Generative Landscapes and Dynamics to Design Multidomain Artificial Transmembrane Transporters

Fernando Montalvillo Ortega[1†], Fariha Hossain[2†], Vladimir V. Volobouev[2], Gabriele Meloni[1*], Hedieh Torabifard[1*], Faruck Morcos[2,3,4*]

[1]Department of Chemistry and Biochemistry, University of Texas at Dallas, Richardson, 75080 TX, USA.

[2]Department of Biological Sciences, University of Texas at Dallas, Richardson, 75080 TX, USA.

[3]Departments of Bioengineering and Physics, University of Texas at Dallas, Richardson, 75080 TX, USA.

[4]Center for Systems Biology, University of Texas at Dallas, Richardson, Texas 75080, USA.

[†]These authors contributed equally to this work

[*]Corresponding author

Protein design is challenging as it requires simultaneous consideration of interconnected factors, such as fold, dynamics, and function. These evolutionary constraints are encoded in protein sequences and can be learned through the latent generative landscape (LGL) framework to predict functional sequences by leveraging evolutionary patterns, enabling exploration of uncharted sequence space. By simulating designed proteins through molecular dynamics (MD), we gain deeper insights into the interdependencies governing structure and dynamics. We present a synergized workflow combining LGL with MD and biochemical characterization, allowing us to explore the sequence space effectively. This approach has been applied to design and characterize two artificial multidomain ATP-driven transmembrane copper transporters, with native-like functionality. This integrative approach proved effective in unraveling the intricate relationships between sequence, structure, and function.

Protein design is becoming a cornerstone of modern biotechnology, with possible applications spanning from therapeutic drug design to industrial biocatalysis. Despite its potential, the immense combinatorial space of possible sequences and the challenge of preserving crucial interactions for functional and structural stability make the design process time-consuming and computationally challenging. However, recent advances in machine learning (ML) along with increasing sequence data availability have revolutionized our ability to model biological systems from evolutionary clues. Current applications range from protein structural prediction to the design of novel proteins with enhanced functionality (such as modifying antibody binding affinity to their respective targets or improving catalytic capacity of enzymes) and stability (*1–10*).

Researchers have been exploring diverse ML strategies in protein design, ranging from sequence-based and sequence-labeling models to structure-based and hybrid approaches (*11, 12*). Among these techniques, latent generative models, such as Variational Autoencoders (VAEs), have been only recently explored in the context of biological systems (*3–5, 13*). We introduce a framework that learns from extant protein sequence data and harnesses the reconstructive ability of VAEs in conjunction with Direct Coupling Analysis (DCA) to produce maps of generated sequences known as latent generative landscapes (LGLs) (*14*). These landscapes model protein sequence-function relationships, capturing sequence diversity within a family while assessing functional integrity, offering a powerful tool for protein design.

When designing proteins, it is important to preserve crucial motifs essential for protein function, as well as highly coupled interacting residues that form interdependent networks critical for the protein's behavior, even beyond conserved sites. A key strength of the LGL framework lies in its ability to learn these key interactions from only protein-family sequences, scoring generated sequence variants more favorably (greater negative value) when such interactions are preserved (*14*). However, relying solely on sequence-based approaches can limit our understanding of intricate interdependencies within proteins, particularly in complex membrane-embedded nanomachines in which catalysis and multidomain coupling are central to protein function. Coupling this method with molecular dynamics (MD) simulations allows for a deeper exploration of how interacting residues and novel mutations influence the protein's conformational dynamics, domain crosstalk, and functional behavior. This integrated computational approach can bridge the gap between prediction and experimental success. Rather than conducting experiments on numerous designed sequences to identify functional ones, it enables us to achieve remarkable success with a carefully selected subset. This, in turn, allows for dedicating more resources to in-depth validation and characterization of the most promising candidates, an aspect often lacking in currently available protein design studies.

The P-type ATPase superfamily of transmembrane transporters is crucial for cellular homeostasis, coupling ATP hydrolysis to substrate transport across membranes against their electrochemical gradient (*15, 16*). These large, multidomain transmembrane proteins rely on intricate crosstalk between multiple cytosolic and transmembrane domains. Their mechanism couples ATP hydrolysis with auto(de)phosphorylation to trigger dramatic and sequential conformational rearrangements in the cytosolic domains. This results in transmembrane helices movements, which in turn facilitate substrate translocation across the membrane lipid bilayer (*17, 18*). Within this superfamily, the P_{1B}-type ATPase subfamily plays an essential role in maintaining transition metal homeostasis, where the $P_{1B\text{-}1}$-type ATPases stand out as the ubiquitous Cu(I) exporters in all organisms (*19*). Thus, the complexity, functional diversity, and relevance of this system make it an ideal candidate to test the robustness of the synergized protein design strategy.

Here, we apply an interdisciplinary approach to efficiently design artificial $P_{1B\text{-}1}$-type ATPase transporters that maintain their quintessential coupled multidomain complexity and multifunctional properties. First, we employ the LGL framework to produce novel variants of the transporter family, decoding few from the vast sequence space based on favorable Hamiltonian metric and the presence of key subgroup-specific motifs essential for function. Next, we leverage MD to assess the structural stability and dynamics adherence to the expected catalytic scheme. For these generated variants, we demonstrate coupled domain movements crucial for catalytic function, comparable to those identified in the wild-type (WT) protein(s). Finally, we experimentally characterize the preservation of multifunctional properties encompassing membrane embedding, proper fold, copper translocation, and *in vivo* cellular protection from copper toxicity. Our design of complex transmembrane transporters that integrate seamlessly into a lipid bilayer and maintain full functional fidelity advances previous attempts in this field (*20–23*). The rate of success on the selected designs is notable, something rare for sequences with hundreds of mutations, opening new possibilities for creating tailored proteins with specific functional characteristics through an efficient, sequence-driven approach synergized with molecular dynamics.

Latent generative landscape produces non-extant transporter sequences

The LGL framework enables visual and quantitative exploration of the sequence energy landscape, enhancing mutational analysis, the study of phylogenetic relationships, and identification of different functional clusters (*14*). However, the generative ability to produce complex new protein sequences with desired features has been less explored. The incorporation of the Hamiltonian metric, determined from key interactions learned within the protein subfamily using DCA (see *Methods*), hints at relative functional fitness. This definition of fitness incorporates various aspects of this type of multifunctional transmembrane proteins, such as selective transport of specific substrates, proper energy transduction, coupled conformational changes necessary for activity modulation, and intercommunication between domains. Thus, we hypothesized that the LGL can produce new variants of sequences that preserve these multiple layers of complexity such as the ones present in functional $P_{1B\text{-}1}$-type ATPase transporters. These nanomachines couple ATP hydrolysis and auto(de)phosphorylation in the soluble domains, with spatially distant transmembrane substrate translocation via a coupled actuator unit.

The LGL training input consists of a multiple sequence alignment (MSA) of the P_{1B}-type ATPase family, comprising a refined set of approximately 13,500 sequences of 574–661 residue length (fig. S1A). This MSA was curated to minimize noise and sampling bias (see *Methods*). Fig. 1A illustrates the methodology used to construct an LGL map of a total of 250,000 pixels, each representing a generated sequence with assigned Hamiltonian values. These generated sequences (pixels) can be decoded and analyzed in synergy with MD simulations, serving as a powerful integrated tool for protein design that effectively narrows the search space, enabling comprehensive characterization of the artificial transporters. When the P_{1B}-type ATPase training data is plotted on the LGL in Fig. 1B, a clustering pattern emerges, aligning with the established classification of the subgroups based on reported transmembrane motifs that correlate with transition metal cargo selectivities (*18*). Though the LGL allows us to potentially sample thousands of generated sequences within the subfamily, we narrowed down the selection to the $P_{1B\text{-}1}$ subgroup. This highly characterized subgroup, with available crystal structures, enables a robust and meaningful comparison and plays essential physiological roles across all organisms, including humans (*16, 18, 24*). Since $P_{1B\text{-}1}$ sequences from the training dataset appear in the lower half of the map as purple symbols in Fig. 1B, we hypothesize that the sequences decoded from this region are more likely to have P_{1B-1}-like functionality. With this in mind, we focused on the area that contained the well-characterized Cu(I)-pump from *Legionella pneumophila* (*Lp*CopA) sequence for which structural information is available (PDB: 3RFU), thus serving as a reference. Fig. 1C depicts a 3D representation of the region, where the red circle represents the location of WT *Lp*CopA sequence, and the nearby white circles, two decoded generated sequences GS1 and GS2 (selected out of eight characterized by MD, *vide infra*). Additionally, we ensured that the decoded sequences did not directly overlap with existing native sequences to increase the diversity of novel mutations (fig. S1B).

Structural stability and mutation distribution of generated sequences

The domain topology diagram in Fig. 2A illustrates that P_{1B-1}-type ATPases consist of a transmembrane (TM) domain featuring eight TM helices (MA-M6) responsible for metal substrate binding and forming the translocation pathway, three cytosolic domains (N-, nucleotide binding; P-, phosphorylation; A-, actuator), and metal-binding domain(s) (MBDs), located at either the N or C-terminal, serving a regulatory function (*25*). The deletion of MBDs has been shown to reduce the transport rate without affecting the transport mechanism or selectivity (*26, 27*). Therefore, it was excluded from our LGL training dataset and subsequent analysis.

To better understand the distribution of the novel mutations within the described topological framework, Fig. 2C highlights the mutation locations in the generated sequences, with corresponding labeled sequence alignments provided in fig. S2 and fig. S3. To explore the effect of these mutations on the 3D structure and stability, the sequences were first analyzed using TOPCONS to predict membrane topology, number and location of transmembrane helices, and delineate intracellular and extracellular domains (*28*). The predicted domain topology included eight TM helices and two large intracellular soluble regions that closely resemble that of the WT (fig. S4A). Consequently, AlphaFold2 (AF2) and Positioning of Proteins in Membranes 2.0 (PPM2.0) were employed to obtain structural and membrane insertion predictions for GS1 and GS2 presented in Fig. 2B (*1, 29*). The structures obtained showcase the characteristic eight TM helices and three catalytic soluble domains, as well as correct lipid bilayer insertion. The sequences had template modeling scores near 1 as reported in Fig. 2C (*1, 26*). This metric is commonly used in structural biology to quantify fold resemblance, with values ranging from 0 to 1, where 1 indicates a perfect match and an identical fold to the WT protein (*30*).

Relative to *Lp*CopA, both GS1 and GS2 contained approximately 180 novel mutations, despite sharing 70% identity to the WT. These mutations were distributed throughout the protein where approximately 24% of the A-domain, 26% of the P-domain, 42% of the N-domain, and 23% of the TM region were mutated (Fig. 2D). This result implies that the VAE provides each domain with flexibility in accommodating mutations without significantly disrupting the overall structure. A closer inspection revealed that both sequences contained a single insertion at V365 in the N-domain compared to the WT. This was notable, as removing this residue caused distortion of the first N-domain β-sheet (P359 to A367 in GS1 and P358 to A366 in GS2) in the AF2 model (fig. S4B). Further MD simulations showed that the secondary structure was not recovered after 400 ns (fig. S4C). Interestingly, this insertion aligns appropriately with other $P_{1B\text{-}1}$ sequences (see position 397 in fig. S3), highlighting the predictive power of our generative model. Additionally, the generated sequences have 32 mutations among themselves with mutations scattered throughout the protein (Fig. 2, C and D). Despite these variations, the Hamiltonian values reported for each sequence in Fig. 2C are comparable, suggesting that these sequences maintain similar functional properties.

Molecular dynamics identifies crucial dynamics of generated transporters

With high confidence in the structural integrity of the generated sequences, we performed MD simulations to investigate crucial dynamical properties. Structural characterizations of CopA in both eukaryotic and prokaryotic systems reveal a high degree of similarity (PDB: 3RFU, 8Q75, 7SI3, 4BBJ, and 8Q76 represent E2P/E2P$_i$; PDB: 7R0I, 8Q74, and 7XUM represent E1-copper bound; PDB: 7R0G, 8Q73, and 7XUN represent E1-apo), indicating conservation of the Post-Albers mechanism cycle across the kingdoms (Fig. 3A) (*19, 24, 26, 31–33*). This transport mechanism involves the intricate and coupled interplay of soluble domain movements driven by auto(de)phosphorylation events. These movements induce rearrangements within the TM domain, allowing the protein to alternate between two major states: the E1 inward-facing state, which has high affinity towards the substrate, and the E2 outward-facing state with lower affinity (*34*). Particularly, the E2P$_i$ state is the starting point of a significant conformational change of the A-domain as the system evolves towards the E1 state (*26, 31*). This A-domain movement triggers a substantial reorganization of the TM domain, visualized as two blocks of four TM helices each, (MA, MB, M1, M2) and (M3, M4, M5, M6), which move relative to each other, opening and closing the translocation pathways and making the binding site accessible to opposite sides of the lipid bilayer (*31*).

Our initial decoded list comprised of eight generated sequences (GS1-GS8) with varying Hamiltonian values and locations on the LGL map (Fig. 2C and fig. S5, A and D). Numerical Hamiltonian value alone, however, does not fully convey the extent to which interactions are retained or altered, nor their impact on the coupled dynamics. We aimed to explore various LGL regions around the *Lp*CopA area to gain a deeper understanding of the relationship between the calculated Hamiltonian and its impact on the system's dynamics. Integrating MD simulations allowed us to unravel intricate details regarding their A-domain and TM helix bundle rearrangement properties, enabling the evaluation of their consistency with the Post-Albers cycle and available crystal structures.

To assess the viability of the generated sequences as functional candidates, MD studies were conducted for *Lp*CopA and GS1-GS8. Each system underwent several 400 ns trials within an isobaric-isothermal (NPT) ensemble, with the protein embedded in a lipid bilayer and surrounded by a salt-containing aqueous solution. To exemplify the evaluation process, we will focus our discussion on the results of *Lp*CopA (PDB: 3RFU), GS1, and GS2, which were selected due to their native-like dynamics and favorable Hamiltonian values. Fig. 3B illustrates the A-domain progression at different stages of the transition from E2P$_i$ (PDB: 3RFU) to E1 (PDB: 7R0I) (outward- to inward-open) as well as the two metrics used for its movement quantification: the tilt angle (pink lines) and the Δdistance (blue lines). The distribution of these two values over the simulation time for the three systems is shown in Fig. 3C, highlighting significant dynamics in the A-domain (see Movie S1). The angle and distance values were combined into a one-dimensional metric, termed the *A-domain movement score*, using equations S4-S7. Fig. 3D shows that positive values, defined as "progress", correspond to A-domain motions toward the E1 state, while negative values, referred to as "deviation", indicate motions in the opposite direction. Although WT *Lp*CopA did not fully transition to the E1 state due to the limitations associated with capturing large conformational changes using all-atom simulations, the overlapping bimodal distribution among the three proteins unravels similar A-domain dynamics to those of the WT, indicating comparable coupled domain movements characteristic of the Post-Albers catalytic cycle. The results for each system trial are detailed in fig. S6, A and C.

Fig. 3E summarizes the qualitative and quantitative analyses used to assess the TM inter-block rearrangement. The superposition of structures, with purple representing the first TM block and green the second block, was combined with the helix Distance Difference Matrix (DDM) to observe and quantify the rearrangement. Results from the *Lp*CopA simulations (Fig. 3E *top* panel), show a pattern clearly mimicking the proper inter-block distance changes observed in experimentally determined crystal structures in fig. S6B (*19, 26, 31*). In the DDM for the 100% A-domain progress in fig. S6B, a large inter-block rearrangement is observed at the bottom-left and top-right corners, while the intra-block rearrangement showed only small RMSD values. We demonstrated that, as the A-domain of both generated sequences transitions from the E2P$_i$ to the E1 state, the TM block movement begins to resemble that of a coupled native transporter, with GS2 showing this more clearly (Fig. 3E *bottom* panel, Movie S2). GS2 also starts to exhibit an E1-like transmembrane tunnel, as indicated by CAVER 3.0 analysis (Fig. 3E *right-most bottom* panel) (*35*). These results suggest that the modeled systems retain the essential mechanistic features required for their biological activity, revealing dynamic properties that complement the LGL approach.

The A-domain and corresponding helix bundle movement rearrangement analyses were also carried out for the GS3-GS8 sequences (fig. S5B-D; tables S1-S2; see *Supplemental* for detailed analysis). Synchronizing LGL predictions with MD simulations allowed us to identify which aspects of these dynamics in the decoded eight sequences were impacted and to what extent, a valuable insight in addition to the Hamiltonian values. Notably, GS1 and GS2 demonstrated superior performance in preserving these desired dynamic properties and thus were subjected to experimental characterization.

Experimental characterization of preserved WT functionality

The two selected generated sequences were subjected to experimental characterization to validate membrane insertion, correct folding, *in vitro* catalytic ATP hydrolysis activity, and *in vivo* cellular protection from copper toxicity underlying transport capability. TM domain folding was assessed by analyzing membrane localization upon recombinant expression in *E. coli* membrane fraction (MF), while folding of the soluble domains was validated through *in vitro* ATP hydrolysis activity, ensuring the protein retained its ATP hydrolysis and P-domain's autophosphorylation turnover. Copper transport activity was evaluated *in vivo* by testing recombinant GS1 and GS2 expression via intracellular copper quantification and protection from copper toxicity.

Initial results, presented in the Western blot in Fig. 4A, confirmed successful recombinant expression despite each generated protein featuring approximately one-third of its sequence mutated. Both protein constructs were primarily localized to the MF, indicating proper insertion in the lipid bilayer and TM domain folding, further supported by the significantly monodisperse Size Exclusion Chromatography (SEC) profiles upon extraction and purification in detergent micelles (fig. S7A). No significant misfolded proteins were indeed observed in the soluble fraction (SF) or inclusion bodies.

Upon successful purification in detergent micelles (fig. S7A), we determined the specific ATP hydrolysis activity rates as this is a defining catalytic biochemical property of all P-type ATPases. Using *in vitro* malachite green assays, which allow colorimetric quantification of released inorganic phosphate upon ATP hydrolysis, we demonstrated that both GS1 and GS2 exhibit high catalytic turnover rates, despite lacking the MBD. As shown in Fig. 4B, ATPase rate of GS2 (77 $nmol\ min^{-1}\ mg^{-1}$) is comparable to characterized native $P_{1B\text{-}1}$-type ATPases (dashed lines) while GS1 (469 $nmol\ min^{-1}\ mg^{-1}$) exceeded those values (*17, 26, 36*). This raised the question of whether the mutations may have compromised the transporter's ability to autophosphorylate at the conserved DKTGT motif present in the P-domain. To ensure that ATP hydrolysis remained coupled to P-domain phosphorylation, we generated GS1-D349A lacking the conserved aspartate residue which undergoes the required phosphorylation essential to complete the Post-Albers cycle. The absence of any ATPase catalytic activity in GS1-D349A confirmed that ATP hydrolysis in the generated sequence is strictly coupled and dependent on autophosphorylation (Fig. 4B). However, while the ATPase activity in native characterized $P_{1B\text{-}1}$-type ATPase is strictly dependent on the presence of Cu(I) substrates, our *in vitro* assay did not show the expected copper dependency. We proposed that this behavior resulted from the μM level copurification of copper with the protein, suggesting a very high affinity of the generated constructs towards Cu(I). This was further supported by the observed reduction but incomplete abolishment of ATP hydrolysis activity upon treatment with up to 50 μM tetrathiomolybdate (TTM), a copper chelator with K_d of 2.32 x 10^{-20} M (*37*), which effectively competes with Cu(I) availability with the Cu(I)-pumps (Fig. 4B).

Thus, to unambiguously confirm functionality *in vivo*, copper susceptibility and metal quantification assays were performed upon expression of GS1 and GS2 in native and copper-sensitive *E. coli* strains. First, the ability of GS1 and GS2 to efficiently extrude Cu(I) from the cytosol was investigated via *in vivo* intracellular copper quantification by Inductively-Coupled Plasma Mass Spectrometry (ICP-MS) upon cellular growth in media containing increasing copper concentrations. In this assay, we expected functional copper exporters to reduce intracellular copper accumulation. *E. coli* cells that were recombinantly expressing GS1 and GS2 showed reduced copper concentrations compared to cells transformed with a control plasmid not encoding for any Cu(I) P-type ATPase pump (Fig. 4C). In agreement with native-like function, GS1 and GS2's ability to reduce cellular concentration aligned well with cells overexpressing the functional native *Ec*CopA pump. To further validate the ability of GS1 and GS2 to protect cells from copper toxicity by extrusion, *in vivo* copper metal susceptibility assays were performed in the copper-sensitive *E. coli* ECA464 strain lacking the endogenous bacterial copper efflux and detoxification system genes (*copA, cueO, and cusCFBA*) (*38*). In this assay, protection from copper toxicity was determined by monitoring bacterial growth under increasing copper concentrations to evaluate the protective effect of the expressed transporters. As shown in Fig. 4D (with detailed OD_{600} data from a representative trial in fig. S7B), at a concentration of 6.0 mM copper, cells expressing GS1 (74.6% mean-normalized relative growth) and GS2 (80.8%) demonstrated comparable or exceeding protection ability against copper toxicity than cells expressing the native *Ec*CopA (62.3%). Contrarily, cells transformed with a corresponding empty control vector (14.1%) featured significant growth inhibition due to copper toxicity at higher copper concentrations. Together, experimental *in vitro* and *in vivo* characterization indicate that both generated sequences retain the expected functionality of a P_{1B-1}-type ATPase, effectively exporting copper from the cellular cytosol.

Discussion

Cu(I) transmembrane pumps are complex ATP-dependent multidomain transporters responsible for maintaining essential metal homeostasis in cells in all kingdoms of life. Designing membrane proteins of this level of complexity is significantly challenging because it requires addressing intricate structural dynamics, achieving stability in lipid bilayers, and assessing their functional ability in a native-like environment. Through synergy of machine learning, molecular dynamics, and experimental validation, we designed two artificial transmembrane Cu(I) P-type ATPase pumps (GS1 and GS2), each containing approximately 180 mutations from the referenced WT *Lp*CopA sequence. The LGL framework successfully decoded non-extant novel variants by leveraging preserved coevolutionary interactions, learned exclusively from the sequence space. The integration of MD and experimentation allowed characterization of essential dynamics and function, providing insights into structural and functional behaviors expected to be preserved in the generated sequences. Results showed that both were capable of adopting the correct fold required for membrane insertion, maintained coupled domain movements, and provided protection against copper toxicity through export across the membrane. The functional efficacy of these artificial transporters marks a major step forward in the field of evolutionary-inspired membrane protein design, offering new possibilities for future advancements.

The interdisciplinary approach presented here can also be applied to other complex, less characterized systems to advance our understanding and engineering of biological systems. Our learning framework integrates an in-depth exploration of sequence space, dynamics, and the biochemical properties of the protein system under consideration. This has significantly narrowed the gap between design predictions and experimental characterization, a trend likely to continue as we deepen our understanding of the functional patterns within protein families as a whole. The potential applications are vast, offering the promise of inspiring numerous additional approaches aimed at designing protein functions in a logical and time-efficient manner.

An important avenue for further investigation includes the analysis of the 32 mutations distinguishing GS1 and GS2 as a potential means of controlling ATP hydrolysis rate. This energy transduction process drives the conformational changes necessary for ion transport. Investigating how these mutations modulate ATP hydrolysis could provide insights into the energy transduction mechanism and overall catalytic efficiency in transmembrane transition metal pumps, and P-type ATPases at large.

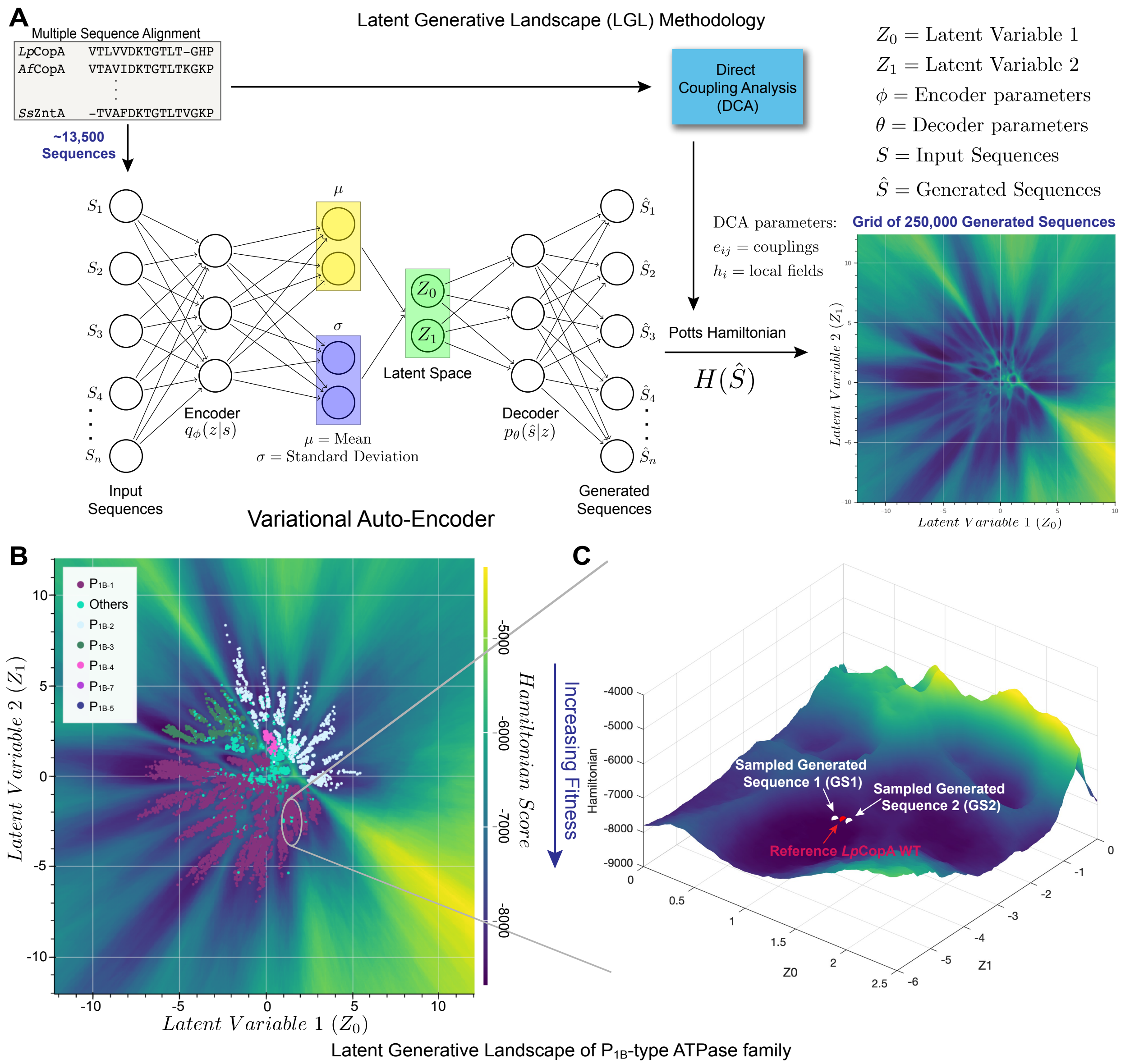
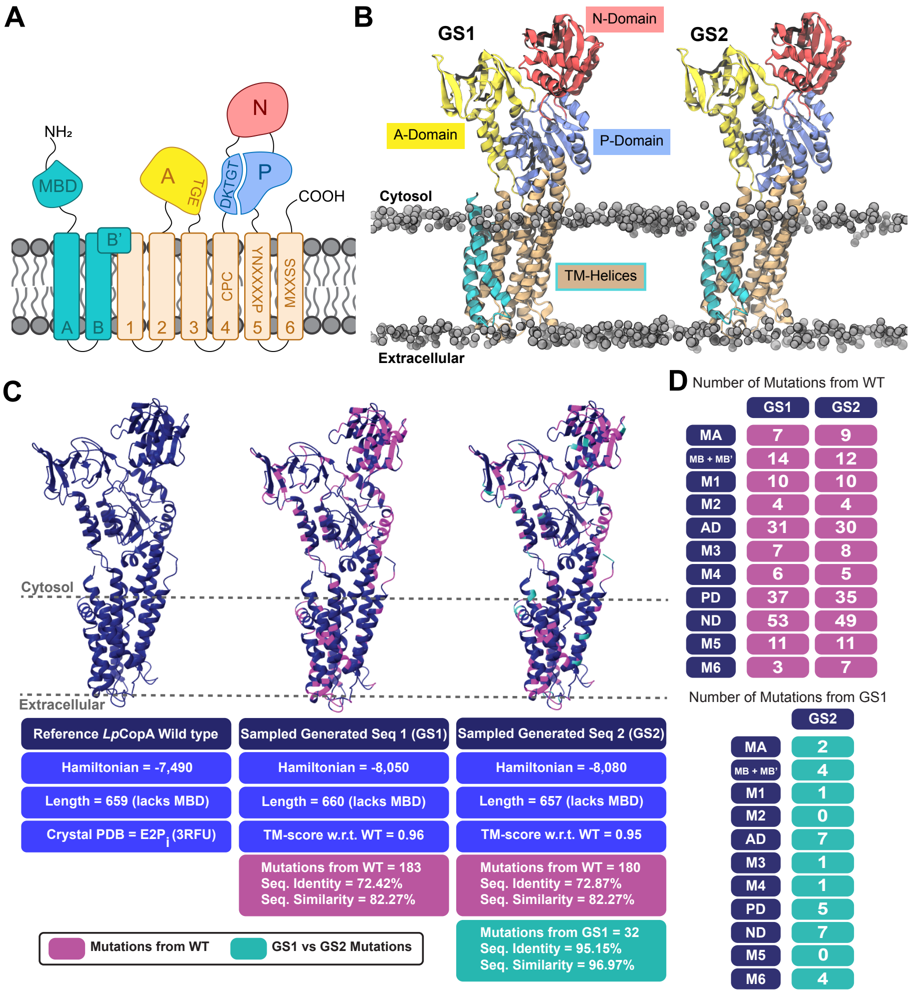

A
Latent Generative Landscape (LGL) Methodology
Multiple Sequence Alignment
LpCopA VTLVVDKTGTLT-GHP
AfCopA VTAVIDKTGTLTKGKP
SsZntA -TVAFDKTGTLTVGKP
~13,500 Sequences
Direct Coupling Analysis (DCA)
Z_0 = Latent Variable 1
Z_1 = Latent Variable 2
ϕ = Encoder parameters
θ = Decoder parameters
S = Input Sequences
$\hat{S}$ = Generated Sequences
S_1 S_2 S_3 S_4 S_n
μ
σ
Z_0 Z_1
Latent Space
$\hat{S}_1$ $\hat{S}_2$ $\hat{S}_3$ $\hat{S}_4$ $\hat{S}_n$
Encoder $q_\phi(z|s)$
μ = Mean
σ = Standard Deviation
Decoder $p_\theta(\hat{s}|z)$
Input Sequences
Generated Sequences
Variational Auto-Encoder
DCA parameters:
e_{ij} = couplings
h_i = local fields
Potts Hamiltonian
$H(\hat{S})$
Grid of 250,000 Generated Sequences
Latent Variable 2 (Z_1)
Latent Variable 1 (Z_0)
-10 -5 0 5 10
B
P1B-1
Others
P1B-2
P1B-3
P1B-4
P1B-7
P1B-5
10 5 0 -5 -10
-10 -5 0 5 10
Latent Variable 2 (Z_1)
Latent Variable 1 (Z_0)
-5000 -6000 -7000 -8000
Hamiltonian Score
Increasing Fitness
C
Hamiltonian
-4000 -5000 -6000 -7000 -8000 -9000
Sampled Generated Sequence 1 (GS1)
Sampled Generated Sequence 2 (GS2)
Reference LpCopA WT
0 0.5 1 1.5 2 2.5
Z0
0 -1 -2 -3 -4 -5 -6
Z1
Latent Generative Landscape of P1B-type ATPase family

Figure 1: **Overview of the Latent Generative Landscape (LGL) methodology.** (**A**) The algorithm combines the generative power of the Variational Auto-Encoder (VAE) and the Potts Hamiltonian to approximate fitness of the generated sequences, resulting in an LGL map. (**B**) The LGL plot was generated using the assembled MSA for the P_{1B}-type ATPase family. Labeled sequences part of the training dataset (13,553 sequences) are plotted on the LGL to identify the corresponding subgroup locations. (**C**) A 3D landscape view of the decoded region with reference *Lp*CopA labeled with a red circle while the two selected decoded sequences (GS1 and GS2) labeled in white.



Reference *Lp*CopA Wild type	Sampled Generated Seq 1 (GS1)	Sampled Generated Seq 2 (GS2)
Hamiltonian = -7,490	Hamiltonian = -8,050	Hamiltonian = -8,080
Length = 659 (lacks MBD)	Length = 660 (lacks MBD)	Length = 657 (lacks MBD)
Crystal PDB = E2P_i (3RFU)	TM-score w.r.t. WT = 0.96	TM-score w.r.t. WT = 0.95
	Mutations from WT = 183 Seq. Identity = 72.42% Seq. Similarity = 82.27%	Mutations from WT = 180 Seq. Identity = 72.87% Seq. Similarity = 82.27%
		Mutations from GS1 = 32 Seq. Identity = 95.15% Seq. Similarity = 96.97%

🟪 Mutations from WT 🟩 GS1 vs GS2 Mutations

Number of Mutations from WT

	GS1	GS2
MA	7	9
MB + MB'	14	12
M1	10	10
M2	4	4
AD	31	30
M3	7	8
M4	6	5
PD	37	35
ND	53	49
M5	11	11
M6	3	7

Number of Mutations from GS1

	GS2
MA	2
MB + MB'	4
M1	1
M2	0
AD	7
M3	1
M4	1
PD	5
ND	7
M5	0
M6	4

Figure 2: **Structural and sequence composition details of the P_{1B-1}-type ATPase subgroup and the two selected decoded generated sequences.** (**A**) A 2D domain topology of the P_{1B}-type ATPase family with highlighted conserved motifs. DKTGT and TGE motifs on the P-domain and A-domain, respectively, are key in the auto(de)phosphorylation catalytic cycle characteristic of P-type ATPase transporters. Specific substrate selectivity motifs associated with the P_{1B-1}-type ATPase subgroup are labeled on M4, M5, and M6. (**B**) A 3D domain topology and membrane insertion of the generated proteins in lipid bilayers, pre-screened via the PPM 2.0 server for membrane compatibility prior to MD simulation. (**C**) Comparison of structural, sequential, and fitness characteristics of GS1 and GS2 relative to WT *Lp*CopA. TM-score values indicate that the AF2-predicted structures closely match the crystal structure. Magenta highlights mutations relative to WT, while teal shows mutations relative to each other between the generated sequences. The Hamiltonian values infer that the generated sequences would maintain native-like functional fitness when compared to *Lp*CopA. (**D**) Domain-specific mutation count comparison for GS1 and GS2 versus *Lp*CopA in the top table, with the bottom table showing mutation counts between GS1 and GS2.

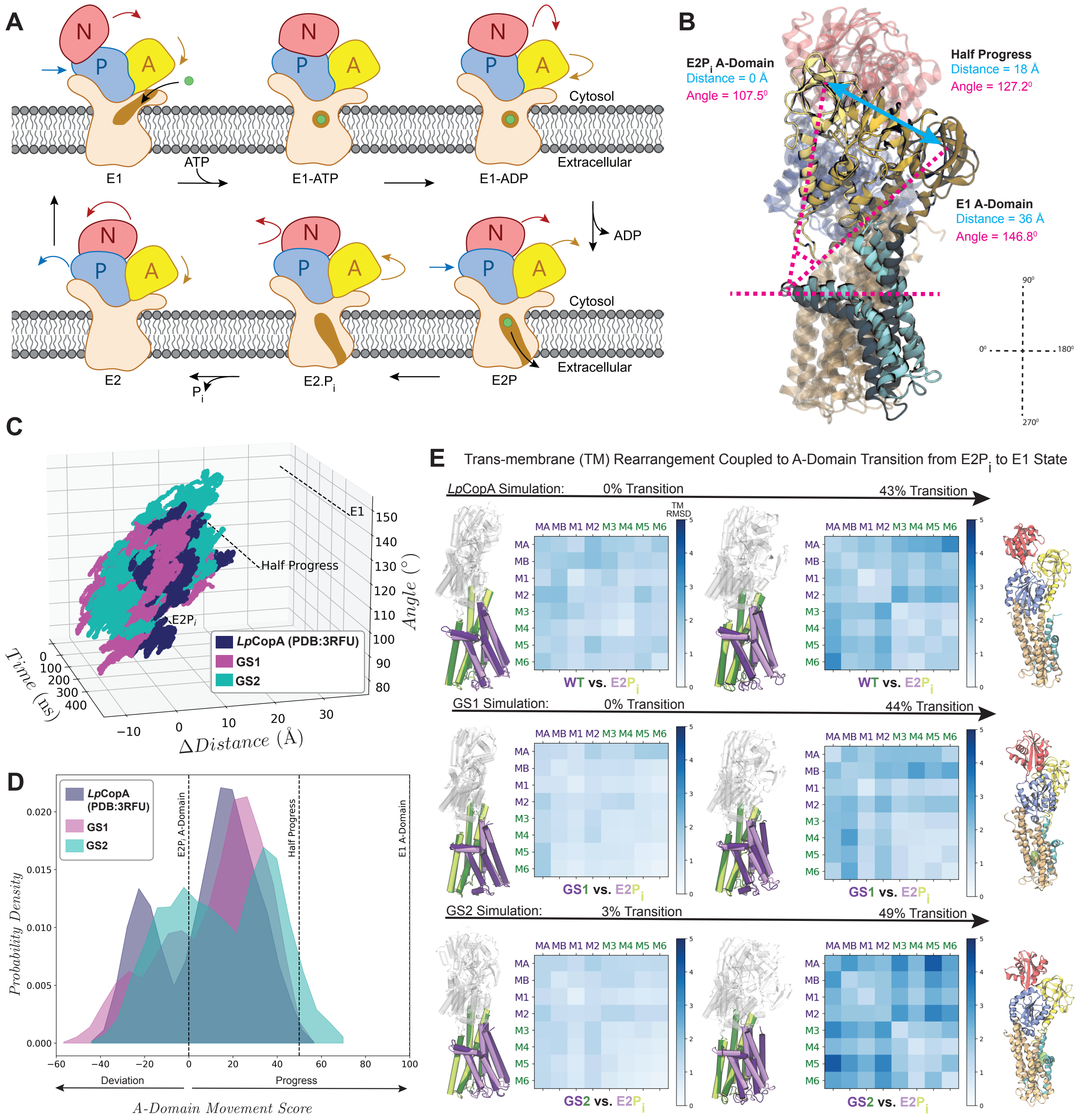


Figure 3: **Molecular dynamics to capture key mechanisms and domain coupling required for metal substrate transport.** (**A**) The P_{1B}-type ATPase family follows the Post-Albers cycle. The pump converts between two states, E1 and E2, in an alternating access mechanism. Metal(s) bind to TM site(s) (E1 state), are occluded within the membrane upon ATP hydrolysis/phosphorylation (E1P), and released on the opposite side in the E2P state, followed by a dephosphorylation transition state (E2P_i) to then regenerate E1. During the E2P_i to the E1 transition, the A-domain undergoes a significant transition that is coupled to a rearrangement of the transmembrane transport pathway. (**B**) Structural alignment of the different stages of the A-domain transition along with labeled vectors used to calculate the tilt angle and Δdistance. (**C**) A-domain dynamics were measured over 400 ns simulations by tracking changes in tilt angle and Δdistance for *Lp*CopA, GS1, and GS2. (**D**) The *A-domain movement scores* for GS1 and GS2 show transitions similar to the WT. (**E**) Transmembrane helices rearrangement panels consist of structural alignments and TM helix distance difference matrices in relation to labeled *A-domain movement scores*. The purple block (MA-M2) and green block (M3-M6) represent the regions of the protein, with the E2P_i state shown in lighter tones and later A-domain stages in darker tones. The top row illustrates the rearrangement in the WT *Lp*CopA simulation followed by GS1 and GS2, respectively. The last column displays CAVER tunnels for each.

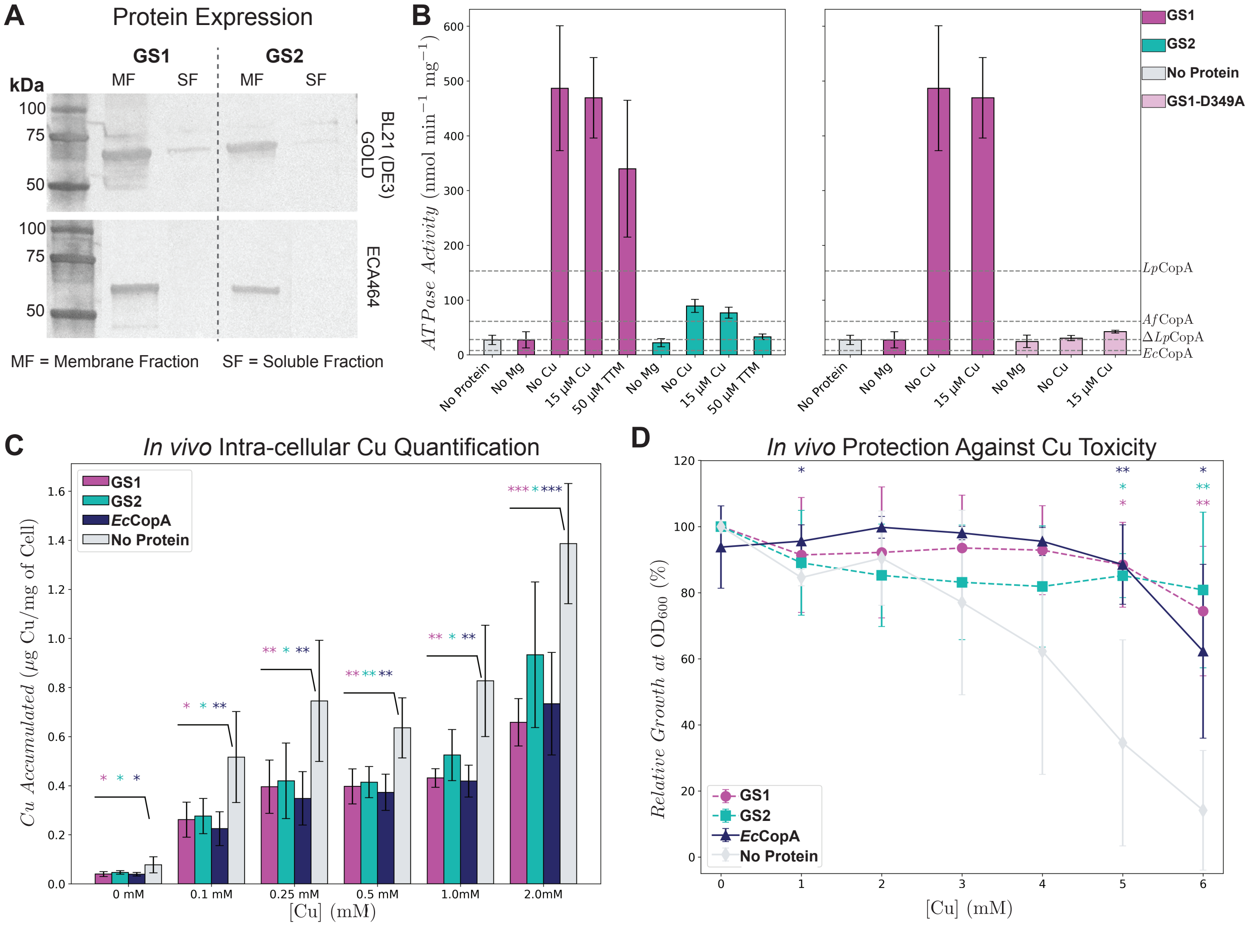


Figure 4: **Experimental characterization of expression, membrane embedding, hydrolysis activity, and copper export for GS1 and GS2.** (**A**) Western blot analysis reveals successful protein expression and insertion of both GS1 and GS2 in the membrane fraction. (**B**) *In vitro* ATPase activity rate for GS1 and GS2 in comparison to several characterized $P_{1B\text{-}1}$ transporters (dashed line). WT Δ*Lp*CopA lacks the MBD, providing an ideal comparison with GS1 and GS2. GS2 exhibited rates that are in range with the reported values for characterized prokaryotic P-type ATPases, while GS1 showed a significantly higher catalytic rate (*17, 26, 36*). The GS1-D349A mutation, which abolished ATP hydrolysis activity, confirmed that the defining auto(de)phosphorylation of the aspartate in the conserved DKTGT motif in the P-domain was preserved. (**C**) *In vivo* efflux assay demonstrates that both GS1 and GS2 can effectively export copper from cells as a function of copper concentrations, similar to WT *Ec*CopA, while the absence of a transporter leads to metal accumulation within the cells. (**D**) *In vivo* copper-susceptibility growth assay confirms that both GS1 and GS2 can protect cells against copper toxicity and promote cell viability as a function of increasing copper concentration, similar to overexpression of WT *Ec*CopA. Statistical significance was calculated by unpaired Student's *t*-test (*** for p <0.001, ** for p <0.01, and * for p <0.05).

References and Notes

1. J. Jumper, *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596** (7873), 583–589 (2021).

2. A. J. Riesselman, J. B. Ingraham, D. S. Marks, Deep generative models of genetic variation capture the effects of mutations. *Nature Methods* **15** (10), 816–822 (2018).

3. J. G. Greener, L. Moffat, D. T. Jones, Design of metalloproteins and novel protein folds using variational autoencoders. *Scientific Reports* **8** (1), 16189 (2018).

4. A. Hawkins-Hooker, *et al.*, Generating functional protein variants with variational autoencoders. *PLoS Computational Biology* **17** (2), e1008736 (2021).

5. S. Lyu, S. Sowlati-Hashjin, M. Garton, Variational autoencoder for design of synthetic viral vector serotypes. *Nature Machine Intelligence* **6** (2), 147–160 (2024).

6. E. Castro, *et al.*, Transformer-based protein generation with regularized latent space optimization. *Nature Machine Intelligence* **4** (10), 840–851 (2022).

7. N. Ferruz, S. Schmidt, B. Höcker, ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications* **13** (1), 4348 (2022).

8. E. Sevgen, *et al.*, ProT-VAE: protein transformer variational autoencoder for functional protein design. *bioRxiv* pp. 2023–01 (2023).

9. C. Frank, *et al.*, Scalable protein design using optimization in a relaxed sequence space. *Science* **386** (6720), 439–445 (2024).

10. Y. Ming, *et al.*, A review of enzyme design in catalytic stability by artificial intelligence. *Briefings in Bioinformatics* **24** (3), bbad065 (2023).

11. P. Notin, N. Rollins, Y. Gal, C. Sander, D. Marks, Machine learning for functional protein design. *Nature Biotechnology* **42** (2), 216–228 (2024).

12. A. Winnifrith, C. Outeiral, B. L. Hie, Generative artificial intelligence for de novo protein design. *Current Opinion in Structural Biology* **86**, 102794 (2024).

13. X. Ding, Z. Zou, C. L. Brooks III, Deciphering protein evolution and fitness landscapes with latent space models. *Nature Communications* **10** (1), 5644 (2019).

14. C. Ziegler, J. Martin, C. Sinner, F. Morcos, Latent generative landscapes as maps of functional diversity in protein sequence space. *Nature Communications* **14** (1), 2222 (2023).

15. M. Dyla, M. Kjærgaard, H. Poulsen, P. Nissen, Structure and mechanism of P-type ATPase ion pumps. *Annual Review of Biochemistry* **89** (1), 583–603 (2020).

16. A. C. Rosenzweig, J. M. Argüello, Toward a molecular understanding of metal transport by P_{1B} ATPases, in *Current Topics in Membranes* (Elsevier), vol. 69, pp. 113–136 (2012).

17. N. Abeyrathna, S. Abeyrathna, M. T. Morgan, C. J. Fahrni, G. Meloni, Transmembrane Cu (I) P-type ATPase pumps are electrogenic uniporters. *Dalton Transactions* **49** (45), 16082–16094 (2020).

18. A. T. Smith, K. P. Smith, A. C. Rosenzweig, Diversity of the metal-transporting P_{1B}-type ATPases. *JBIC Journal of Biological Inorganic Chemistry* **19**, 947–960 (2014).

19. M. Andersson, *et al.*, Copper-transporting P-type ATPases use a unique ion-release pathway. *Nature Structural & Molecular Biology* **21** (1), 43–48 (2014).

20. C. Zhou, P. Lu, De novo design of membrane transport proteins. *Proteins: Structure, Function, and Bioinformatics* **90** (10), 1800–1806 (2022).

21. S. Sowlati-Hashjin, A. Gandhi, M. Garton, Dawn of a new era for membrane protein design. *BioDesign Research* **2022**, 9791435 (2022).

22. C. A. Goverde, *et al.*, Computational design of soluble and functional membrane protein analogues. *Nature* pp. 1–10 (2024).

23. B. J. Hardy, P. Curnow, Computational design of de novo bioenergetic membrane proteins. *Biochemical Society Transactions* **52** (4), 1737–1745 (2024).

24. R. Bitter, *et al.*, Structure of the Wilson Disease Copper Transporter ATP7B, *Science Advances* **2022**, 8 (9), eabl5508.

25. O. Sitsel, *et al.*, Structure and function of Cu(I)- and Zn(II)-ATPases. *Biochemistry* **54** (37), 5673–5683 (2015).

26. P. Gourdon, *et al.*, Crystal structure of a copper-transporting P_{1B}-type ATPase. *Nature* **475** (7354), 59–64 (2011).

27. M. J. Gallenito, G. W. Irvine, L. Zhang, G. Meloni, Coordination promiscuity guarantees metal substrate selection in transmembrane primary-active Zn^{2+} pumps. *Chemical Communications* **55** (73), 10844–10847 (2019).

28. K. Tsirigos, C. Peters, N. Shu, L. Käll, A. Elofsson, The TOPCONS web server for combined membrane protein topology and signal peptide prediction. *Nucleic Acids Research* **43**, W401–W407 (2015).

29. M. A. Lomize, I. D. Pogozheva, H. Joo, H. I. Mosberg, A. L. Lomize, OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Research* **40** (D1), D370–D376 (2012).

30. Y. Zhang, J. Skolnick, Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics* **57** (4), 702–710 (2004).

31. N. Salustros, *et al.*, Structural basis of ion uptake in copper-transporting P_{1B} ATPases. *Nature Communications* **13** (1), 5121 (2022).

32. Z. Guo, *et al.*, Diverse roles of the metal binding domains and transport mechanism of copper transporting P-type ATPases. *Nature Communications* **15** (1), 2690 (2024).

33. G.-M. Yang, *et al.*, Structures of the human Wilson disease copper transporter ATP7B. *Cell Reports* **42** (5) (2023).

34. R. Aguayo-Ortiz, L. M. Espinoza-Fonseca, Linking biochemical and structural states of SERCA: achievements, challenges, and new opportunities. *International Journal of Molecular Sciences* **21** (11), 4146 (2020).

35. E. Chovancova, *et al.*, CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures (2012).

36. A. K. Mandal, W. D. Cheung, J. M. Arguello, Characterization of a Thermophilic P-type Ag+/Cu+-ATPase from the ExtremophileArchaeoglobus fulgidus. *Journal of Biological Chemistry* **277** (9), 7201–7208 (2002).

37. J. Smirnova, *et al.*, Copper (I)-binding properties of de-coppering drugs for the treatment of Wilson disease. α-Lipoic acid as a potential anti-copper agent. *Scientific Reports* **8** (1), 1463 (2018).

38. C. Große, G. Schleuder, C. Schmole, D. H. Nies, Survival of Escherichia coli cells on solid copper surfaces is increased by glutathione. *Applied and Environmental Microbiology* **80** (22), 7071–7078 (2014).

39. S. R. Eddy, Accelerated profile HMM searches. *PLoS Computational Biology* **7** (10), e1002195 (2011).

40. F. Morcos, *et al.*, Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences* **108** (49), E1293–E1301 (2011).

41. M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, T. Hwa, Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences* **106** (1), 67–72 (2009).

42. R. M. Levy, A. Haldane, W. F. Flynn, Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Current Opinion in Structural Biology* **43**, 55–62 (2017).

43. S. Alvarez, *et al.*, In vivo functional phenotypes from a computational epistatic model of evolution. *Proceedings of the National Academy of Sciences* **121** (6), e2308895121 (2024).

44. D. P. Kingma, M. Welling, *et al.*, An introduction to variational autoencoders. *Foundations and Trends in Machine Learning* **12** (4), 307–392 (2019).

45. A. Waterhouse, *et al.*, SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research* **46** (W1), W296–W303 (2018).

46. N. Guex, M. C. Peitsch, T. Schwede, Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *Electrophoresis* **30** (S1), S162–S173 (2009).

47. R. Anandakrishnan, B. Aguilar, A. V. Onufriev, H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Research* **40** (W1), W537–W541 (2012).

48. J. Myers, G. Grothaus, S. Narayanan, A. Onufriev, A simple clustering algorithm can be accurate enough for use in calculations of pKs in macromolecules. *Proteins: Structure, Function, and Bioinformatics* **63** (4), 928–938 (2006).

49. J. C. Gordon, *et al.*, H++: a server for estimating pK$_a$s as and adding missing hydrogens to macromolecules. *Nucleic Acids Research* **33** (suppl_2), W368–W371 (2005).

50. D. A. Case, *et al.*, *Amber 2021* (University of California, San Francisco) (2021).

51. S. Schott-Verdugo, H. Gohlke, PACKMOL-memgen: a simple-to-use, generalized workflow for membrane-protein–lipid-bilayer system building. *Journal of Chemical Information and Modeling* **59** (6), 2522–2528 (2019).

52. L. Martínez, R. Andrade, E. G. Birgin, J. M. Martínez, PACKMOL: A package for building initial configurations for molecular dynamics simulations. *Journal of Computational Chemistry* **30** (13), 2157–2164 (2009).

53. J. A. Maier, *et al.*, ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of Chemical Theory and Computation* **11** (8), 3696–3713 (2015).

54. I. Gould, A. Skjevik, C. Dickson, B. Madej, R. Walker, Lipid17: A comprehensive AMBER force field for the simulation of zwitterionic and anionic lipids. *Manuscript in Preparation* (2018).

55. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79** (2), 926–935 (1983).

56. I. S. Joung, T. E. Cheatham III, Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *The Journal of Physical Chemistry B* **112** (30), 9020–9041 (2008).

57. R. J. Loncharich, B. R. Brooks, R. W. Pastor, Langevin dynamics of peptides: The frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. *Biopolymers: Original Research on Biomolecules* **32** (5), 523–535 (1992).

58. J.-P. Ryckaert, G. Ciccotti, H. J. Berendsen, Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **23** (3), 327–341 (1977).

59. W. Humphrey, A. Dalke, K. Schulten, VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* **14**, 33–38 (1996).

60. J. Stone, An Efficient Library for Parallel Ray Tracing and Animation, Master's thesis, Computer Science Department, University of Missouri-Rolla (1998).

61. J. A. Licht, S. P. Berry, M. A. Gutierrez, R. Gaudet, They all rock: A systematic comparison of conformational movements in LeuT-fold transporters. *bioRxiv* (2024).

62. X. Robert, P. Gouet, Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Research* **42** (W1), W320–W324 (2014).

63. F. Sievers, *et al.*, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7** (1), 539 (2011).

64. M. Goujon, *et al.*, A new bioinformatics analysis tools framework at EMBL–EBI. *Nucleic Acids Research* **38** (suppl_2), W695–W699 (2010).

Acknowledgments:

The authors thank the Office of Information Technology and the Cyber Infrastructure Research Computing (CIRC) at the University of Texas at Dallas as well as the Texas Advanced Computing Center (TACC) at the University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper.

The authors also thank Prof. Dietrich Nies, Martin-Luther University Halle-Wittenberg, Germany, for providing the copper-sensitive *E. Coli* ECA464 strain for this study.

Funding: This research was funded by the University of Texas at Dallas (to H.T.), the NSF Faculty Early Career Development (CAREER) Program grant numbers MCB-1943442 (to F.M.) and CLP-2045984 (to G.M.), and the Robert A. Welch Foundation AT-2073-20210327 (to G.M). Furthermore, research reported in this publication was supported by the National Institutes of Health under award number R35GM155106 (to H.T.), R35GM128704 (to G.M.), and R35GM133631 (to F.M.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions: F.M.O., F.H., G.M., H.T., and F.M. designed research; F.M.O., F.H., and V.V.V. performed research; F.M.O., F.H., V.V.V., G.M., H.T., and F.M. analyzed data; and F.M.O., F.H., G.M., H.T., and F.M. wrote the paper.

Competing interests: The authors declare no competing interest.

Data and materials availability: The LGL framework used in this study is publicly available at https://github.com/morcoslab/LGL-VAE/. The LGL training dataset, the produced LGL model, a cumulative list of decoded sequences (GS1-GS8) along with MD simulation input files and experimental raw data are uploaded to Zenodo: 10.5281/zenodo.14783470. These files can be accessed through the following link: https://tinyurl.com/4fjwcuhj.