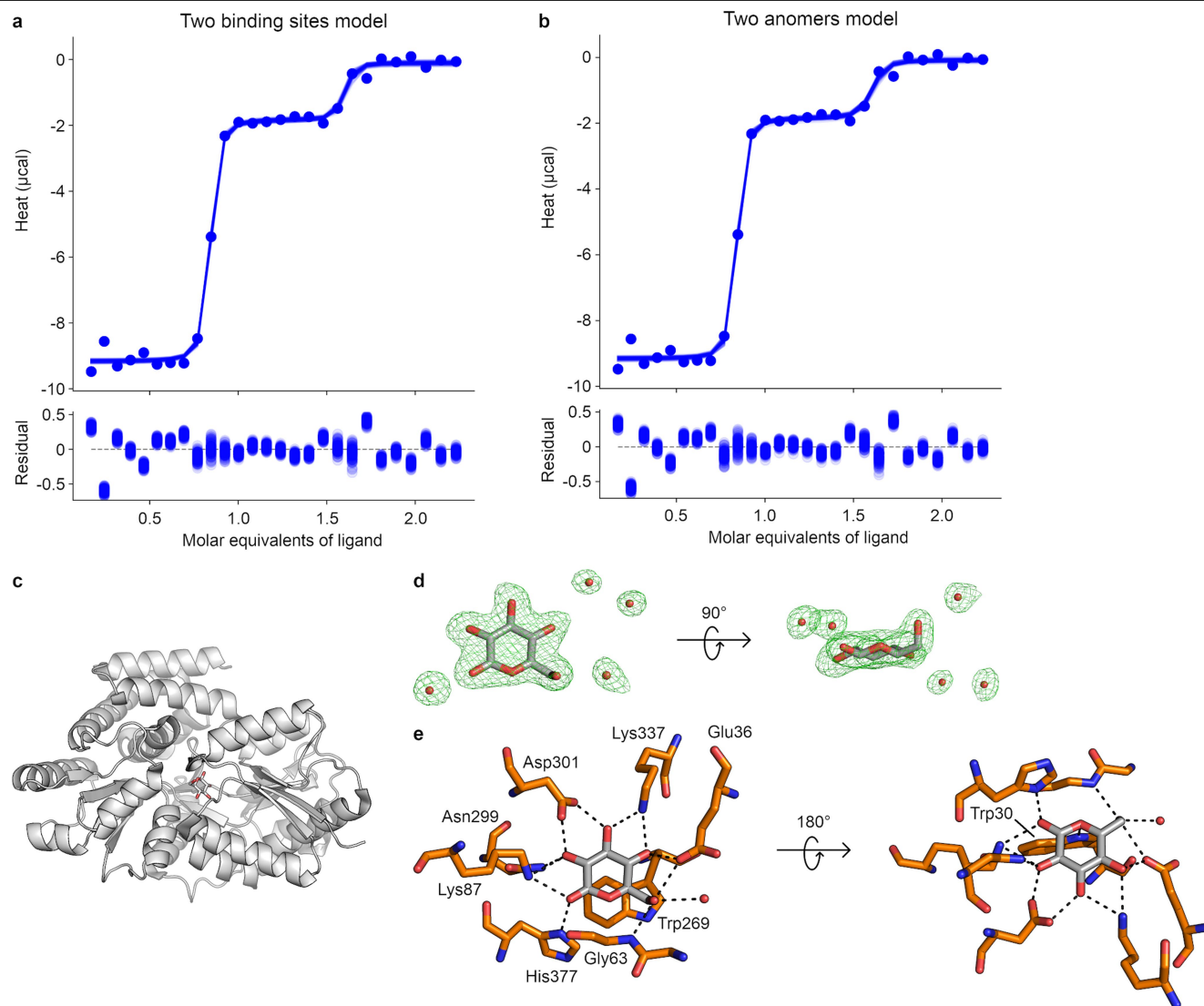
Extended Data Fig. 3 | Identification of co-purified SBP ligands by GC-MS and X-ray crystallography. For each protein, the left-hand graph shows an extracted ion chromatogram of the protein sample (i.e., the co-purified ligand extracted from the protein and subjected to trimethylsilyl derivatization) (top) and the corresponding chemical standard (bottom, reflected). The right-hand graph shows the mass spectrum of the major chromatographic peak for the protein sample (top) and the corresponding standard (bottom, reflected). i.s., internal standard (α -methylglucoside). For SAR11_1361, due to the lack of a distinctive ion in the mass spectrum of bis(trimethylsilyl) succinate, the

presence of succinate is illustrated by comparison with a blank run (reagent only) shown in orange. The three additional peaks in the protein sample correspond to glycerol (from protein purification), succinate, and α -methylglucoside (internal standard), in that order. For SAR11_0655 (**a**) and SAR11_0769 (**b**), further evidence for the identity of the co-purified ligand was attained from the crystal structure of the protein obtained without addition of exogenous ligand. Electron density for the co-purified ligand is shown by $mF_o - dF_c$ omit maps contoured at various levels (in the case of SAR11_0655, to distinguish between carbon atoms and oxygen/nitrogen atoms).



Extended Data Fig. 4 | Non-canonical binding mode of L-arginine to SAR11_1210. (a-c) Crystal structure of SAR11_1210 complexed with L-arginine (1.32 Å). The large domain, small domain, and hinge regions are shown in green, orange, and purple, respectively. (a) Overall structure. (b) Electron density for the L-arginine molecule, shown by an $mF_o - dF_c$ omit map contoured at $+3\sigma$. (c-d) Comparison of binding modes of L-arginine to (c) SAR11_1210 and (d) a homologous lysine-/arginine-/ornithine-binding protein from *Geobacillus stearothermophilus* (K_d 39 nM, PDB ID 2Q2A), which shows the amino acid

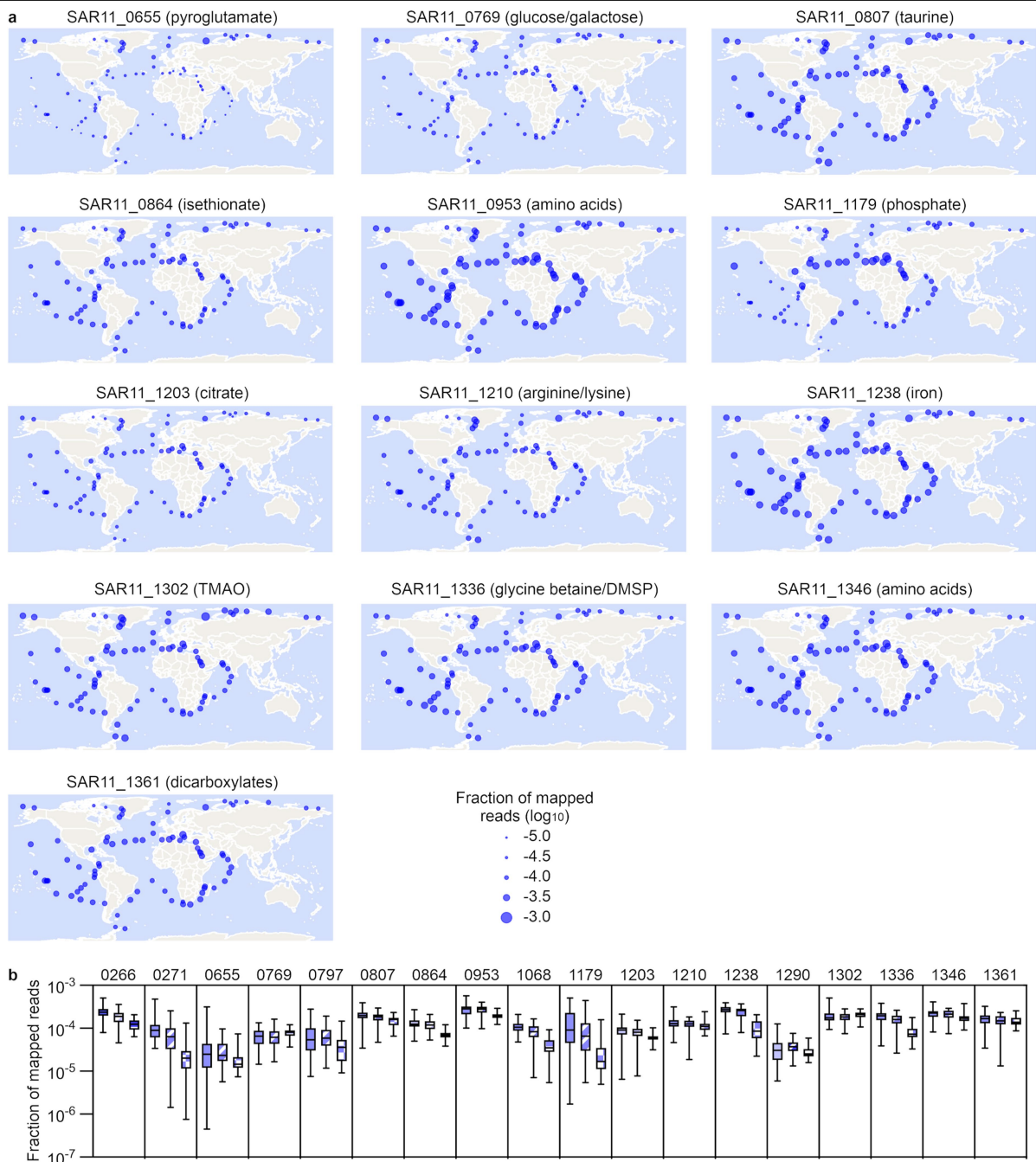
binding motif typical of this SBP family. Residues are numbered according to the homologous positions in SAR11_1210. (e-g) Effect of E108A substitution on binding affinity of SAR11_1210 for L-arginine. (e, f) Competitive ITC titration of an equimolar mixture of SAR11_1210 and SAR11_1210(E108A) with L-arginine. (g) Thermodynamic parameters for interaction of L-arginine with the WT and E108A variants of SAR11_1210. Bars represent mean of technical replicates (separate titrations), shown as individual data points (WT, $n = 5$; E108A, $n = 3$).



Extended Data Fig. 5 | Binding mode of β -D-glucose to SAR11_0769.

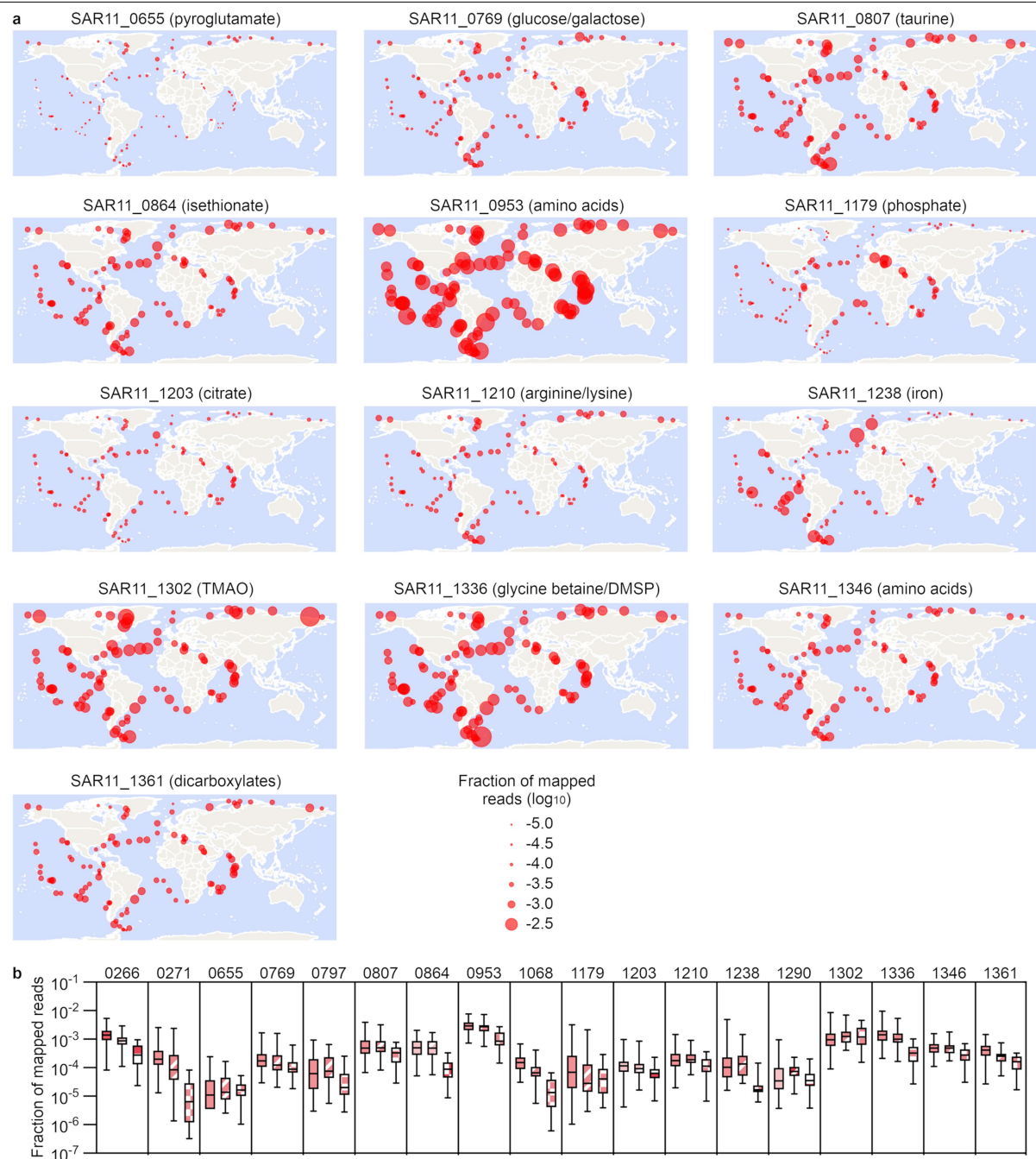
(a-b) Representative ITC data for titration of SAR11_0769 with D-glucose, fitted to (a) the two-sets-of-sites binding model, or (b) a competitive binding model accounting for the two anomeric forms of D-glucose. (c-e) Crystal structure of SAR11_0769 complexed with β -D-glucose (1.86 Å). (c) Overall structure.

(d) Electron density for the β -D-glucose molecule and neighboring water molecules, shown by an $mF_o - dF_c$ omit map contoured at $+3\sigma$. The density for the anomeric hydroxyl group is clearly resolved. (e) Binding mode of β -D-glucose.



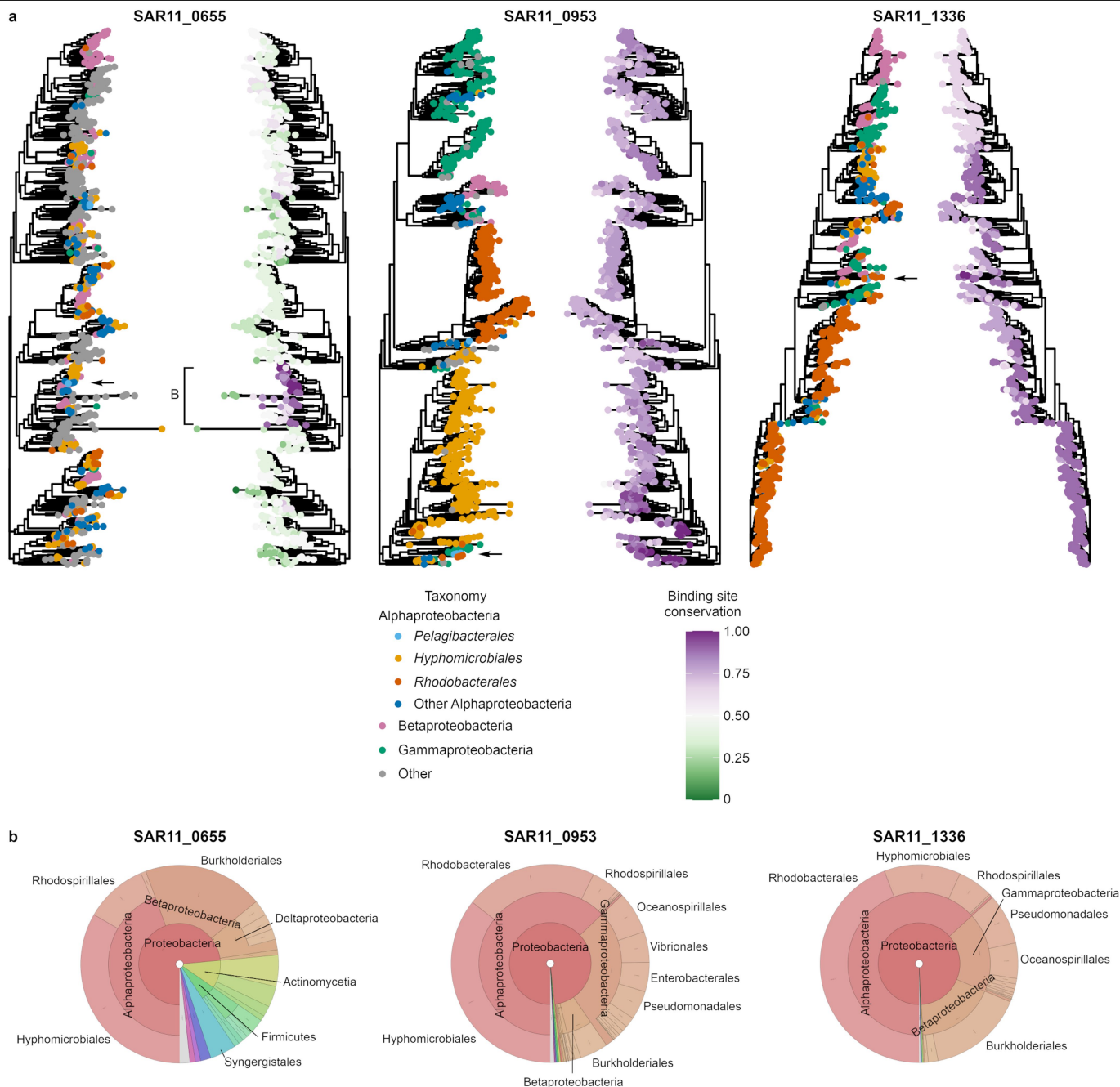
Extended Data Fig. 6 | Abundance of SBP genes from SAR11 (*Ca. P. ubique* HTCC1062) in the global ocean metagenome. (a) The abundance of each gene in surface samples from the *Tara* Oceans OM-RGCv2+G metagenomic dataset is shown. Data were obtained from the Ocean Gene Atlas v2.0 using an e-value cut-off of 10^{-30} and filtered using a sequence identity threshold of 40% for ABC SBPs and 55% for TRAP and TTT SBPs. Abundance at each location is expressed as the fraction of mapped reads and represented by point area on a

linear scale. (b) Box-and-whisker plots comparing abundance of each SBP in epipelagic/surface (SRF, $n = 83$, solid color), deep chlorophyll maximum (DCM, $n = 53$, stripes), and mesopelagic (MES, $n = 38$, checkerboard) samples from the OM-RGCv2+G metagenome dataset (center line, median; box limits, 25th and 75th percentiles; whiskers, maximum and minimum). Statistical comparisons are given in Supplementary Table 12.



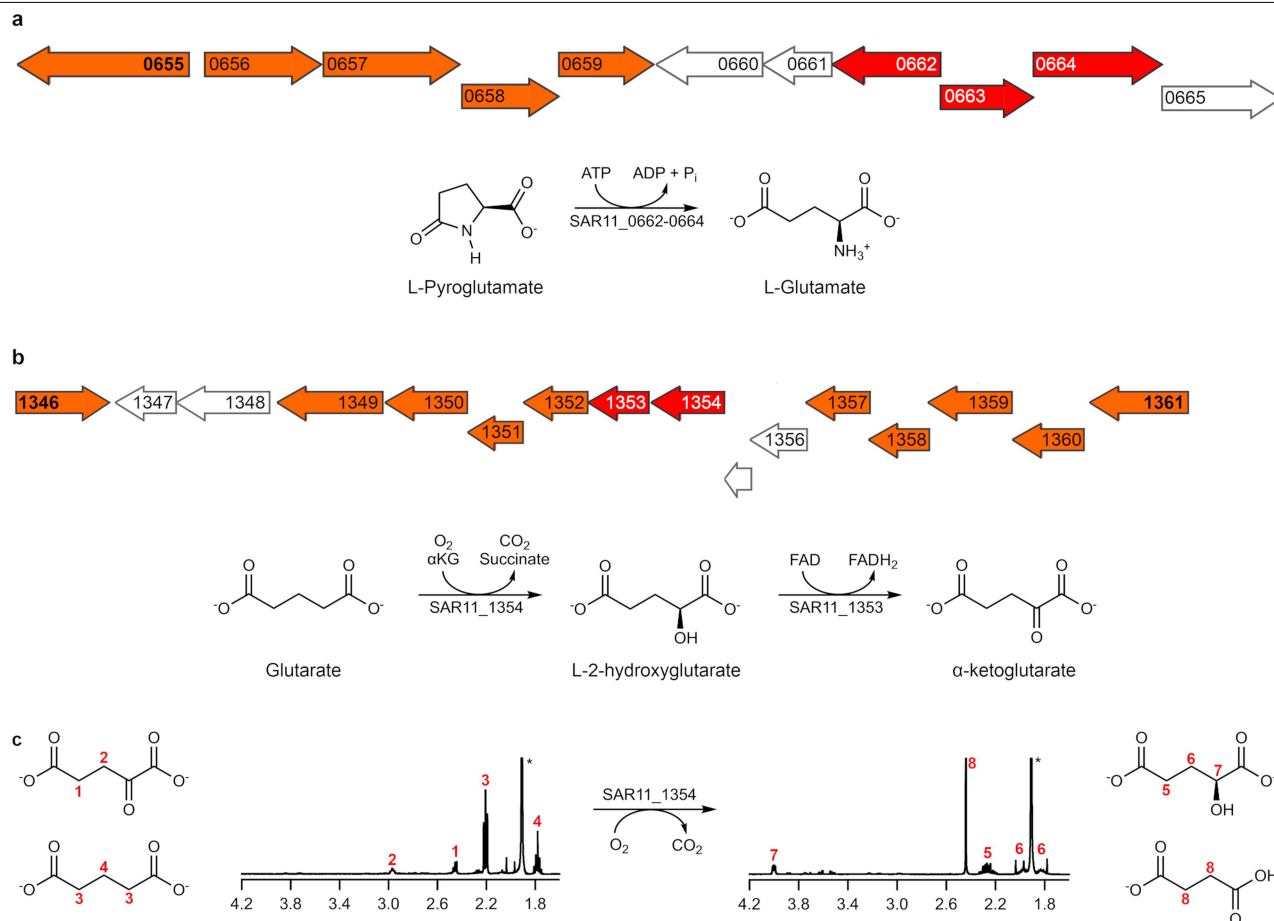
Extended Data Fig. 7 | Abundance of SBP genes from SAR11 (*Ca. P. ubique* HTCC1062) in the global ocean metatranscriptome. (a) The abundance of each gene in surface samples from the *Tara* Oceans OM-RGCv2+T metatranscriptomic dataset is shown. Data were obtained from the Ocean Gene Atlas v2.0 using an e-value cut-off of 10^{-30} and filtered using a sequence identity threshold of 40% for ABC SBPs and 55% for TRAP and TTT SBPs. Abundance at each location is expressed as the fraction of mapped reads and

represented by point area on a linear scale. (b) Box-and-whisker plots comparing abundance of each SBP in epipelagic/surface (SRF, $n = 103$, solid color), deep chlorophyll maximum (DCM, $n = 49$, stripes), and mesopelagic (MES, $n = 26$, checkerboard) samples from the OM-RGCv2+T metatranscriptomic dataset (center line, median; box limits, 25th and 75th percentiles; whiskers, maximum and minimum). Statistical comparisons are given in Supplementary Table 12.



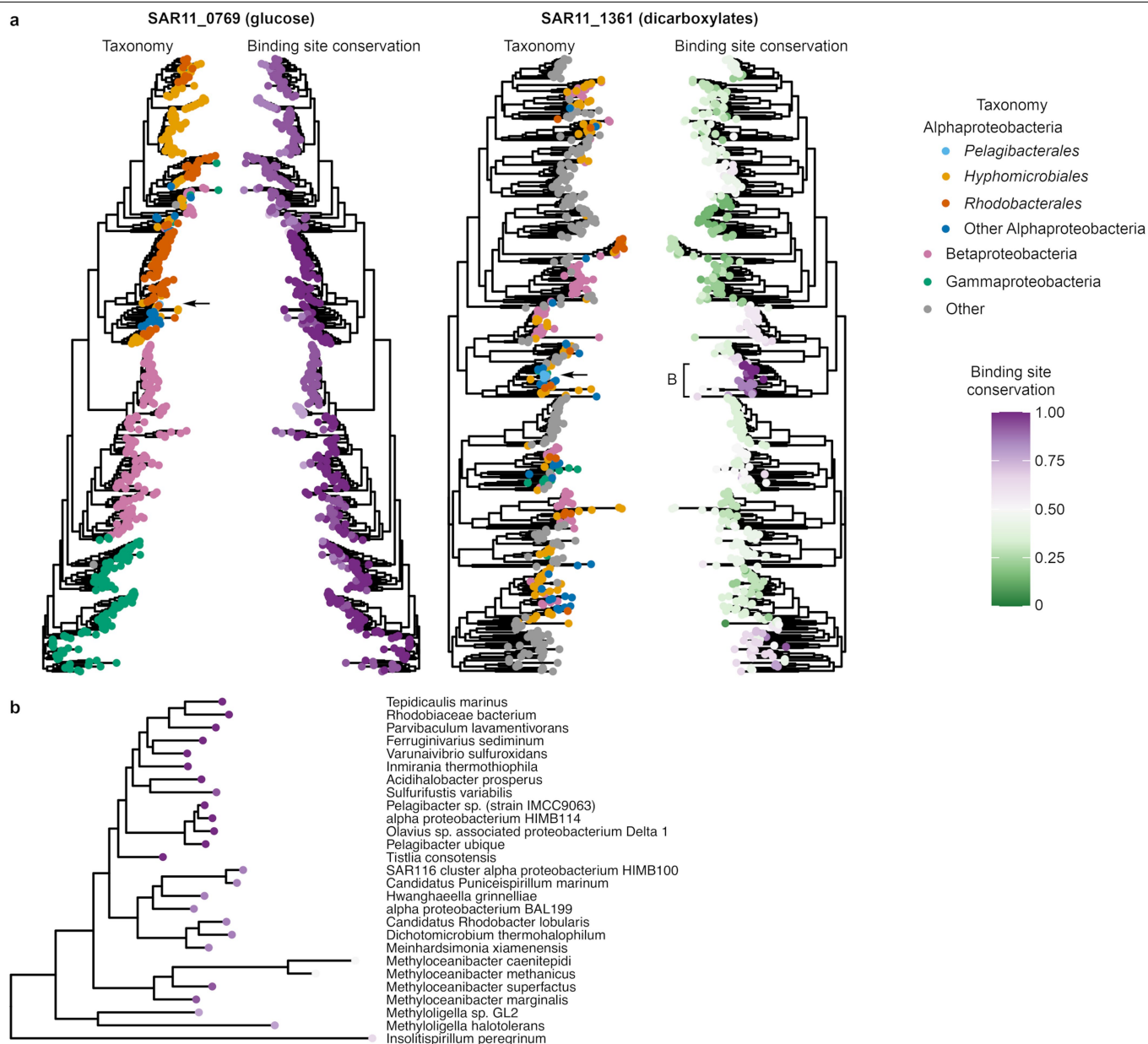
Extended Data Fig. 8 | Phylogenetic analysis of selected SBPs from *Ca. P. ubique* HTCC1062. (a) Maximum-likelihood phylogenies of 1000 homologs of SAR11_0655, SAR11_0953, and SAR11_1336 from the UniProtKB Reference Proteomes and Swiss-Prot databases. The positions of the SAR11 SBPs are indicated by an arrow. Nodes are colored by taxonomy and the

fraction of binding site residues conserved relative to the corresponding protein in *Ca. P. ubique* HTCC1062. **(b)** Taxonomic distribution of sequences displayed in (a). In the case of SAR11_0655, only sequences belonging to the clade indicated in (a) were considered.



Extended Data Fig. 9 | Genome context of SAR11_0655 and SAR11_1361 suggests additional metabolic capabilities of *Ca. P. ubique* HTCC1062. ABC transporter genes (including SBP genes) are shown in orange, while genes putatively involved in metabolism of the transported substrates are shown in red. The genomic regions shown are bounded by non-coding regions of ≥ 75 bp. **(a)** Genome context of SAR11_0655 (9,248 bp). SAR11_0662–SAR11_0664 are homologous to the *pxpABC* genes from *E. coli* (sequence identity 30.0% overall). *pxpABC* encodes 5-oxoprolinase, which catalyzes ATP-dependent hydrolysis of L-pyrroglutamate to L-glutamate⁹⁰. **(b)** Genome context of SAR11_1346 and SAR11_1361 (14,635 bp). SAR11_1354 shows high sequence identity (41.5%) to *csiD* from *E. coli* encoding glutarate 2-hydroxylase, which converts glutarate to L-2-hydroxyglutarate and is involved in catabolism of L-lysine to α -ketoglutarate⁹¹. Although the operon in *Ca. P. ubique* is lacking

the remainder of the the L-lysine catabolic pathway, it does contain a FAD-dependent oxidoreductase of unknown function (SAR11_1353), which may convert L-2-hydroxyglutarate to α -ketoglutarate by analogy with the L-lysine catabolic pathway. Abbreviations: α KG, α -ketoglutarate. **(c)** Experimental confirmation of in vitro glutarate 2-hydroxylase activity of SAR11_1354 by ¹H-NMR (500 MHz, D₂O). Complete conversion of α -ketoglutarate and glutarate to succinate and 2-hydroxyglutarate was observed. Reaction conditions: 1 mM glutarate, 1 mM α -ketoglutarate, 100 μ M Fe(NH₄)₂(SO₄)₂, 5 μ M SAR11_1354 in 20 mM ammonium acetate buffer, 16 h incubation at 24 °C. Right, reaction mixture; left, no enzyme control. The asterisked peak corresponds to acetate. Functional validation of SAR11_0662–0664 and SAR11_1353 was also attempted, but SAR11_0662 and SAR11_1353 could not be expressed in soluble form in *E. coli*.



Extended Data Fig. 10 | Contrasting phylogenetic distributions of SAR11_0769 and SAR11_1361. (a) Maximum-likelihood phylogenies of 500 homologs of SAR11_0769 and SAR11_1361 from the UniProtKB Reference Proteomes and Swiss-Prot databases. The positions of SAR11_0769 and SAR11_1361 are indicated by an arrow. Nodes are colored by taxonomy and the fraction of binding site residues conserved relative to the corresponding

protein in *Ca. P. ubique* HTCC1062. SAR11_0769 is widely distributed among bacteria, while SAR11_1361 appears to be limited mainly to SAR11 bacteria and a small range of other marine Alphaproteobacteria. (b) Expanded view of a clade of the SAR11_1361 phylogeny (indicated in a) showing protein sequences with a similar binding site to SAR11_1361 (suggesting a similar function).

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	ITC: MicroCal PEAQ-ITC control software v1.41 DSF: StepOne Software v2.3
Data analysis	ITC: MicroCal PEAQ-ITC analysis software v1.41, pytc (Jun 15, 2017) DSF: StepOne software v2.3, R v4.2.2 GC-MS: ChromaTOF v4.72.0.0 Phylogenetics: MUSCLE v3.8.3, trimAl v1.2, FastTree v2.1.11, HMMER v3.4, R v4.2.2, ggtree v3.6.2, Krona (Dec 21, 2021) Crystallography: XDS (Jun 30, 2023), MOLREP v11.9.02, REFMAC v5.8.0405, Phenix v1.20.1, COOT v0.8.9.3, Pymol v2.3.4, LigPlot+ v2.2. ZOO and KAMO were accessed on 2022-07-28 (PDB 8HQR), 2022-10-25 (PDB 8HQQ), 2023-06-09 (PDB 8KDO), or 2023-06-28 (PDB 8WCH). Various analyses: R v4.2.2, GraphPad Prism v10.1.0 Custom scripts used for data analysis are available via the Open Science Framework (https://doi.org/10.17605/OSF.IO/47TR5). See text for citations and further details.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Coordinates and structure factors for each crystal structure have been deposited in the Protein Data Bank under the following accession codes: SAR11_0655/L-pyroglutamate (8WCH), SAR11_0769/D-glucose (8HQQ), SAR11_0769/D-galactose (8KD0), SAR11_1210/L-arginine (8HQR). Experimental protein structures referenced in the text are available from the Protein Data Bank under accession codes 1EU8, 1GLG, 2B3B, 2B3F, 2FVY, 2PFY, 2PFZ, 2Q2A, 3OO6, 3ZKK, 4R2B, 4UAB, 4UA8, 5DVI, 5DVJ, 6WGM. Structural models are available from the AlphaFold Protein Structure Database using the following UniProt accessions: SAR11_0266, Q4FP02; SAR11_0271, Q4FNZ7; SAR11_0655, Q4FMW4; SAR11_0769, Q4FMK2; SAR11_0953, Q4FM26; SAR11_1290, Q4FL44; SAR11_1336, Q4FKZ8; SAR11_1346, Q4FKY9; SAR11_1361, Q4FKX4; SAR11_0271*, A0A1X7H7N5; SAR11_1346*, A0A1X7H1Y9; A2cp1_3084, B8JG16. Raw ITC data, DSF data for Biolog assays, GC-MS data, phylogenetic and biogeographical data, and source data for figures are available via the Open Science Framework (<https://doi.org/10.17605/OSF.IO/47TR5>).

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

The research did not involve human participants, their data, or biological material.

Reporting on race, ethnicity, or other socially relevant groupings

See above

Population characteristics

See above

Recruitment

See above

Ethics oversight

See above

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences

☐ Behavioural & social sciences

☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Sample size was not predetermined. For ITC experiments, we generally used sample sizes of 2-4 technical replicates (exact number of replicates for each experiment is given in Supplementary Data 1). Given the high reproducibility of ITC experiments using purified protein, two replicates were considered sufficient, but if there was sufficient material (i.e., protein) available or baseline stability issues with one of the replicates, additional replicates were added. A single replicate was performed for high-throughput screening by DSF; hits were further validated in separate DSF experiments.

Data exclusions

Outliers in DSF and ITC data were excluded as appropriate. DSF data was most commonly excluded due to the lack of a clear melting transition in the data, usually as a result of dye aggregation, which could be resolved using a new dye aliquot. ITC data was excluded in the case of baseline instability, as a result of an unclean sample cell or bubbles in the sample cell. Exclusion criteria were not predetermined.

Replication

The number of replicates for each experiment is discussed under "sample size". All attempts at replication were successful.

Randomization

Not relevant - the study did not involve group allocation.

Blinding

Not relevant - the study did not involve group allocation.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.