Adam Mickiewicz University
Poznań
Faculty of Biology

Developing and exploiting state-specific descriptors of conformational cycle of ABC transporters.

Master's Thesis in Bioinformatics

Maciej Orski-Hryc
machry@st.amu.edu.pl
ORCID: 0009-0009-6707-7727

Laboratory of Biomolecular Interactions and Transport
Institute of Molecular Biology and Biotechnology

Supervisor: dr hab. Jan Brezovský

July 06, 2026

I'd like to give my sincere thanks to prof. Brezovský, who took me in and provided valuable guidance through the whole project. I'd also like to thank the lab members: Aaftaab, Abraham, Igor, Nitheesh and Ola, for all the discussions we've had, their advice and friendly atmosphere.

Abstract

ATP-binding cassette (ABC) transporters move substrates through biological membranes by cycling through inward-facing (IF) and outward-facing (OF) conformations, powered by ATP hydrolysis without protein phosphorylation. While the general principles of their mechanism are known, precise residue level characterization is still incomplete. Low number of experimental structure models, covering only a fraction of the protein-conformation space, makes *in silico* study of ABC transport mechanism a difficult task.

Recently a coevolution-powered machine learning method of finding conformation specific residue contacts in transporters of the sugar porter was developed by **(Mitrovic et al., 2023)**. The method relies on coevolutionary filtering of protein contact maps, and convolutional neural network classifier coupled with layerwise relevance propagation, to assign per conformation importance to contact map positions.

This work attempts to extend their approach to the G subfamily of ABC transporters, with the aspirational goal of applying identified state specific contactsfor *in silico* studies of *Medicago truncatula* ABCG46 computer predicted structure. Specifics of the data and studied system made a direct reproduction impossible, leading to development of a modified method based on using linear kernel support vector machine classifier (LinearSVC) coefficients for assigning feature importance.

Poor classification performance and high coefficient instability make the utility of this modified approach inconclusive.

I discuss aspects of ABCG biology and available data we believe are the most problematic, and propose technical improvements for future reference.

List of Abbreviations

- ABC — ATP binding cassette
- ABCG — ATP binding cassette G subfamily
- AF — AlphaFold
- APC — average product correction
- CNN — convolutional neural network
- DCA — direct coupling analysis
- EMDB — electron microscopy data bank
- FAB — fragment antigen-binding
- IF — inward facing (conformation)
- KEGG — Kyoto encyclopedia of genes and genomes
- LOO — Leave-One-Out cross validation
- MD — molecular dynamics
- MDR — multidrug resistance
- ML — machine learning
- MRF — markov random field
- NBD — nucleotide binding domain
- NBS — nucleotide binding site
- OF — outward facing (conformation)
- OPM — orientation of proteins in membranes database
- PDB — protein data bank
- PDR — pleiotropic drug resistance
- RCSB — research collaboratory for structural bioinformatics
- SVC — Support Vector machine Classifier
- SVM — support vector machine
- TED — the encyclopedia of domains
- TMD — transmembrane domain

Contents

List of Figures

List of Tables

1. Introduction

1.1. Biological background

ATP Binding Cassette transporters (ABC transporters) is a superfamily of ubiquitous transmembrane transporters **Figure 1**, that includes both uptake and efflux systems **(Thomas et al., 2020)**. ABCs function as active transporters using ATP hydrolysis without protein phosphorylation as energy source **Equation 1 (Thomas et al., 2020)**.

Upon binding of ATP molecules to nucleotide binding domains (NBDs), ABC exporters undergo a conformational transition from an inward facing (IF) to an outward facing (OF) state via chemo-mechanical coupling. The ATP binding energy is converted into distortion energy of several transmembrane helices **(Hayashi et al., 2014)**, **(Arai et al., 2017)**.

$$\text{Substrate (side A)} + \text{IF} + 2\text{ATP} \rightarrow \text{Substrate (side B)} + \text{OF} + 2\text{ADP} + 2P_i \qquad 1.$$

All ABC transporters share a common architecture consisting of 2 highly conserved nucleotide-binding domains (NBDs) and 2 family specific transmembrane domains (TMDs) which are the basis for modern ABC classification **(Alam & Locher, 2023; Thomas et al., 2020)**. ABC architecture can be divided into two broad classes: half-size transporters, where a functional system is a homo- or heterodimer of two subunits with one NBD and TMD each (eg. human ABCG2 homodimer or human ABCG5/G8 heterodimer) **(Xin et al., 2024)**, or full-size transporters, where a single chain roughly double the length of a half-size monomer folds into a full system with two NBDs and two TMDS (eg. *S. cerevisiae* Pdr5 or *A. thaliana* ABCG40) **(Harris et al., 2021)**. A functional ABC has an approximate two-fold rotational symmetry along the axis perpendicular to the membrane, but transport mechanism is often asymmetric, even in homodimer systems **(Aittoniemi et al., 2010)**, **(Locher, 2016)**, **(Jones & George, 2023)**.

1.1.1. The G family (ABCG)

Subject of this study is the G subfamily of ABC transporters, which is classified as type V transporter by **(Thomas et al., 2020)**, and is one of the two families with highest number of experimental structures (50+) **(Lomize et al., 2011)**.

ABCGs are particularly numerous in plants **Table 1**. It's proposed that their evolution enabled transport of compounds critical to plant terrestrial lifestyle, including: phytohormones, lignin precursors and hydrophobic compounds for coating leaf surface **(Do et al., 2018)**, **(Hwang et al., 2016)**.

In humans only five ABCGs are known: G1, G2, G4, G5, G8 **(Vasiliou et al., 2008)**, **(Almén et al., 2009)**.

Scientific name	ABCG-half	ABCG-full
Saccharomyces cerevisiae	1	9
Physcomitrella paten	23	19
Selaginella moellendorffii	28	24
Oryza sativa	30	20
Arabidopsis thaliana	28	15
Caenorhabditis elegans	9	0
Drosophila melanogaster	15	0
Mus musculus	6	0
Homo sapiens	5	0

Table 1: No. of ABCG transporter genes in selected species. Heterodimer genes counted separately. Table adapted from **(Hwang et al., 2016)**.

1.1.1.1. Examples of ABCG transporters

ABCGs are responsible for transporting a wide variety of substrates, and perform different biological functions, eg. single substrate specific phytohormone transporter or export of wide variety of drugs or toxins.

Examples of ABCGs with different substrate specificity and function:

- human ABCG1 is involved in cholesterol efflux in macrophages.
- human ABCG2, also known as breast cancer resistance protein (BCRP) **(Natarajan et al., 2012)**, functions as a multidrug resistance transporter. It's capable of transporting steroids (cholesterol, oestradiol, progesterone and testosterone) and certain chlorophyll metabolites, as well as organic anions. Overexpression of ABCG2 in tumour tissues leads to export of anti-cancer drugs, interfering in chemotherapy **(Hasanabady & Kalalinia, 2016)**, making the study of ABCG2 transport mechanism a relevant problem of biomedical sciences.
- human ABCG5 and ABCG8 fucntion as a heterodimer ABCG5/G8. Both appear to limit intestinal absorption and promote biliary excretion of sterols. Mutations in either of these two genes lead to sitosterolaemia **(Vasiliou et al., 2008)**.
- *Arabidopsis thaliana* ABCG25, ABCG40, ABCG30 and ABCG31 function as abscisic acid (ABA) transporters in seeds, mediating ABA secretion from endosperm and uptake into the embryo **(Kuromori & Shinozaki, 2010)**, **(Kuromori et al., 2010)**, **(Kang et al., 2015)**.
- *Saccharomyces cerevisiae* pleiotropic drug resistance protein 5 (Pdr5) is responsible for efflux of a wide variety of small weakly charged organic compounds, including herbicides, mycotoxins and antibiotics **(Harris et al., 2021)**. Pdr5 has only one functional NBD.

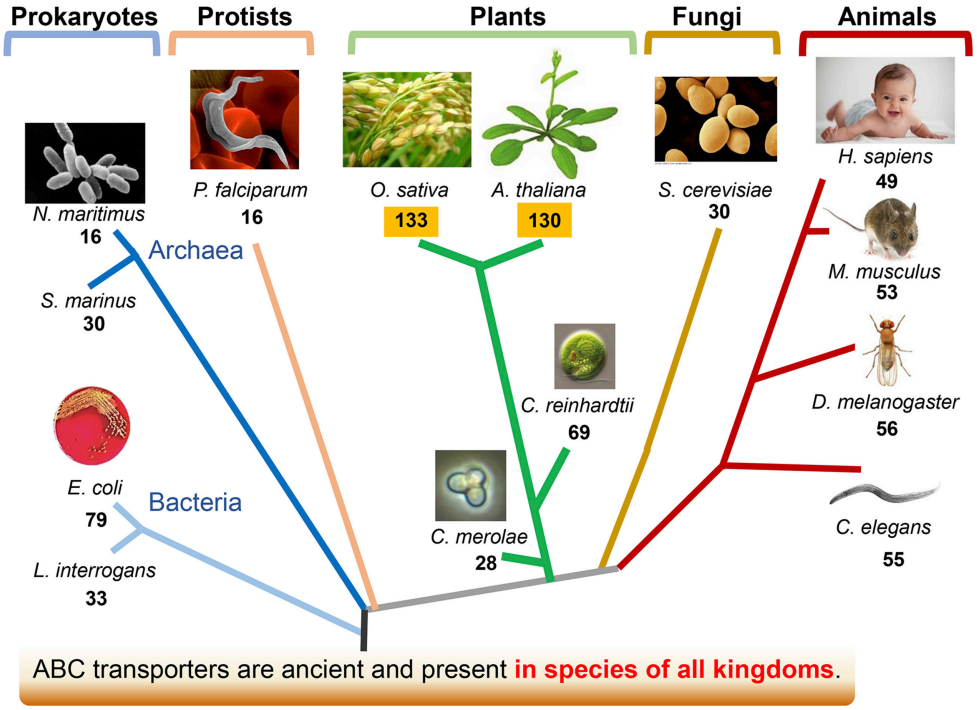


Figure 1: Numbers below the species name indicate the numbers of ABC transporters in that organism. For example, rice has 133 different ABC transporters, while humans have only 49. The number of ABC transporters in higher plant species is highlighted in yellow. Figure adapted from **(Hwang et al., 2016)**.

1.1.2. ABCG Topology and structure

All ABCGs have a common TMD architecture of 6+6 helices anchoring the transporter in biological membrane **Figure 5** and creating a transport channel through the membrane **Figure 3**. Between helices 5 and 6 there's a characteristic V-shaped re-entry helix **Figure 3**, **Figure 4**.

The ABCG NBD is the same as in other ABCs, and contains motifs typical for ATP binding subunits, for eg. the Walker motifs **Figure 2**, **(Schneider & Hunke, 1998)** and a LSGGQ signature motif.

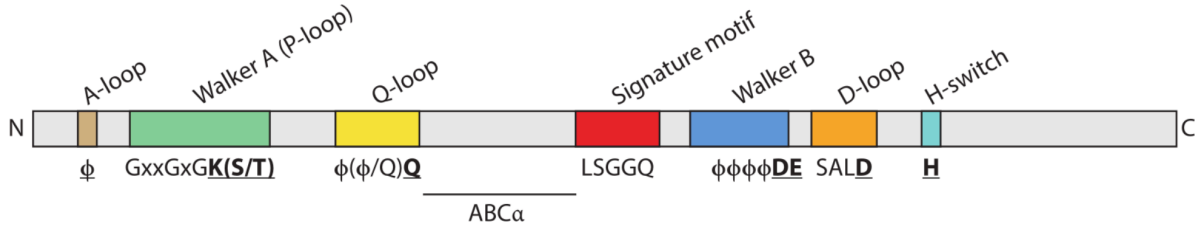


Figure 2: Motifs found in a functional NBD of ABC transporters. Figure adapted from **(Thomas & Tampé, 2020)**.

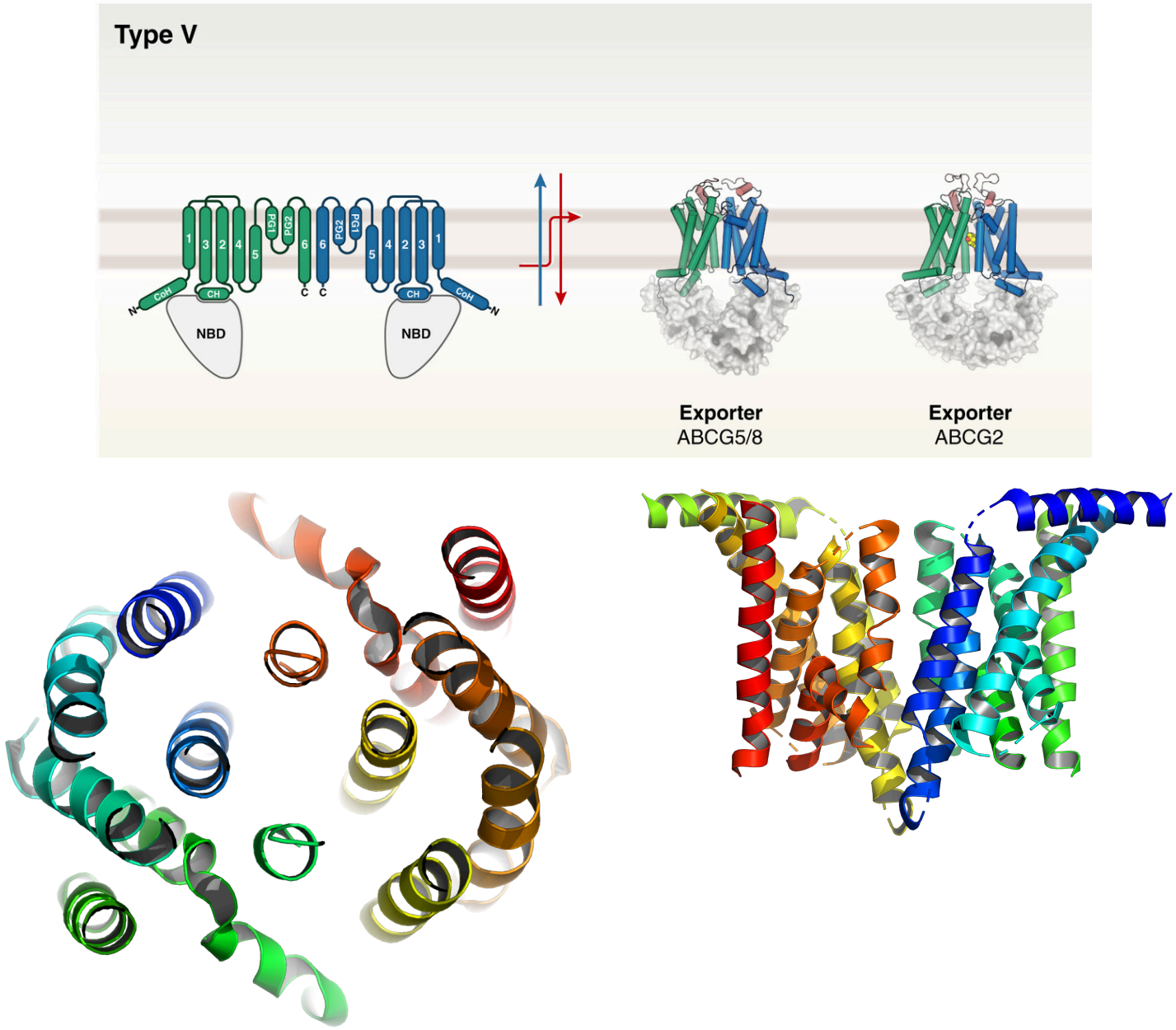


Figure 3: **Top**: General fold of ABC type V transporters with two representative experimental structures (human ABCG5/8 and ABCG2). Each monomer's transmembrane domain (TMD) has six α-helices (1-6), characteristic re-entry helix (here labelled **PG1/PG2**) and elbow helix (here labelled **CoH**). Direction of substrate transport indicated with arrows, blue arrow indicates the direction of transport in all members of ABC G subfamily (ABCG), which are exclusively exporters. **Bottom left**: Orthographic projection perpendicular to the channel formed by helices 2 and 6. **Bottom right**: Orthographic projection of TMD dimer in inward facing conformation viewed from the side. Extracellular side oriented downwards. **Top**: Adapted from **(Thomas et al., 2020)**; structure accession codes: ABCG5/8: 5DO7 **(Lee et al., 2016)**; ABCG2: 6HCO **(Manolaridis et al., 2018)**; **Bottom**: original work.

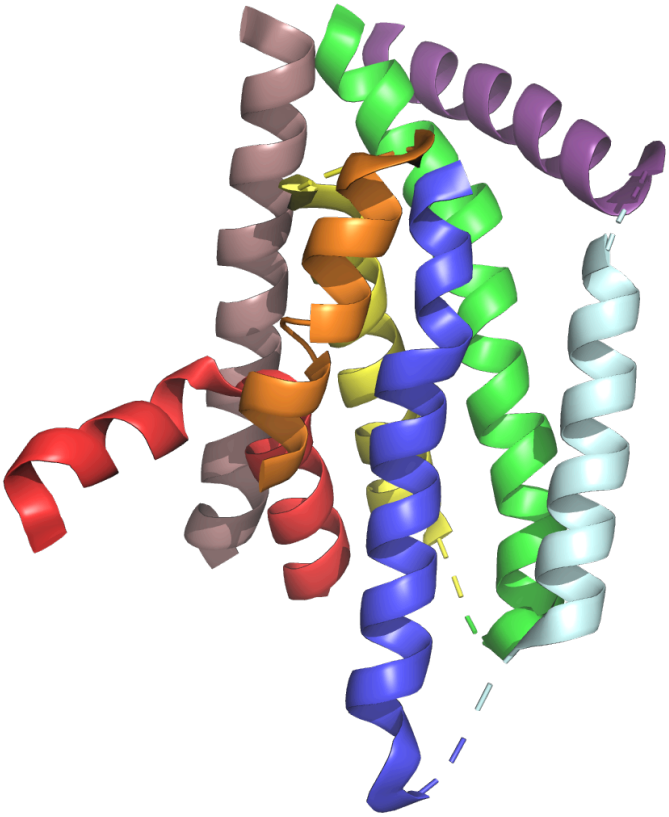


Figure 4: Single transmembrane domain (TMD) of Pdr5 of *Saccharomyces cerevisiae*. Colour codings: **purple**: elbow, **light blue**: 1st helix, **dark** blue:} 2nd helix, **green**: 3rd helix, **yellow**: 4th helix, **orange**: 5th helix, **red**: re-entry helix, **brown**: 6th helix. Original work annotation based on **(Harris et al., 2021)**.

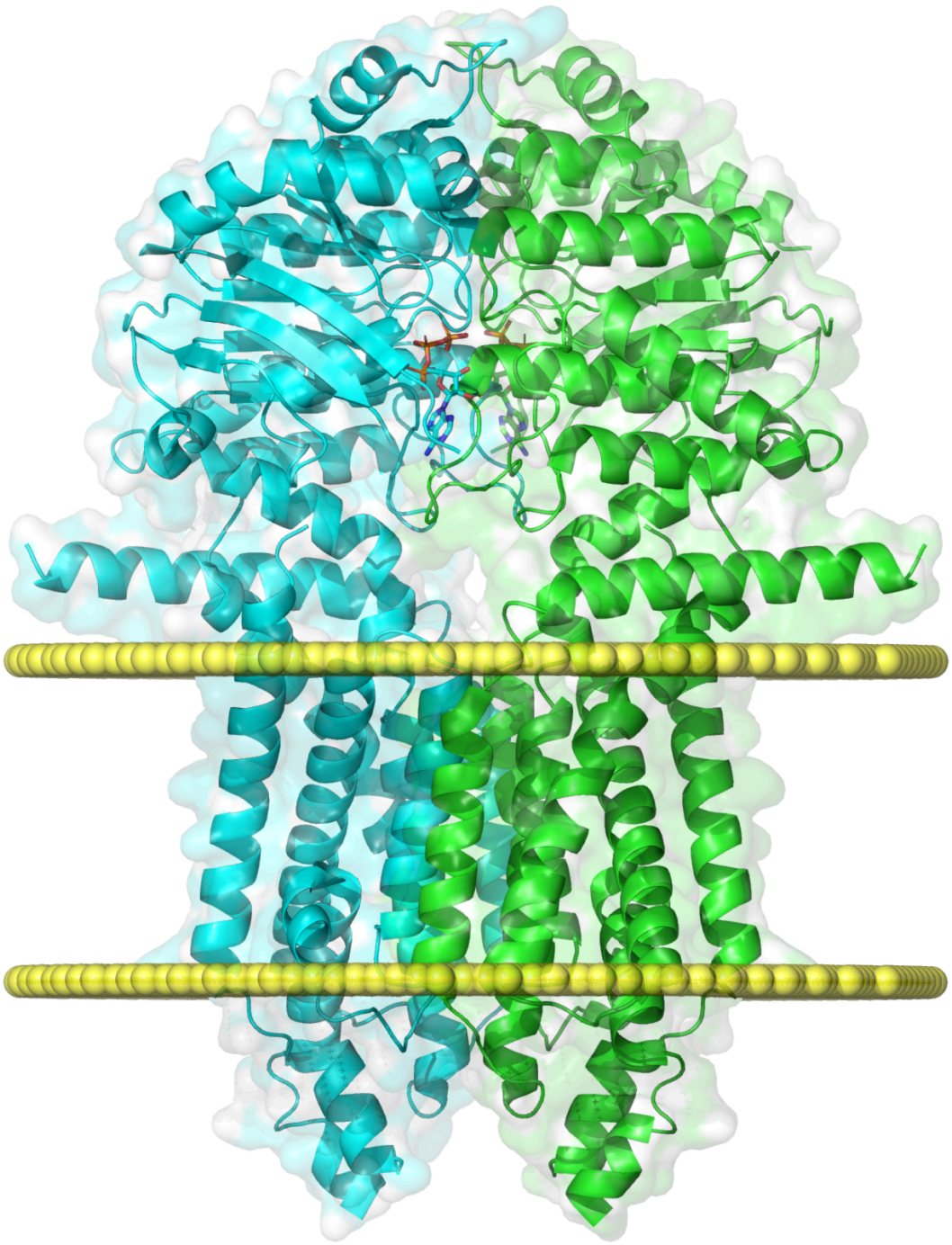


Figure 5: Structure of *A. thaliana* ABCG25 ABA exporter in outwards facing conformation, with two molecules of ATP bound in its NBDs. Yellow spheres show orientation of ABCG25 in cellular membrane as computed by Positioning of Proteins in Membranes (PPM 3.0) web server **(Lomize et al., 2021)**. PDB accession code 8WAM **(Xin et al., 2024)**. Original work.

1.1.3. ABCG TRANSPORT MECHANISM

In the resting state IF conformation, ABCGs have roughly parallel TMD helices and NBDs far apart, allowing the entry of substrate and ATP. ATP molecules act as molecular glue and induce dimerization of NBDs that drives the conformational change from IF to OF. Movement

of TMDs exposes substrate to extracellular fluid which washes the substrate out of the transporter. ATP hydrolysis energizes the return to apo-IF resting conformation, and closes the cycle **Figure 6**, **Figure 7**, **(Yu et al., 2021)**, **(Peng et al., 2024)**, **(Khunweeraphong & Kuchler, 2025)**.

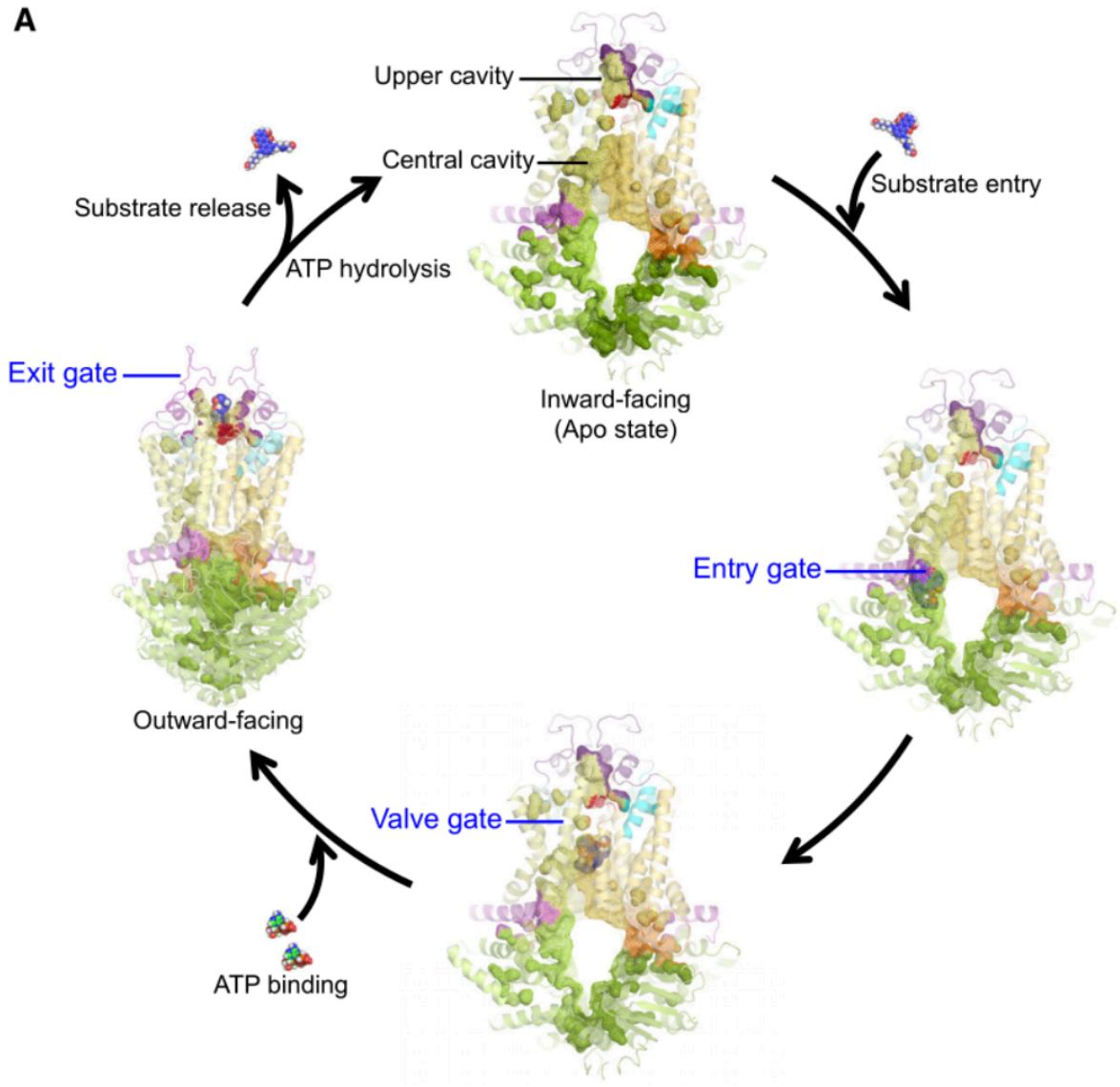


Figure 6: ABCG2 transport mechanism proposed by **(Khunweeraphong & Kuchler, 2025)** highlighting three gates controling drug extrusion.

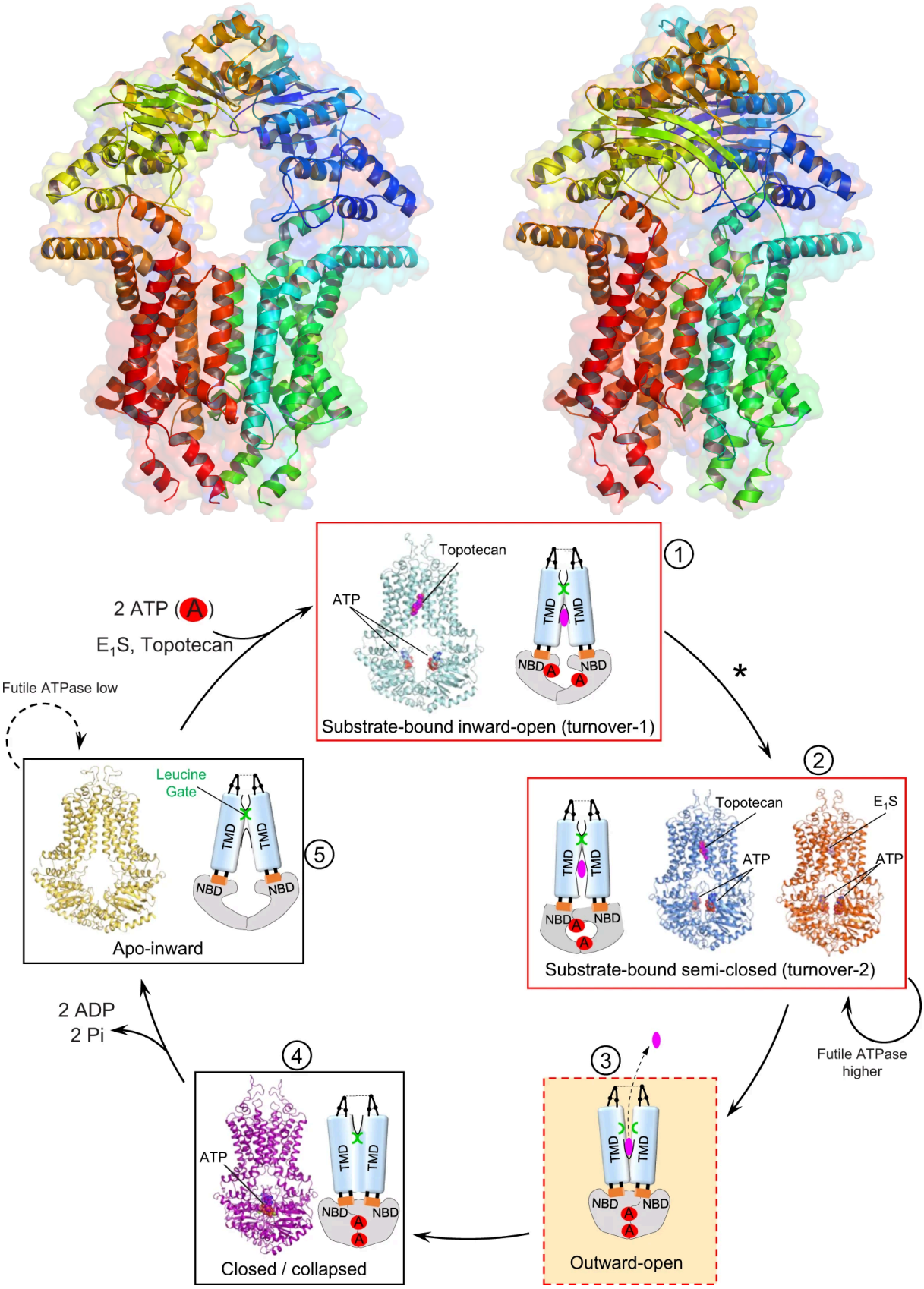


Figure 7: **Top**: Examples of Inward Facing (IF; left) and Outward Facing (OF; right) conformations from *Arabidopsis thaliana* ABCG25. **Bottom**: ABCG2 transport mechanism proposed by **(Yu et al., 2021)** with a hypothesized missing outward open structure. **Top**: original work, accession codes: PDB: 8WD6, 8WAM **(Xin et al., 2024)**; UniProt: Q84TH5; **Bottom**: figure adapted from **(Yu et al., 2021)**.

1.2. Goals of the project

In 2023 **(Mitrovic et al., 2023)** published a paper titled *"Reconstructing the transport cycle in the sugar porter superfamily using coevolution-powered machine learning"* describing a method of identifying state-specific residue-residue contacts, and using them to predict structures of all known conformations in sugar porter transporters **Figure 8**, **Figure 9**, **Figure 10**.

Outline of the method:

1. collect a large (serveral thousand) number of protein sequences of a particular functional group,
2. compute their multiple sequence alignment (MSA),
3. apply direct coupling analysis (DCA) to predict residue-residue contacts from evolutionary couplings (brief explanation of theory behind this method in **Section 1.3**)
4. prepare a dataset of relevant experimental structures,
5. compute experimental structure contact maps,
6. select contact map positions with high evolutionary coupling scores,
7. train a convolutional neural network (CNN) for conformation classifiaion based on filtered contact maps,
8. apply layerwise relevance propagation (LRP) **(Bach et al., 2015)** to identify contact map pixels contributing the most to classification of each conformation,
9. identified conformation specific contacts can be used for state specific structure prediction in rosetta, or for driving molecular dynamics simulations (MD simulations).

This project aims to extend the outlined approach to ABCG transporters, which can be considered a more difficult system for the following reasons:

1. complex architecture — unlike sugar porters, ABCGs have distinct transmembrane and nucleotide binding domains, that perform distinct functions,
2. architecture variants — ABCGs can function as single chain full-size transporters, or homo- or heterodimers of smaller half-size transporters,
3. larger size — functional ABCG transporter on average are composed of roughly twice as many amino acids compared to sugar porters.

Work was carried out with the ultimate objective of applying identified ABCG state specific residue contacts to study the transport mechanism of *Medicago truncatula* ABCG46.

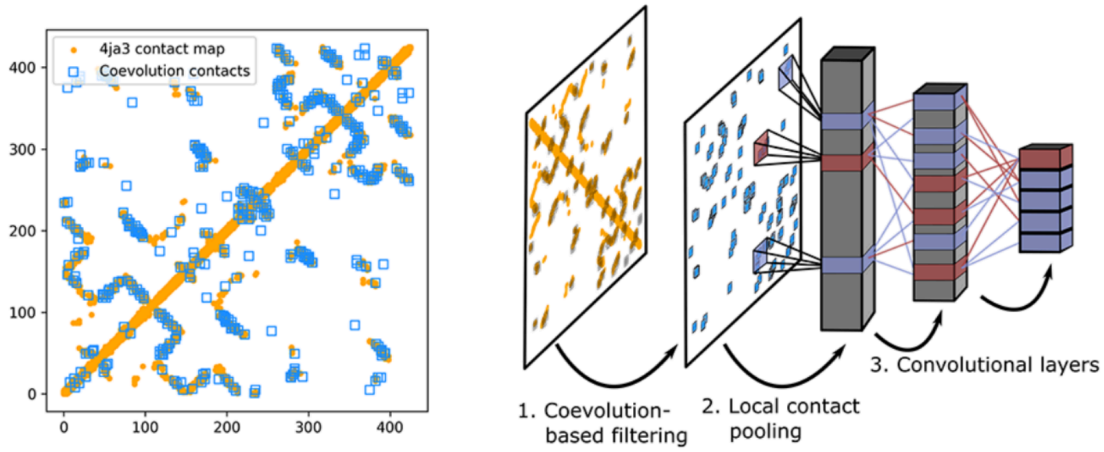


Figure 8: **Left**: Example sugar porter experimental contact map overlaid with coevolutional contacts identified by DCA. **Right**: outline of the method. Contact maps filtered by DCA coevolution scores are input for convolutional neural network (CNN) classifying transporter conformations. Figure adapted from **(Mitrovic et al., 2023)**.

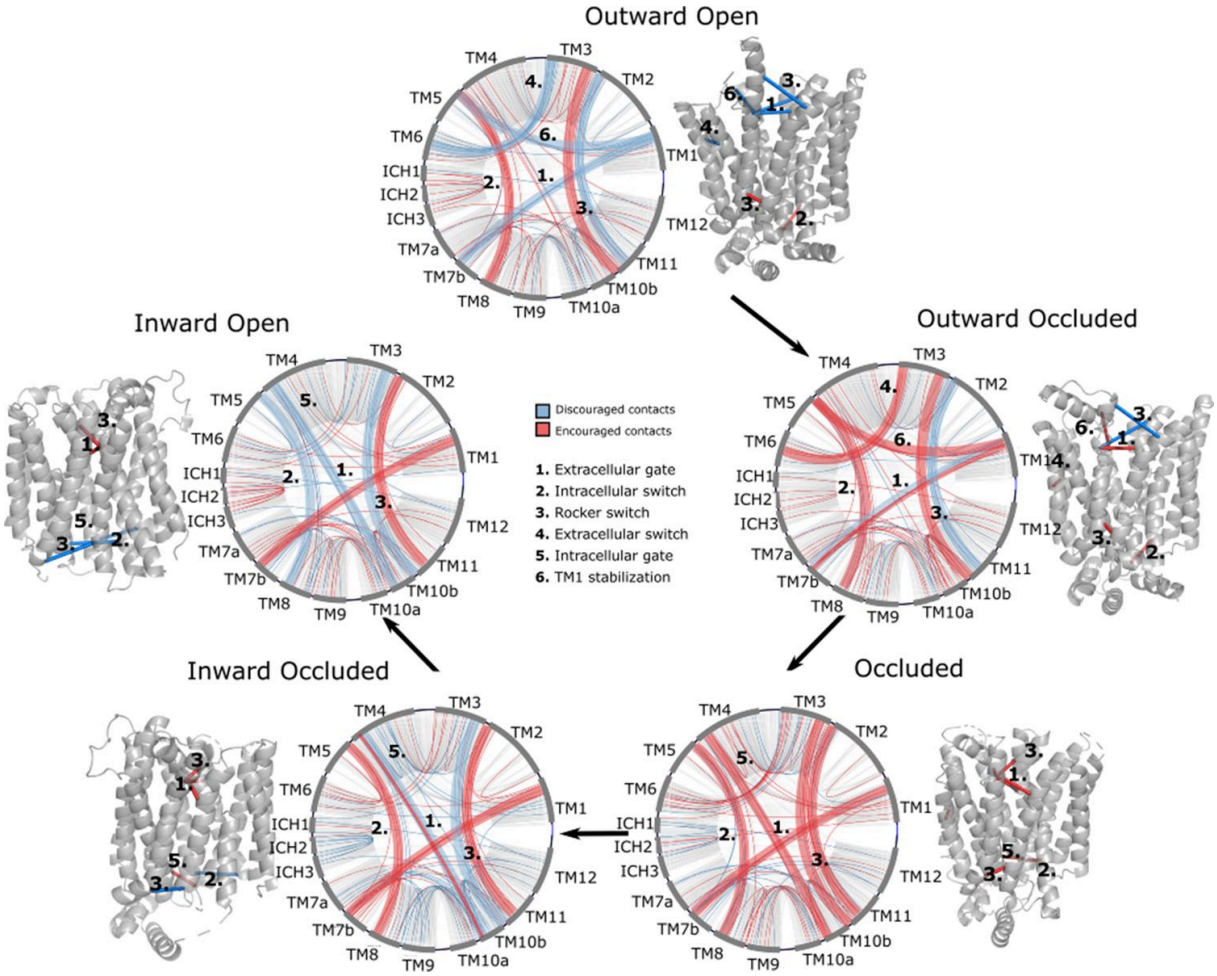
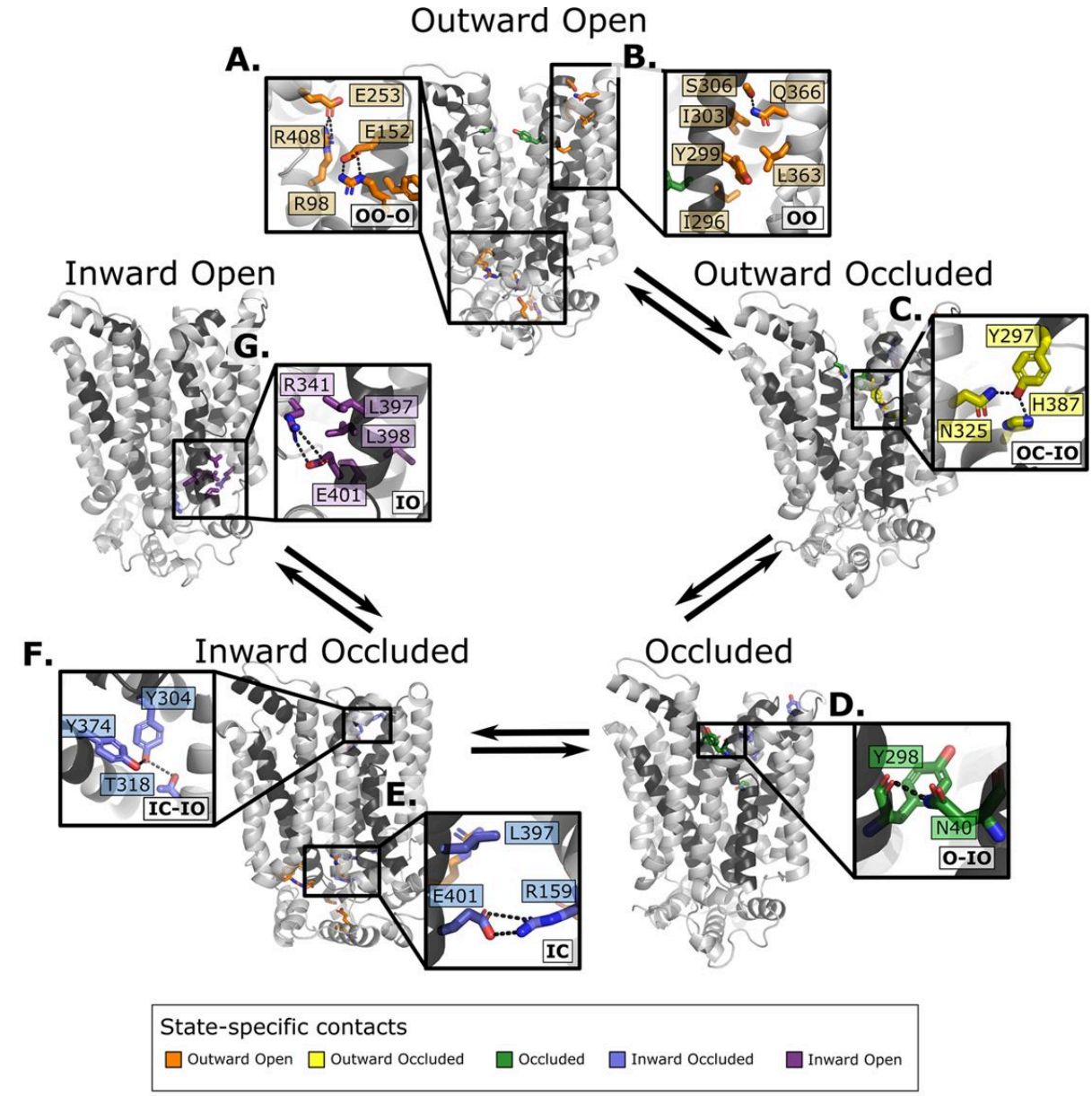


Figure 9: Network representation of state specific contacts from layer-wise relevance propagation (LRP). Blue — discouraged contacts, red — encouraged contacts. Figure adapted from **(Mitrovic et al., 2023)**.



H.

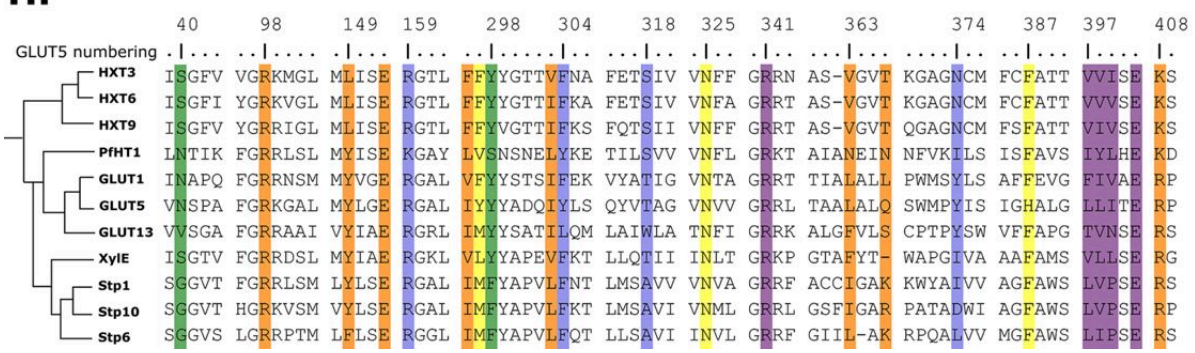


Figure 10: Visualization of state specific contacts. **Top**: sugar porter conformations with relevant contacts highlighted. **Bottom**: MSA view of state specific contacts. Figure adapted from **(Mitrovic et al., 2023)**.

1.3. Predicting residue contacts with probabilistic graphical models

If positions in protein sequence exert selection pressures on each other, for eg. by charge or size interactions, they are expected to evolve in tandem. This process of coevolution leads to correlations between positions in multiple sequence alignments (MSA) of protein families. Such coevolutionary information can be utilized to study 3D structure of protein families **Figure 11**, **(Morcos et al., 2011)**, **(Sułkowska et al., 2012)**, **(Nugent & Jones, 2012)**, **(Juan et al., 2013)**.

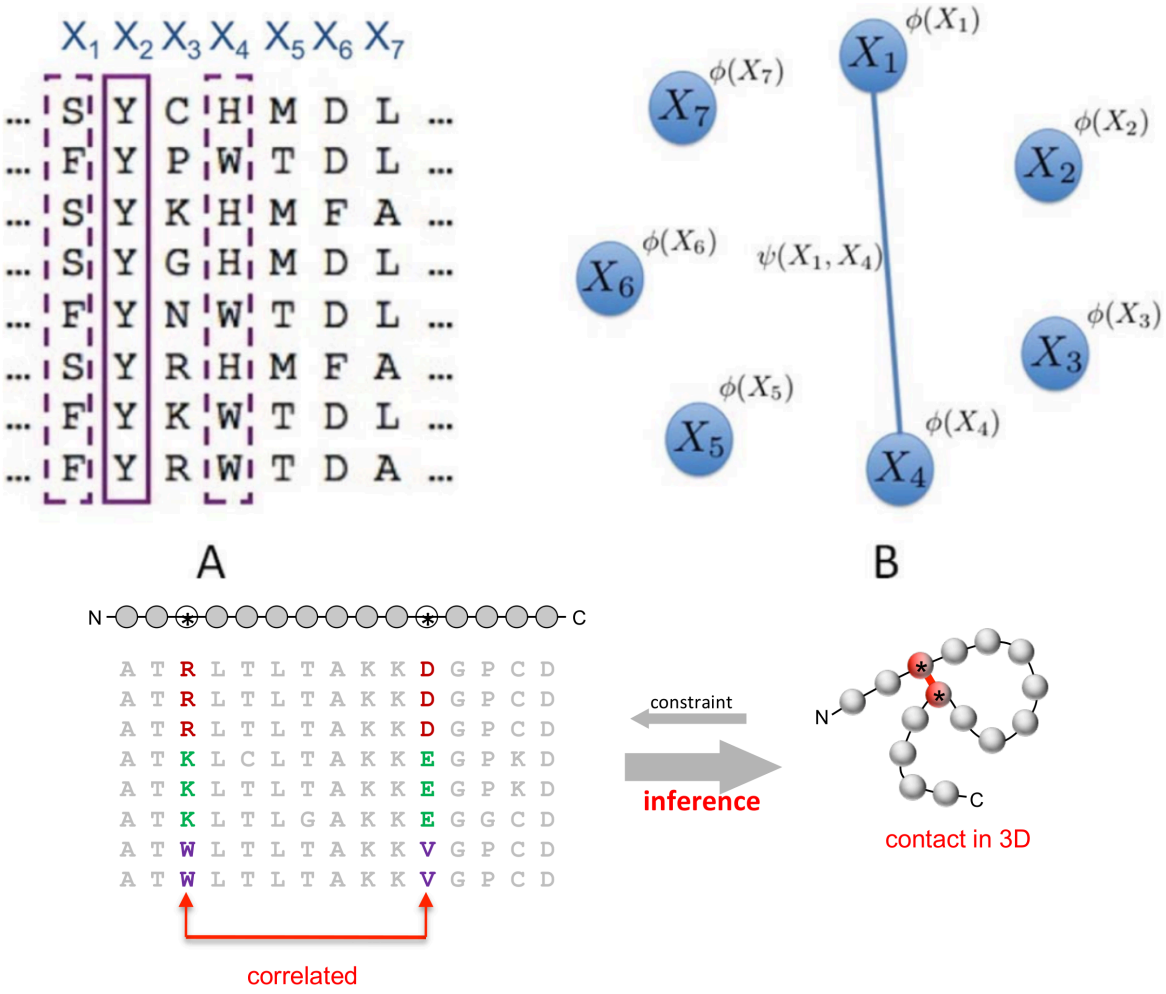


Figure 11: **Top**: (**A**) A multiple sequence alignment (MSA) for a hypothetical domain family. Positions X_1 and X_4 strongly coevolve, only appearing in pairs (S, H) and (F, W) (**B**) The Markov Random Field encoding the conservation in and the coupling in the MSA. The edge $\psi(X_1, X_4)$ between random variables X_1 and X_4 reflects the coupling between positions 1 and 4 in the MSA. **Bottom**: (A) MSA showing correlated positions (B) visualization of contact between correlated positions imposing a constraint and resulting in observed coevolution. Inference of 3D contacts from MSA correlation is the goal of direct coupling analysis. **Top**: figure adapted from **(Balakrishnan et al., 2011)**; **Bottom**: figure adapted from **(Marks et al., 2011)**.

1.3.1. Mathematical formulation, markov random fields (MRFs)

In case of protein MSA a markov random field (MRF) is a tuple:

$$\mathcal{M} = (\mathcal{V}, \mathcal{E}, \varphi, \psi) \tag{2.}$$

where $\mathcal{V}$ is a set of vertices representing MSA columns, and $\mathcal{E}$ is a set of edges connecting MSA columns. Together $\mathcal{V}$ and $\mathcal{E}$ form an undirected graph encoding probability distributions of amino acids at each MSA position and couplings between MSA positions. The graph must satisfy the Markov property — given the variables in the neighborhood of a vertex, its probability distribuion is conditionally independent of all other variables in the graph.

The Markov property of MRFs allows them to identify direct couplings resulting from physical contacts, rather than correlations arising from non-direct interactions.

φ and ψ are a set of node and edge potentials respectively. φ_i is a vector of size k of potentials assigned to ith column of MSA, and ψ_{ij} is a $k \times k$ matrix of coupling potentials between ith and jth columns of MSA, where k is the size of the alphabet **Equation 3**. For protein MSA $k = 21$ (20 amino acids + gap indexed as the last character) **(Daphne Koller, 2009)**, **(Balakrishnan et al., 2011)**.

$$\varphi_i = \begin{pmatrix} v_1^i \\ v_2^i \\ \vdots \\ v_k^i \end{pmatrix}; \quad \psi_{ij} = \begin{pmatrix} w_{11}^{ij} & w_{12}^{ij} & \dots & w_{1k}^{ij} \\ w_{21}^{ij} & w_{22}^{ij} & \dots & w_{2k}^{ij} \\ \vdots & \vdots & \ddots & \vdots \\ w_{k1}^{ij} & w_{k2}^{ij} & \dots & w_{kk}^{ij} \end{pmatrix} \tag{3.}$$

Process of learning an MRF is achieved by optimizing a loss function (log-likelihood or pseudo log-likelihood) and can constitute tuning both the architecture (the set $\mathcal{E}$ of MSA column connections) **Figure 12** and edge potentials, or just the edge potentials of a model with a predefined architecture, eg. a fully connected model **(Daphne Koller, 2009)**, **(Rosset et al., 2025)**.

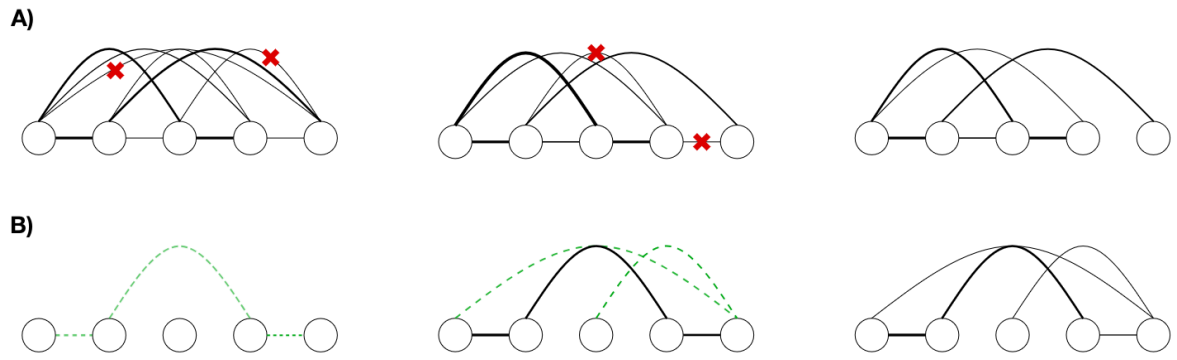


Figure 12: Two examples of training a sparse protein MRF: (**A**) pruning a fully connected model by removing weakest couplings, (**B**) adding connections to a set of vertices. Figure adapted from **(Rosset et al., 2025)**.

GREMLIN (Generative REgularized ModeLs of proteINs) **(Balakrishnan et al., 2011)**, a DCA implementation used in this work following methods outlined in **(Mitrovic et al., 2023)**, learns a fully connected MRF and reports coupling values computed as follows. For each coupling a raw score is the Frobenius norm of the coupling matrix ψ — square root of sum of squares of coupling potentials excluding gaps **Equation 4**.

$$\text{raw}_{ij} = \|\psi_{ij}\|_F = \sqrt{\sum_{a=1}^{k-1}\sum_{b=1}^{k-1}\left(\psi_{ij}\right)_{ab}^2} \tag{4.}$$

The average product corrected score — `apc` — used for ordering couplings, is calculated by subtracting the product of row and column averages divided by total average of `raw` couplings matrix **Equation 5**.

$$\text{apc} = \text{raw} - \frac{\overline{\text{raw}}_{\cdot j}\overline{\text{raw}}_{i\cdot}}{\overline{\text{raw}}} \tag{5.}$$

An apc coupling matrix is then filtered for a desired range of values, eg. top 200 strongest coevolving pairs, to extract contact information.

2. Materials and methods

2.1. Computational resources

Adapted from `fastfetch` output on used machines

Workstation1

```
OS: Ubuntu 24.04.4 LTS x86_64
Host: HP Z2 Tower G4 Workstation
Kernel: Linux 6.8.0-110-generic
CPU: Intel(R) Core(TM) i7-8700 (12) @ 4.60 GHz
GPU: NVIDIA Quadro P1000 [Discrete]
Memory: 31.20 GiB
Swap: 47.68 GiB
```

Fedora laptop

```
OS: Fedora Linux 44 (Workstation Edition) x86_64
Host: 21AJS69V0Y (ThinkPad T14 Gen 3)
Kernel: Linux 7.0.9-205.fc44.x86_64
CPU: 12th Gen Intel(R) Core(TM) i7-1265U (12) @ 4.80 GHz
GPU: Intel Iris Xe Graphics @ 1.25 GHz [Integrated]
Memory: 31.01 GiB
Swap: 22.90 GiB
```

CNN training was carried out on Google Colab with T4 GPU runtime **(Google, n.d.)**.

2.2. Software used through whole project

Python programming language, version 3.14 with libraries:

- NumPy **(Harris et al., 2020)**,
- SciPy **(Virtanen et al., 2020)**,
- pandas **(team, 2020)**, **(McKinney, 2010)**,
- Numba **(Lam et al., 2015)**,

- matplotlib **(Hunter, 2007)**,
- seaborn **(Waskom, 2021)**,
- scikit-learn **(Pedregosa et al., 2011)**,
- tensorflow **(Martín~Abadi et al., 2015)**, **(Developers, 2026)**,
- keras **(Chollet & others, 2015)**.

JupyterLab interactive development environment for data science notebooks **(Thomas et al., 2016)**.

R statistical programming language **(R Core Team, 2026)**, with libraries:
- dplyr **(Wickham et al., 2026)**.

PyMOL Open Source for molecular visualization system **(Schrödinger, LLC, 2015)**.

Jalview for multiple sequence alignment editing, visualisation and analysis **(Waterhouse et al., 2009)**.

2.3. Experimental structure library construction

Experimental structures of ABCG transporters were obtained from databases building on top of Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) **(Berman, 2000)**:

1. PDBsum **(Laskowski et al., 1997)** — by querying *"ABCG"*,
2. Orientation of Proteins in Membranes (OPM) **(Lomize et al., 2011)** — by searching the OPM classification based on Transporter Classification Database (TCDB) **(Saier, 2006)**, **(Saier et al., 2009)**, **(Saier et al., 2015)**, **(Saier et al., 2020)**,
3. Electron Microscopy Data Bank (EMDB) **(Turner et al., 2023)** — by querying *"ABCG"*.

Figure 13 shows example results from OPM and PDBsum. PDB IDs were extracted from html page sources or raw copied text.

ABC transporter G family

- *Type:* Transmembrane (3 classes)
- *Class:* Alpha-helical polytopic (156 superfamilies)
- *Superfamily:* ABC transporters (27 families) 3.A.1 (TCDB)
- *Family:* ABC transporter G family (34 proteins show) 3.A.1.204 (TCDB) PF01061 IPR013525 PDBsum

Proteins (34 entries)

Search table... search

Protein Name	PDB ID	Resolution	Species	Localization	Num. TM Subunits	Num. TM. Sec. Structs.	Hydrophobic Thickness or Depth (Å)	Tilt Angle (°)	Radius of curv. (Å)	$\Delta G_{transfer}$ (kcal/mol)
ABCG1 transporter	7r8c	3.7 EM	*Homo sapiens*	Eykaryo. plasma	2	16	29.8	2	undef.	-96.7
ABCG1 transporter, with ATP	7r8e	3.7 EM	*Homo sapiens*	Eykaryo. plasma	2	14	30.6	0	undef.	-81.9
ABCG1 transporter, with ATP and choleste...	7oz1	4 EM	*Homo sapiens*	Eykaryo. plasma	2	16	30.4	7	undef.	-96.3
ABCG1 transporter, with cholesterol	7fdv	3.26 EM	*Homo sapiens*	Eykaryo. plasma	2	14	29.6	3	undef.	-80.5
ABCG1 transporter, with cholesterol	7r8d	3.2 EM	*Homo sapiens*	Eykaryo. plasma	2	14	29.6	1	undef.	-106.5

PDBsum

Search-string [ABCG]

5do7 - *Crystal structure of the human sterol transporter abcg5/abcg8*

TITLE : CRYSTAL STRUCTURE OF THE HUMAN STEROL TRANSPORTER ABCG5/ABCG8
SOURCE: GENE: ABCG5;
SOURCE: GENE: ABCG8;
KEYWDS: ATP-BINDING CASSETTE TRANSPORTER, ABCG, STEROL EFFLUX,
DATE: 10-SEP-15

5niv - *Crystal structure of 5d3 fab*

KEYWDS: FAB, ABCG2, INHIBITOR, IMMUNE SYSTEM
DATE: 27-MAR-17

5nj3 - *Structure of an abc transporter: complete structure*

SOURCE: GENE: ABCG2, ABCP, BCRP, BCRP1, MXR;
DATE: 28-MAR-17

Figure 13: Screenshots showing OPM (**top**) and PDBsum (**bottom**) results. Original work.

2.3.1. Structures batch download

RCSB PDB offers a script for performing batch downloads for a list of provided structure IDs. I developed an optimized fork of the script with greatly reduced runtime, github: PDB_fast_batch.

2.3.2. Structure conformation labelling

Most structures were not explicitly labelled in database entries. But in almost all cases their conformation was cleary stated in the source paper or its supplementary materials. In case no explicit label was given by authors, a pairwise structural alignment with labeled structures was performed, and conformation label was copied from the best aligned structure.

2.3.3. Descriptive statistics

Barplots showing no. of structures grouped by conformation and protein/species were prepared. Structure resolution information was obtained from structure metadata and visualized as a histogram.

To validate that structures labeled with opposite conformations (IF, OF) are structuraly distinct, I performed in-group structural alignment in PyMOL, and applied Student's t-test on measured distance between H-switch histidines of ABCG NBDs. In case of asymmetric full-size transporters with one NBD modified (eg. Pdr5, Cdr1), I considered the distance betwen H-switch histidine of one NBD and tyrosine at the structuraly equivalent position in the opposite NBD.

2.4. Protein sequence library construction

Protein sequences for DCA contact prediction came from two sources: (1) data set published by one of the colaborators of our group, (2) PSI-BLAST search.

2.4.1. Full-size plant ABCGs

1840 full-size plant ABCG transporter protein sequences found in One Thousand Plant Transcriptomes (1KP) transcript data by **(Pakuła, 2023)**.

2.4.2. Picking sequences for PSI-BLAST search

I prepared a list of protein sequences matching experimental structures from **Section 2.3** which resulted in 8 unique sequences, half of which came from human genes **Table 2**.

	UniProt ID	protein	species
1	Q9H221	ABCG8	human
2	Q9H222	ABCG5	human
3	Q9UNQ0	ABCG2	human
4	P45844	ABCG1	human
5	P33302	PDR5	*S. cerevisiae*
6	Q84TH5	AB25G	*A. thaliana*
7	Q9M2V7	AB16G	*A. taliana*
8	Q5ANA3	CDR1	*C. albicans*

Table 2: Protein sequences of all ABCG transporters with known experimental structures with their UniProt IDs. Original work.

2.4.3. PSI-BLAST

Each sequence from **Table 2** was used as query for PSI-BLAST (Position Specific Iterative BLAST) search against Clustered-NR database. PSI-BLAST is better at finding sequences with conserved motifs, because on each iteration it refines a position specific subsitution matrix (PSSM) based on previous hits **(Altschul, 1997; Altschul et al., 2005)**.

For 4 of them running the 2nd iteration (construction of the first PSSM) resulted in Error: CPU usage limit was exceeded. For the remaining 4 (human ABCG1,ABCG2, ACBG8 and AtABCG16, **Table 2**) we've ran five more iterations, increasing number of target sequences by 100-200 each time the number of new identified sequences was small.

Algorithm parameters used are listed in table **Table 3**.

Parameter	Value
Database	ClusteredNR
Max target sequences	1000
expect threshold	0.05
Word size	3
Max matches in a query range	0
Matrix	Blosum62
Gap cost	Existance: 11 Extension 1
Compositional adjustments	Conditional compositional score matrix adjustment
PSI-BLAST Threshold	0.005
Pseudocount	0

Table 3: Initial parameters used for PSI-BLAST search on NCBI website.

In total 5500 clusters were identified, after removing duplicates — same cluster identified in different runs — 4696 remained **Table 4**.

#	protein	No. cluster representatives
1	ABCG8	1000
3	ABCG2	1500
4	ABCG1	1500
7	AB16G	1500
total	—	5500
unique	—	4696

Table 4: No. of sequences identified by PSI-BLAST. Original work.

2.4.4. Cleaning the sequence library

2.4.4.1. Filtering nonstandard amino acid codes

Some of the plant full-size ABCGs from **Section 2.4.1** contained nonstandard amino acids: ?, *, the "unknown" and "STOP codon". GREMLIN DCA accepts only standard 20 amino acid codes + gap symbol. I trimmed terminal * and removed sequences containing ? at any position, or * at position any other than the C terminus. Procedure yielded **1418** viable sequences.

2.4.4.2. Separating half-size and full-size transporters

Most of the sequence length density in PSI-BLAST sequences was concentrated within the range 650 ± 160, consistent with our knowledge about half-size ABCG sequence length. Any sequences bellow 490 or above 810 were excluded, and remaining sequences were concatenated head-to-tail, to simulate full-size sequences.

Resulting sequences covered size range identical to **(Pakuła, 2023)** plant full-size transporters **Figure 14**.

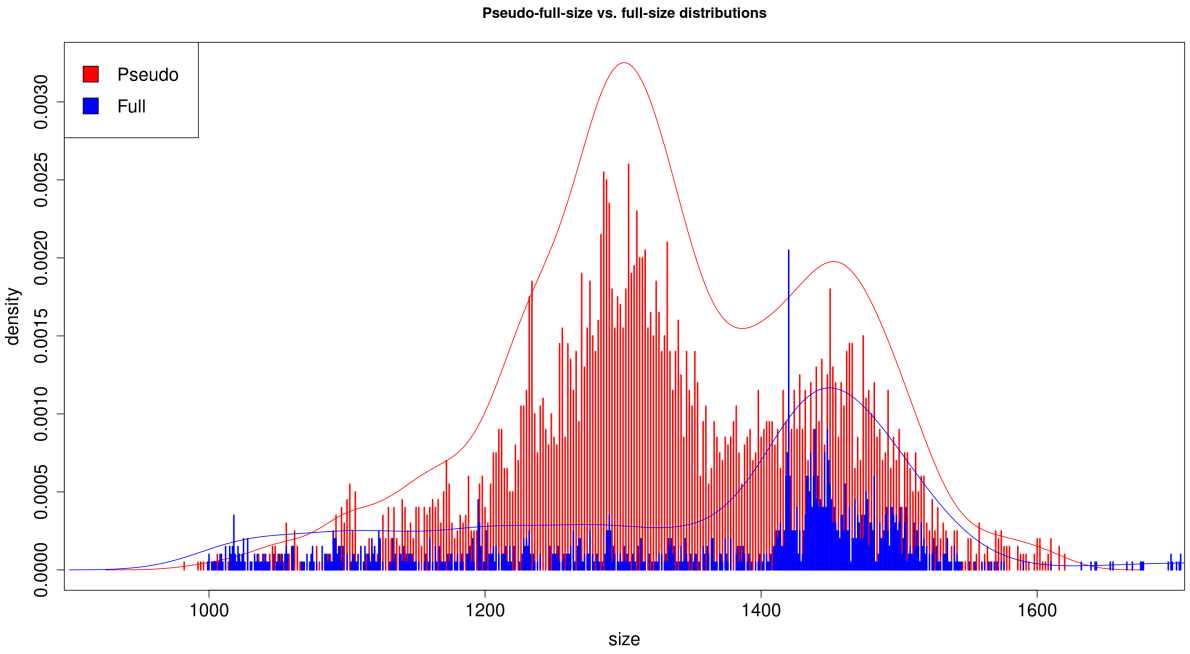


Figure 14: Plot showing kernel density estimated probability density (KDE with gaussian kernel), frequency table barplot was scaled to fit under the curve. Color coding: **red** — concatenated half-size transporters (pseudo-full-size) prepared from PSI-BLAST results; **blue** — plant full-size transporters from **(Pakuła, 2023)**. Original work.

2.4.4.3. PROSITE motif filtering

Sequences found by PSI-BLAST were filtered on PROSITE webserver **(Sigrist et al., 2012)** for presence of the ABC motif (**PS50893**).

2.4.4.4. Structure sequences

Sequences of proteins identified in **Section 2.3**, and sequence of *M. truncatula* ABCG46, were added. Half-size transporters were concatenated head-to-tail, heterodimers were concatenated as homodimers, eg. ABCG5/G8 as ABCG5-ABCG5 and ABCG8-ABCG8.

Described procedure resulted in **5624** viable sequences.

2.5. Multiple Sequence Alignment (MSA)

Processed sequence library was used to construct a multiple sequence alignment with MUSCLE5 using Super5 algorithm **(Edgar, 2022)**. MSA columns with $> 20\%$ gap contet were dropped, following methods outlined in **(Mitrovic et al., 2023)**.

FAMSA2 **(Gudyś et al., 2026)** was used as a newer alternative, using an approximated medoid guide tree, and dropping columns with >80% gaps.

2.6. Experimental structure residue distance and contact maps

A distance map is a matrix $M \in (\mathbb{R}^+)^{n \times n}$, where M_{ij} stores the distance between i-th and j -th residue expressed in Angstrem (Å).

Two distance metrics were used:

1. Euclidean alpha carbon distance (C_α), faster to compute, less affected by lower structure resolution, but insensitive to rotamers and providing only approximate measure of physical residue contacts,
2. minimal interatomic Euclidean distance between residues, slower to compute, less robust for lower resolution structures with wrong rotamers, but able to identify physical contacts withing structural models.

2.6.1. Overcoming time complexity of minimal interatomic distance computation

For a protein of n residues $\binom{n}{2}$ unique pairs need to be evaluated. To find minimal interatomic distances between two residues with k and l heavy atoms respectively, $k \times l$ distances need to be computed, with best case of 25 (Gly-Gly) and worst case of 225 (Trp-Trp).

I utilized bounding sphere based pruning, to limit the costly minimization only to a small number of residue pairs with a chance of contact.

For each residue a bounding sphere is computed with centre at its geometric centre and radius equal to the distance between centre and furthest atom (code: **Listing 1**; visual example: **Figure 15**).

Interatomic distance minimization (code: **Listing 2**) is performed only for residue pairs if distance between surfaces of their bounding spheres is less than 12 Å.

For higher performance both functions were accelerated with LLVM JIT compilation using Numba **(Lam et al., 2015)**.

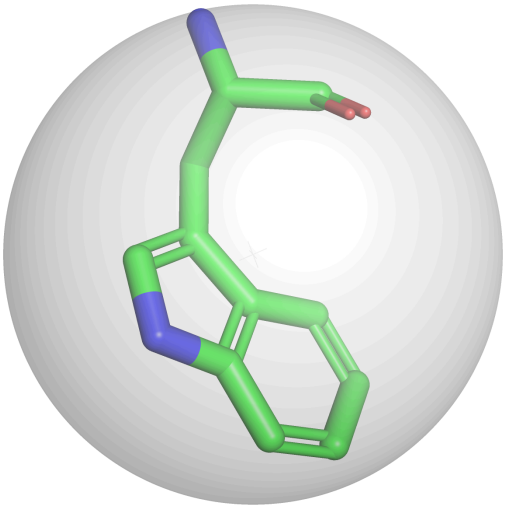


Figure 15: Tryptophan with smallest bounding sphere centered at its geometric centre

```
@jit(nopython=True)
def get_sphere(residue: np.ndarray):
    """ residue is a matrix of xyz coordinates of n atoms
    with shape (n, 3) """
    centroid = np.mean(residue, axis=0)
    radius = np.max(np.linalg.norm(residue - centroid, axis=1))

    return centroid, radius
```

Listing 1: Compute bounding sphere

```
    @jit(nopython=True)
    def _minimize_distance(r1: np.ndarray, r2: np.ndarray):
        """ r1 and r2 are matrices of shape (n, 3) with xyz
        coordinates of residue atoms """
        # initialize to value larger than anything expected
        min_distance = 2048.0

        for i in range(r1.shape[0]):
            for j in range(r2.shape[0]):
                d = np.linalg.norm(r1[i] - r2[j])

                if d < min_distance:
                    min_distance = d

        return min_distance
```

Listing 2: Minimize interatomic distance

2.6.2. Experimental contact map estimation

Experimental contact maps were calculated from distance maps by applying a sigmoid function (**Equation 6**) which introduces a smoothened cutoff at chosen distance. Several values of cutoff were tested:

1. 4 Å following **(Mitrovic et al., 2023)**,
2. 6 Å, as a an alternative to limit the number of pixels with values very close to 0, that could interfiere with CNN training,
3. 8 Å and 12 Å for C_α maps, used in CNN and LinearSVM training.

Formula for sigmoid transformation of distance (d) with specified cutoff (d_0):

$$C(d) = \frac{1}{1 + \exp(d - d_0)} \tag{6.}$$

Applying a sigmoid function transforms distances into open interval $]0, 1[$, where values can be interpreted as strength of contact.

2.7. Direct coupling analysis (DCA)

Following **(Mitrovic et al., 2023)** for DCA I used GREMLIN **(Balakrishnan et al., 2011)**, **(Kamisetty et al., 2013)** with gap cutoff of 20%, LBGFS optimizer and 1000 optimization

iterations. Because web server version of GREMLIN is no longer available, a locally installed Cpp implementation was used: https://github.com/sokrypton/GREMLIN_CPP.

It's important to point out that GREMLIN Cpp doesn't implement early stopping, and will continue until the set number of iterations is met, only saving the output after finishing.

Documentation mentions the possibility of OMP multithreading, but no significant difference in runtime was observed between `OMP_NUM_THREADS=1` and `OMP_NUM_THREADS=8` runs.

DCA was performed on:
1. whole dataset MUSCLE5 MSA,
2. only full-size plant ABCGs from **(Pakuła, 2023)** MUSCLE5 MSA,
3. whole dataset FAMSA2 MSA.

GREMLIN output inlcludes effenctive number of sequences (NEFF). Each sequence is assigned weight of 1 divided by no. sequences with Hamming similarity greater than 80%. NEFF is the sum of individual weights.

2.7.1. Coupling score selection

Initial analysis showed that highest coevolving pairs follow the main diagonal, and for datasets including pseudo-full-size concatenated half-size transporters from PSI-BLAST additional two smaller diagonals representing the main diagonal shifted in phase. I consider those to be artifacts resulting from previously described pseudo-full-size sequence generation, as identical halves will appear to perfectly coevolve along their entire length.

After visual inspection the FAMSA2 DCA most closely resembled experimental contact maps and was chosen for downstream map filtering.

Average product corrected (APC) values were used for ordering contacts. Procedures described by GREMLIN developers suggest to z-score transform APC values and select n residue pairs with highest z-scores.

Distribution of z-scores was roughly normal around 0, with a very long right tail reaching up to z-score of 59. Highest coupling scores exclusively belonged to diagonal artifacts, therefore instead of taking n highest scoring residue pairs I selected those with z-scaled coupling values between 1 and 5.

Additionally to exclude diagonal artifacts I applied filtering with a boolean matrix (code: **Listing 3**).

```
MSA_width = 1146 # Width of FAMSA2 MSA
F = np.ones((MSA_width, MSA_width))

bandwidth = 16 # width of filter lines
diag = 580 # position of phase shifted diagonal

for i, j in itertools.product(range(MSA_width), range(MSA_width)):
    dist = abs(i - j)
    if (dist < bandwidth or abs(dist - diag) < bandwidth):
        F[i,j] = 0
```

Listing 3: Code for diagonal artifact filtering matrix. Original work.

2.8. Machine learning data engineering

2.8.1. Feature extraction

Models deposited in protein structure databases almost never cover the whole sequence, eg. 1st residue in structure might be 13th amino acid in translated sequence. Because of that two different models of the same protein often have differing lengths, and produce noncomparable contact maps. Therefore it's necessary to select equivalent positions before passing contact maps as input for machine learning models.

(Mitrovic et al., 2023) doesn't specify the method of mapping DCA predicted contacts to experimental contact maps, therefore I applied two approaches:

1. based on MSA mapping:
    1. First, from each structure I extracted its sequence as a single chains. We added *M. truncatula* ABCG46 to be used for reference with DCA coupling matrix,
    2. then I computed MSA using FAMSA2 with medoid tree and no column dropping,
    3. finally, from each contact map I select pixels that corresponded to interactions betweem MSA columns with 0 gaps.
2. based on pairwise flexible structural alignment:
    1. first I computed flexible structural alignment between ABCG46 and every other structure with FATCAT 2.0 **(Li et al., 2020)**,
    2. then I identified sequence ranges forming rigid alignment blocks,
    3. finally from each experimental structure I selected positions that aligned with the same sequence ranges in ABCG46.

2.8.2. DCA contact filtering

From each map I selected distance values for residue pairs present in set obtained in **Section 2.7.1**. First sequence of the MSA — *M. truncatula* ABCG46 was used for mapping between DCA coupling matrix positions and aligned contact maps.

2.8.3. DCA-free feature selection

I tested training the LinearSVC model with DCA filtering substituted for variance based feature selection. Selecting aligned contact map positions with variance within the top 10% and at least 10% of values above 0.5.

2.8.4. Preparing labels for classification

Because some conformations were represented by only a few structures or even just a single structure, I decided to simplify the problem into binary classification with broad IF-like conformation and OF-like conformation classes, grouping conformations without accounting for finer details like substrate binding state or ATP binding.

All inward facing (IF) and IF-like conformations were labeled `0` and the rest (OF and OF-like) as `1`.

2.9. Finding state specific contacts with support vector classifiers (SVCs)

I decided to focus our approach on support vector machine linear kernel classifier (LinearSVC) rather than convolutional neural networks (CNNs) for the following reasons:

1. SVMs are known to perform well on problems where no. of features is larger than no. of samples,
2. SVMs readily offer explainability via decision hyperplane coefficients

To correct for class imbalance I trained models with `balanced` class weights to prevent the models from classifying everything as the majority class.

I tested 64 values of C hyperparameter equaly spaced in the log-space at range from −10 to 2 (10^{-10} to 10^{2}) with 32 stratified random shuffle splits per C for validation.

I selected the C value that resulted in mean recall equal to the inflection point of mean recall curve.

To measure coefficient stability I performed Leave-One-Out cross validation of the final model, calculated coefficient mean, standard deviation and normalized maximum difference between coefficient's value.

I tested training the model on **Equation 6** transformed distances (contact maps) before and after Z-score scaling with `StandardScaler()`.

2.10. Training a convolutional neural network (CNN)

Following the methods outlined in **(Mitrovic et al., 2023)** I trained a convolutional neural network (CNN) using Keras **(Chollet & others, 2015)** and Tensorflow **(Martín~Abadi et al., 2015)**, **(Developers, 2026)** on Google Colab T4 GPU runtime.

Models were trained on aligned contact maps without DCA filtering.

Architecture of the final model:

1. 2D convolutional layer, 16 filters with 3×3 kernel and ReLU activation function,
2. 2D max pooling with 3×3 pool size,
3. 2D convolutional layer, 128 filters with 5×5 kernel and ReLU activation function,
4. 2D max pooling with 3×3 pool size,
5. 256 dimensional dense layer with ReLU activation function,
6. single output neuron with sigmoid activation function.

Figure 16 shows a diagram of used CNN architecture (layer sizes not to scale).

I trained the model using binary crossentropy loss, ADAM optimizer and validation loss early stopping with patience of 10 epochs.

I tested using one or two convolutional layers (with max pooling layer after each one) and one or two dense layers, and several hand picked values of learning rate in the range from 10^{-5} to 10^{-3}. I did not carry out a systematic architecture or hyperparameter optimization, since available compute on Google Colab was insufficient.

To mitigate class imbalance I tested training the network on dataset consisting of all minority class samples and random selection of majority class to achieve roughly 1:1 ratio of classes in the training set.

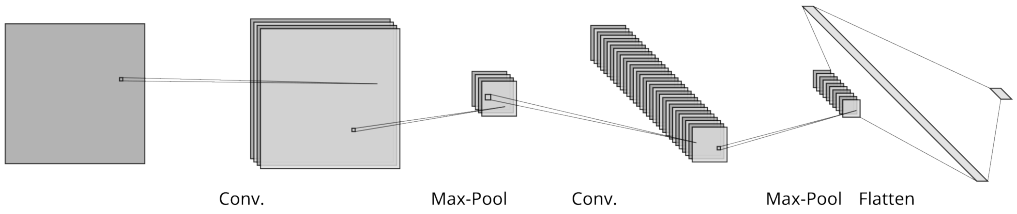


Figure 16: Diagram of final CNN architecture: $2 \times (\text{2D Conv.} \rightarrow \text{Max-Pooling}) \rightarrow \text{flatten} \rightarrow \text{dense} \rightarrow \text{output neuron}$. Original work made with NN-SVG LeNet.

2.11. Layerwise Relevance Propagation of CNN

I did not manage to find a Python library capable of performing layerwise relevance propagation (LRP) that is compatible with current versions of Tensorflow and Keras.

From the existing implementations of LRP, iNNvestigate **(Alber et al., 2019)** is the most promising one, but would require rewriting CNN code in older version of Keras.

(Mitrovic et al., 2023) made their own LRP implementation tailored to sugar porter data, that cannot be applied to ABCG data without refactoring.

3. Results

3.1. Experimental Structures

In total 67 full ABCG structures were found. Full list with citations at **Section 3.1.1**.

Table 5 shows structure accession codes with protein names and conformation as reported by authors.

Figure 17 shows no. of structures grouped by protein and color coded by species, and no. of structures per conformation.

Figure 18 shows distribution of structures' resolution. Surprisingly crystalographic models had one of the lowest resolutions (5DO7: 3.93 and 8CUB: 4.05).

Figure 19 shows histogram of distances between NBD H-switch histidines color coded by broad conformation (IF or OF), and t-test results. Excluding outliers, both groups are clearly identifiable by H-switch distances alone.

Figure 20 structural alignments of structures labeled as IF and OF. Broad structural features specific to each conformation are visible: roughly parallel TMD helices and a visible gap between NBDs for IF, TMD helices brought closer on the intracellular side forming a V-shape, NDBs are close to or dimerized in OF.

Figure 21 shows the orthoscopic view down the TMD channel from NBD side. Color coded by conformation (IF or OF). While the conformations appear distinguishible, there is considerable variance of position withint the groups.

3.1.1. List of structures with citations

Structures obtained with X-ray diffraction are marked with astrisk (*), the rest come from Cryo-EM.

Human ABCG1: 7FDV **(Xu et al., 2022)**, 7OZ1 **(Skarda et al., 2021)**, 7R8C, 7R8D, 7R8E **(Sun et al., 2021)**.

Human ABCG2: 7NEQ, 7NEZ, 7NFD **(Kowal et al., 2021)**, 8P7W, 8P8A, 8P8J **(Irobalieva et al., 2023)**, 8PXO, 8PY4, 8Q7B, 8QCM **(Yu et al., 2024)**, 5NJ3 **(Taylor et al., 2017)**, 6VXF, 6VXH, 6VXI, 6VXJ **(Orlando & Liao, 2020)**, 7OJ8, 7OJH, 7OJI **(Yu et al., 2021)**, 8I38, 8I39, 8I3A, 8I3B, 8I3C, 8I3D **(Huang et al., 2023)**, 6HBU, 6HCO, 6HZM **(Manolaridis et al., 2018)**, 8U2C **(Chen et al., 2023)**, 6ETI, 6FEQ, 6FFC, 6HIJ **(Jackson et al., 2018)**, 8BHT, 8BI0 **(Rasouli et al., 2022)**.

Human ABCG5/G8: 8CUB* **(Farhat et al., 2022)**, 5DO7* **(Lee et al., 2016)**, 7JR7 **(Zhang et al., 2021)**, 7R87, 7R88, 7R89, 7R8A, 7R8B **(Sun et al., 2021)**.

***A. thaliana* ABCG16**: 8WTM, 8WTN, 8WTO, 8WTP **(An et al., 2024)**.

***A. thaliana* ABCG25**: 8WAM, 8WBA, 8WBX, 8WD6 **(Xin et al., 2024)**, 8IWJ, 8IWK, 8IWN, 8K0X, 8K0Z **(Ying et al., 2023)**.

***S. cerevisiae* Pdr5**: 7P03, 7P04, 7P05, 7P06 **(Harris et al., 2021)**.

***C. albicans* Cdr1**: 9IUK, 9IUL, 9IUM, **(Peng et al., 2024)**.

PDB ID	protein	conformation	PDB ID	protein	conformation
7FDV	ABCG1	IF	7OZ1	ABCG1	IF Open
7R8C	ABCG1	IF	7R8D	ABCG1	IF
7R8E	ABCG1	OF	7NEQ	ABCG2	IF
7NEZ	ABCG2	IF	7NFD	ABCG2	IF
8P7W	ABCG2	IF	8P8A	ABCG2	IF
8P8J	ABCG2	IF	8PXO	ABCG2	IF Open
8PY4	ABCG2	IF Open	8Q7B	ABCG2	IF Open
8QCM	ABCG2	IF Open	5NJ3	ABCG2	IF Open
6VXF	ABCG2	Apo Closed	6VXH	ABCG2	IF
6VXI	ABCG2	IF	6VXJ	ABCG2	IF
7OJ8	ABCG2	Turnover 2	7OJH	ABCG2	Turnover 1
7OJI	ABCG2	Turnover 2	8I38	ABCG2	IF
8I39	ABCG2	IF	8I3A	ABCG2	OF
8I3B	ABCG2	OF	8I3C	ABCG2	IF
8I3D	ABCG2	IF	6HBU	ABCG2	OF
6HCO	ABCG2	IF	6HZM	ABCG2	OF
8U2C	ABCG2	IF*	6ETI	ABCG2	IF
6FEQ	ABCG2	IF	6FFC	ABCG2	IF
6HIJ	ABCG2	IF	8BHT	ABCG2	Turnover 1
8BI0	ABCG2	Turnover 2	8CUB	ABCG5/G8	IF
5DO7	ABCG5/G8	IF	7JR7	ABCG5/G8	IF
7R87	ABCG5/G8	IF	7R88	ABCG5/G8	IF
7R89	ABCG5/G8	IF	7R8A	ABCG5/G8	IF
7R8B	ABCG5/G8	IF	8WTM	ABCG16	IF
8WTN	ABCG16	OF Occluded	8WTO	ABCG16	OF Open
8WTP	ABCG16	IF	8WAM	ABCG25	OF
8WBA	ABCG25	IF	8WBX	ABCG25	IF
8WD6	ABCG25	IF	8IWJ	ABCG25	IF
8IWK	ABCG25	IF	8IWN	ABCG25	OF
8K0X	ABCG25	IF	8K0Z	ABCG25	IF
7P03	Pdr5	IF	7P04	Pdr5	IF
7P05	Pdr5	IF	7P06	Pdr5	OF
9IUK	Cdr1	IF	9IUL	Cdr1	IF Open
9IUM	Cdr1	IF			

Table 5: PDB accession codes, protein name and conformation label. 8U2C is marked as 'IF*' because conformation was not specified in the paper and was assigned by structural alignment with IF and OF representatives

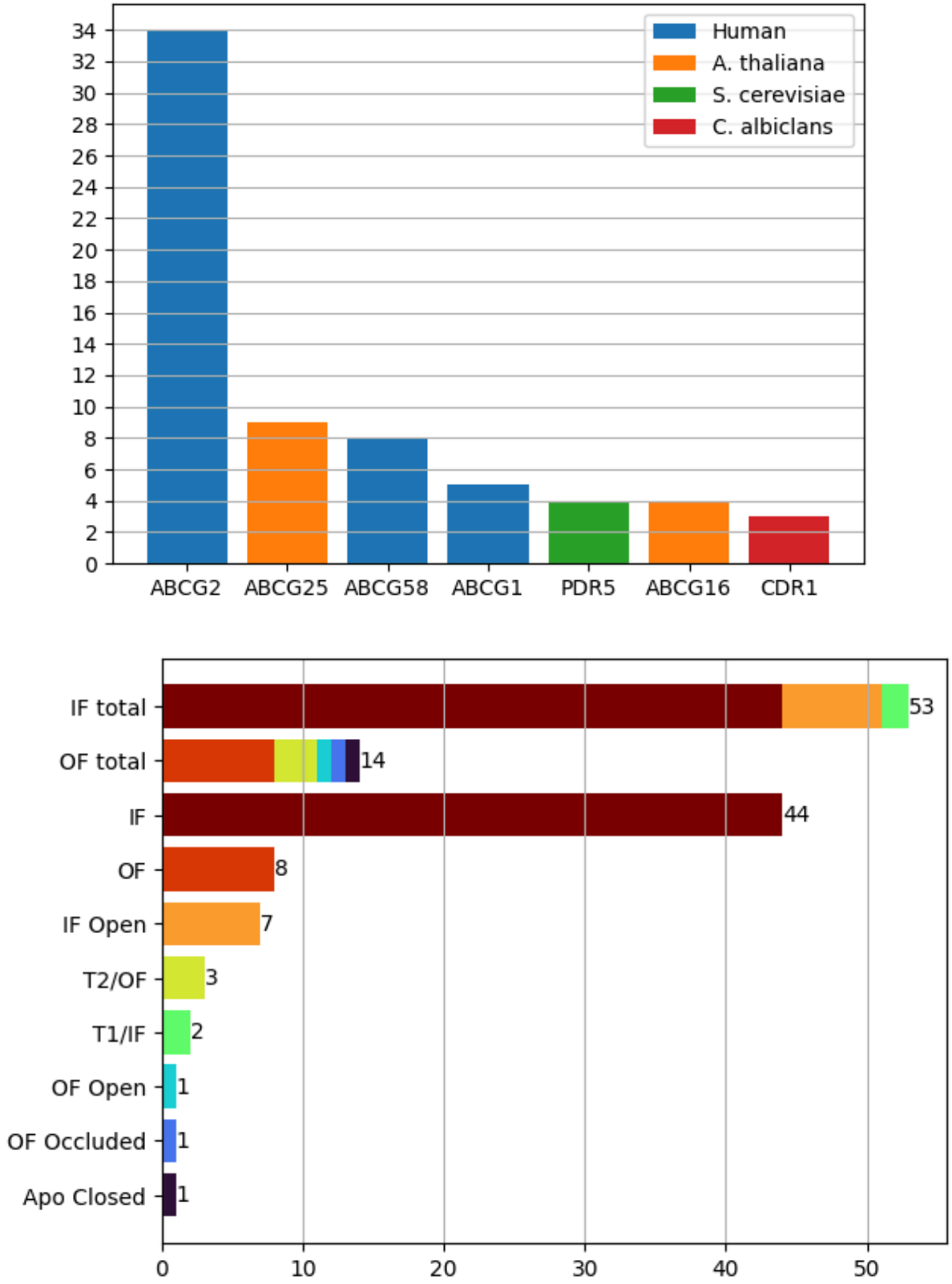


Figure 17: **Top**: total number of structures identified per protein. Color coded by species. **Bottom**: No. of structures with specified conformation labels. IF total and OF total bars show total No. of structures that can be considered as IF or OF conformation. Colors show bar composition. Original work.

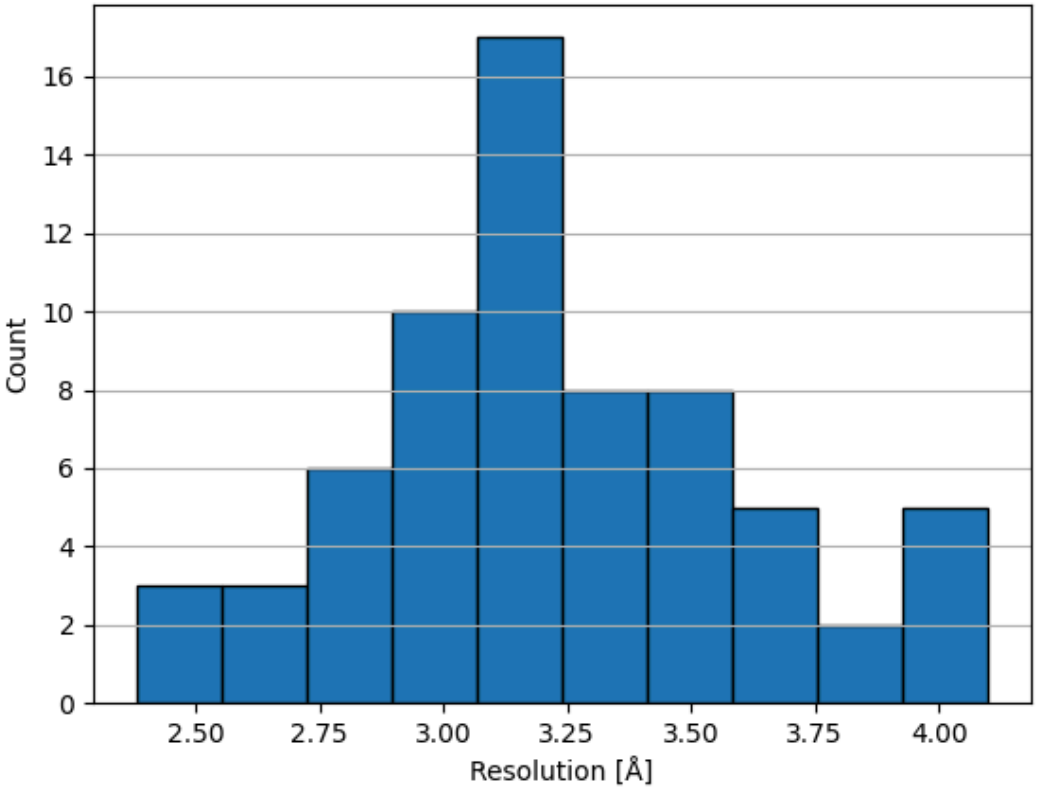


Figure 18: Histogram of structure resolutions as reported in their metadata. Original work.

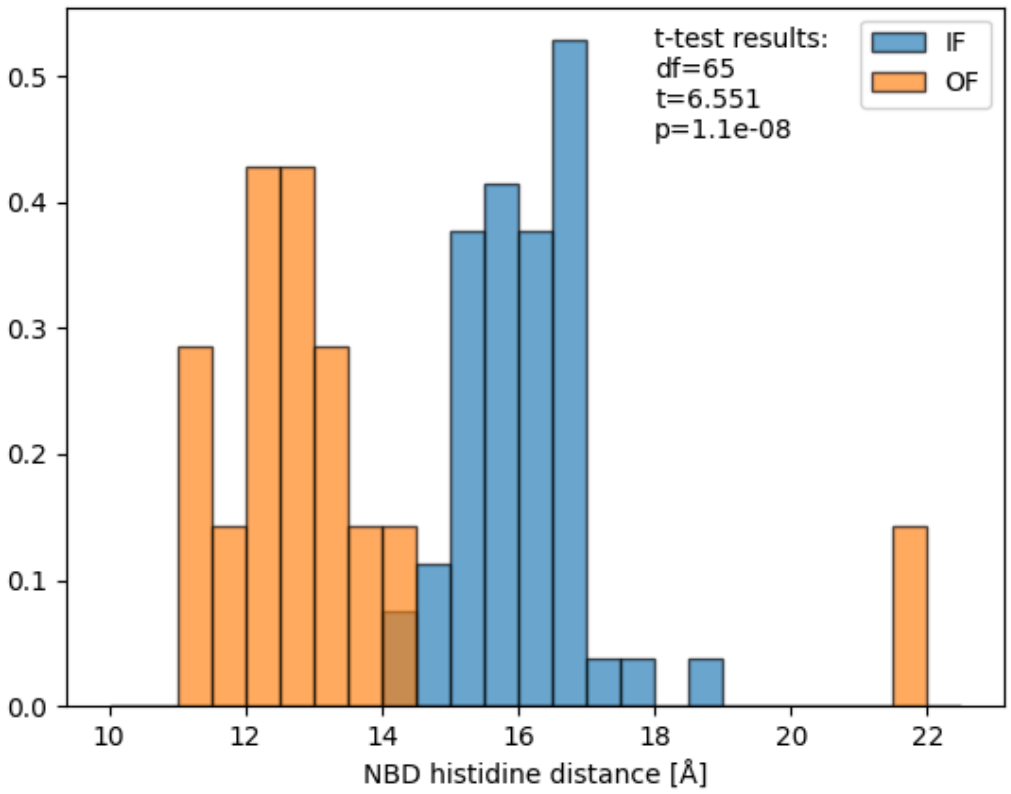


Figure 19: Histogram of H-switch histidine NBD-NBD distance. Excluding outliers, IF and OF conformations are clearly identifiable. Original work.

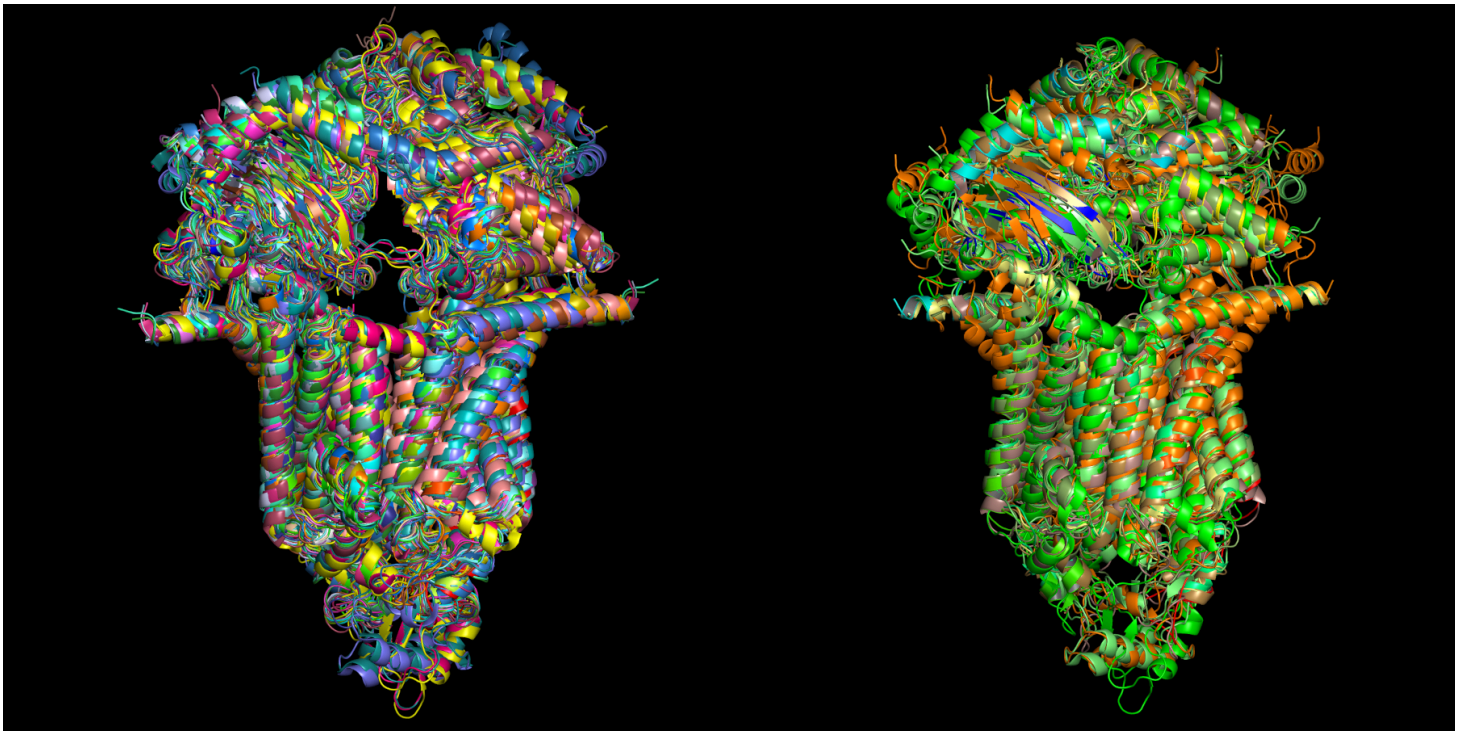


Figure 20: Structural alignment of structures belonging to broad IF (**left**) and broad OF (**right**) conformations. Original work.

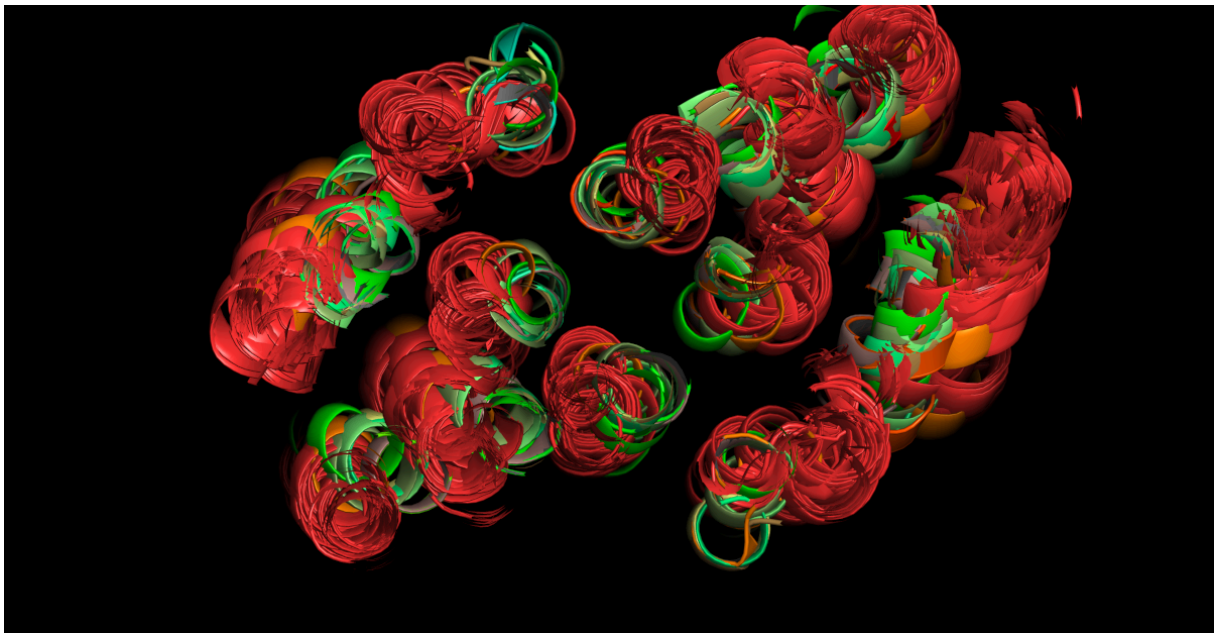


Figure 21: Structural alignment of all structures, orthoscopic view down the TMD channel from NBD side. Color codings: **red**: OF, **green**: IF

3.2. Experimental structure distance and contact maps

Figure 22 shows an example raw distance map and it's corresponding contact map. Visible checkerboard pattern in the raw distance map can be used to identify interactions between domains **Figure 23**.

Figure 24 shows variance of distance at each of the aligned contact map positions for all distance maps. Oval regions of high variance correspond to interactions between NBDs of ABCG halves.

Figure 25 shows difference of means between distance maps grouped by conformation (IF or OF), and per position standard error for difference of means with unequal variances. Calculated difference of means in the oval high variance region of NBD interaction is characterized by large standard error, likely reflecting low resolution of the data.

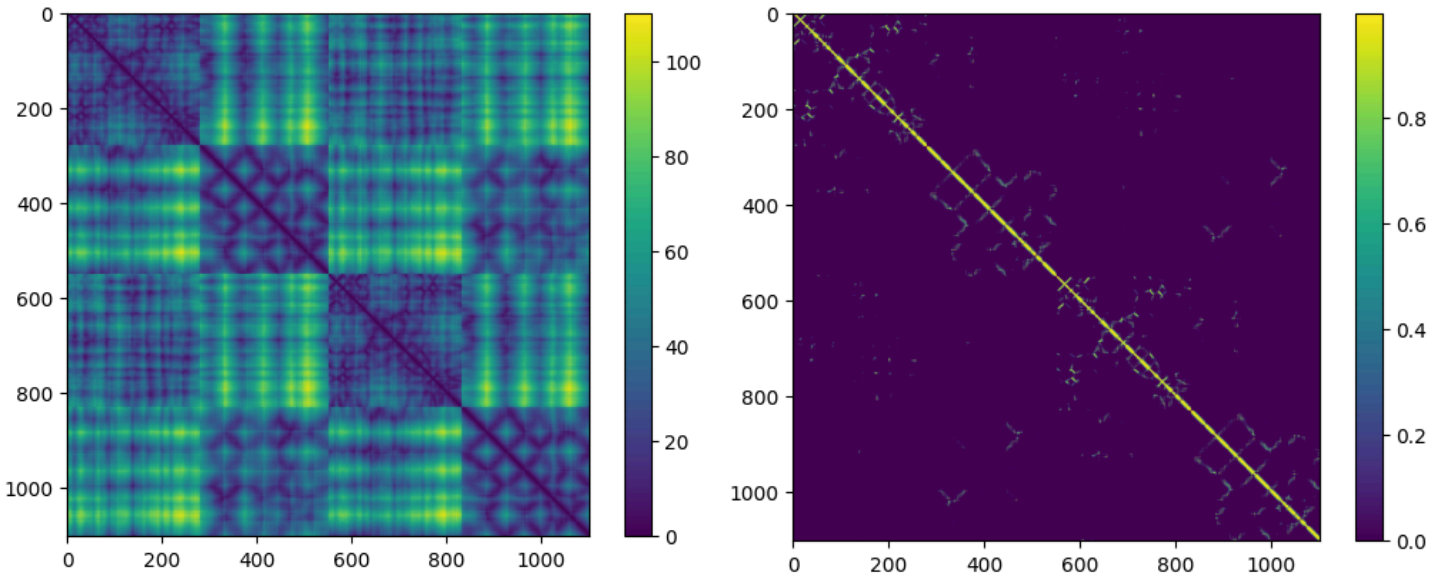


Figure 22: **Left**: full distance map of AtABCG25 (PDB ID: 8WAM). **Right**: same distance map after sigmoid transformation. Original work.

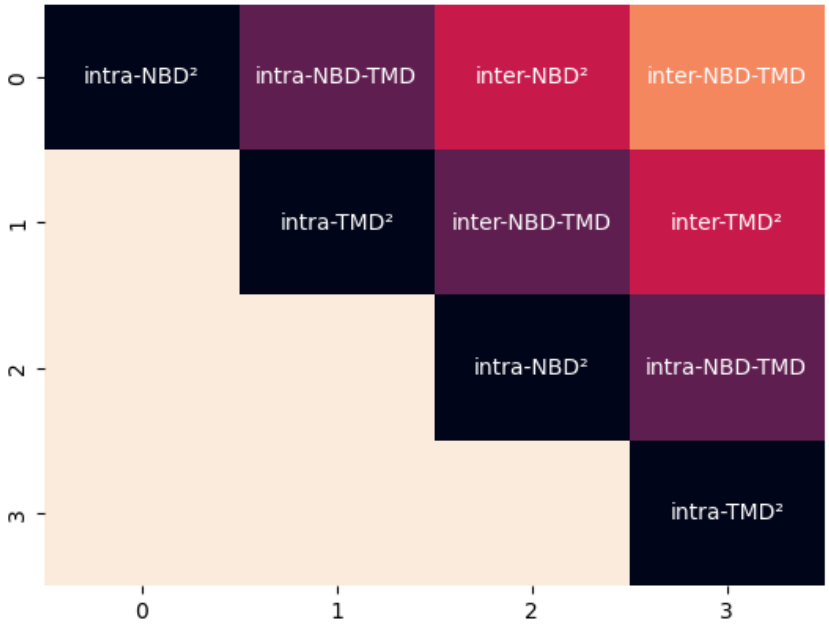


Figure 23: Interpretation of ABCG contact matrices. Because NBD and TMD are roughly the same length, contact matrices can be divided into a 4×4 grid of different NBD and TMD interactions. Interaction encoding scheme: intra — interaction withing the monomer or the same half of full-size transporter, inter — interaction between monomers or halves of full-size transporter, NBD^2 — interaction between NBDs, TMD^2 — interaction between TMDs, NBD-TMD — interaction between NBD and TMD. Lower triangle is identical, hence not shown. Original work.

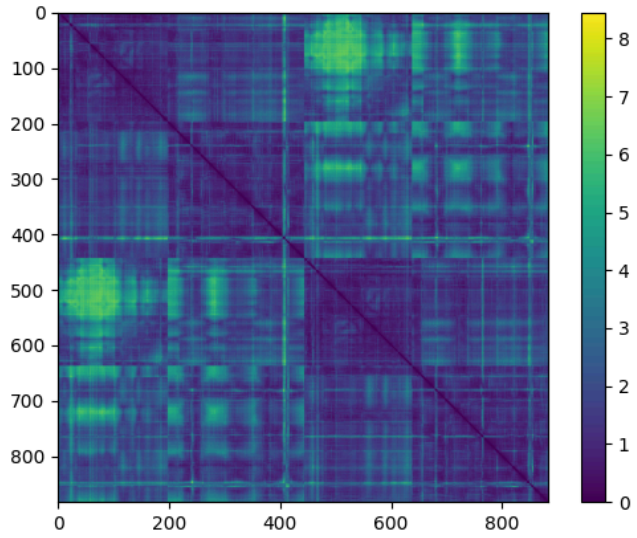


Figure 24: Variance of raw distance values at positions within alignet distance maps. Original work.

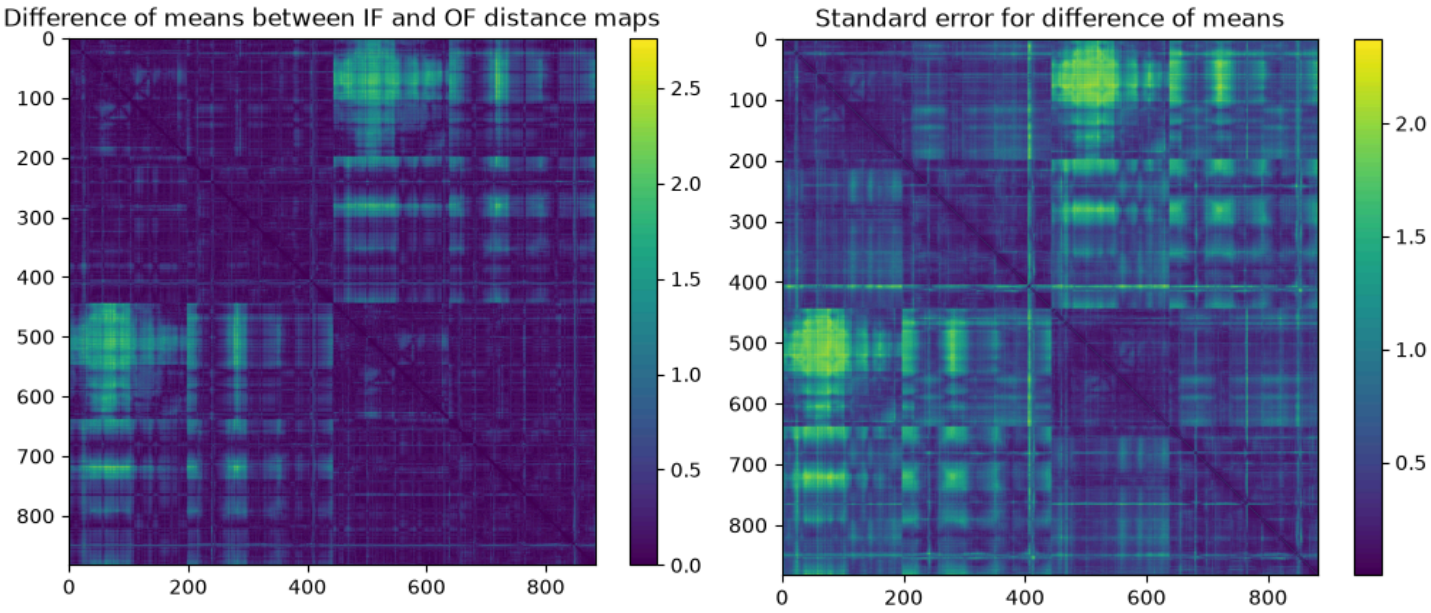


Figure 25: **Left**: per position difference of means between aligned distance maps grouped by conformation (IF or OF). **Right**: standard error for the difference of two means with unequal variances. Original work.

3.3. MSA

FAMSA2 outperformed MUSCLE5 in runtime and quality of produced MSA. Taking less than 1 minute to produce MSA of 5624 sequence of lengths in 1000-1500 range, with smaller no. of columns (20808 vs 34366).

3.4. DCA

MRF built from FAMSA2 MSA used 1146 columns (all with gap content $< 20\%$). Computed effective number of sequences was 1577.7.

Figure 26 shows coupling matrix positions with APC values above set thresholds. Visual inspection shows that threshold of roughly 0.01 is the best compromise between removing noise and capturing non-trivial couplings.

Figure 27 shows couplings within 1 to 5 standard deviations from the mean APC value (APC range from 0.014 to 0.071), and diagonal filter mask.

Figure 28 shows the final filtered coupling matrix of 17072 predicted contacts.

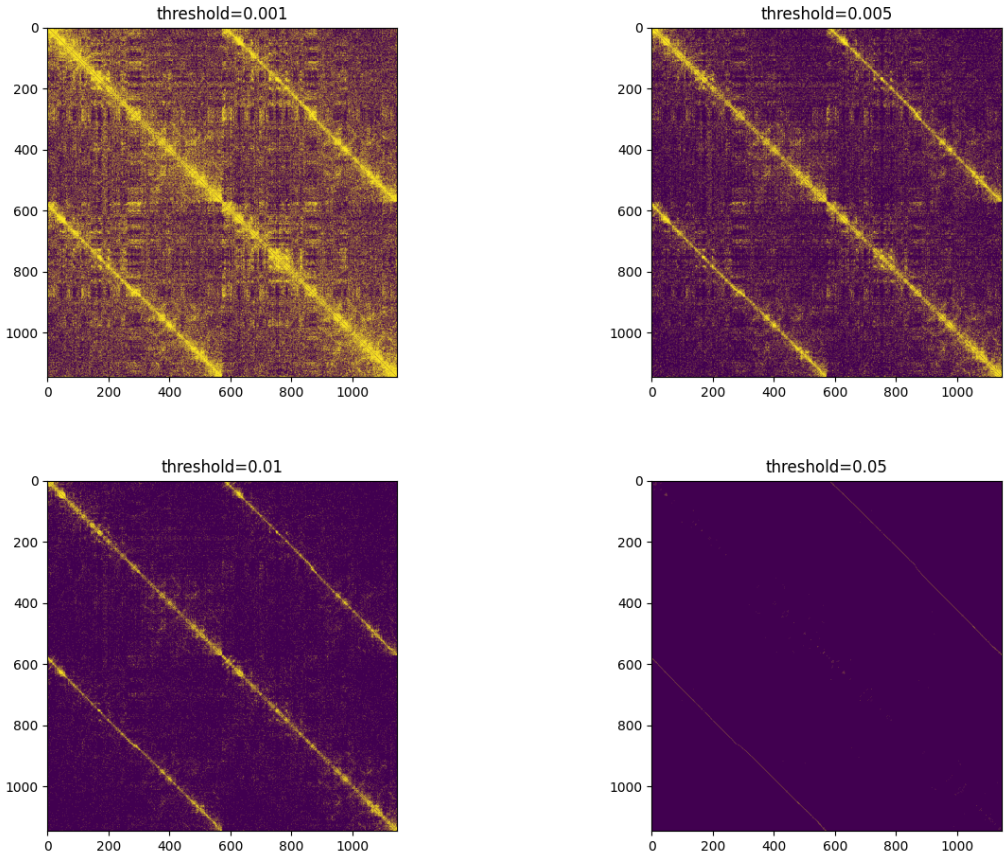


Figure 26: Boolean matrices of couplings with APC values above set threshold. threshold $=$ 0.01 results in the least amount of noise and retains features reminiscent of experimental contact maps. Strong diagonal signal can be filtered before using the matrix for feature selection. Original work.

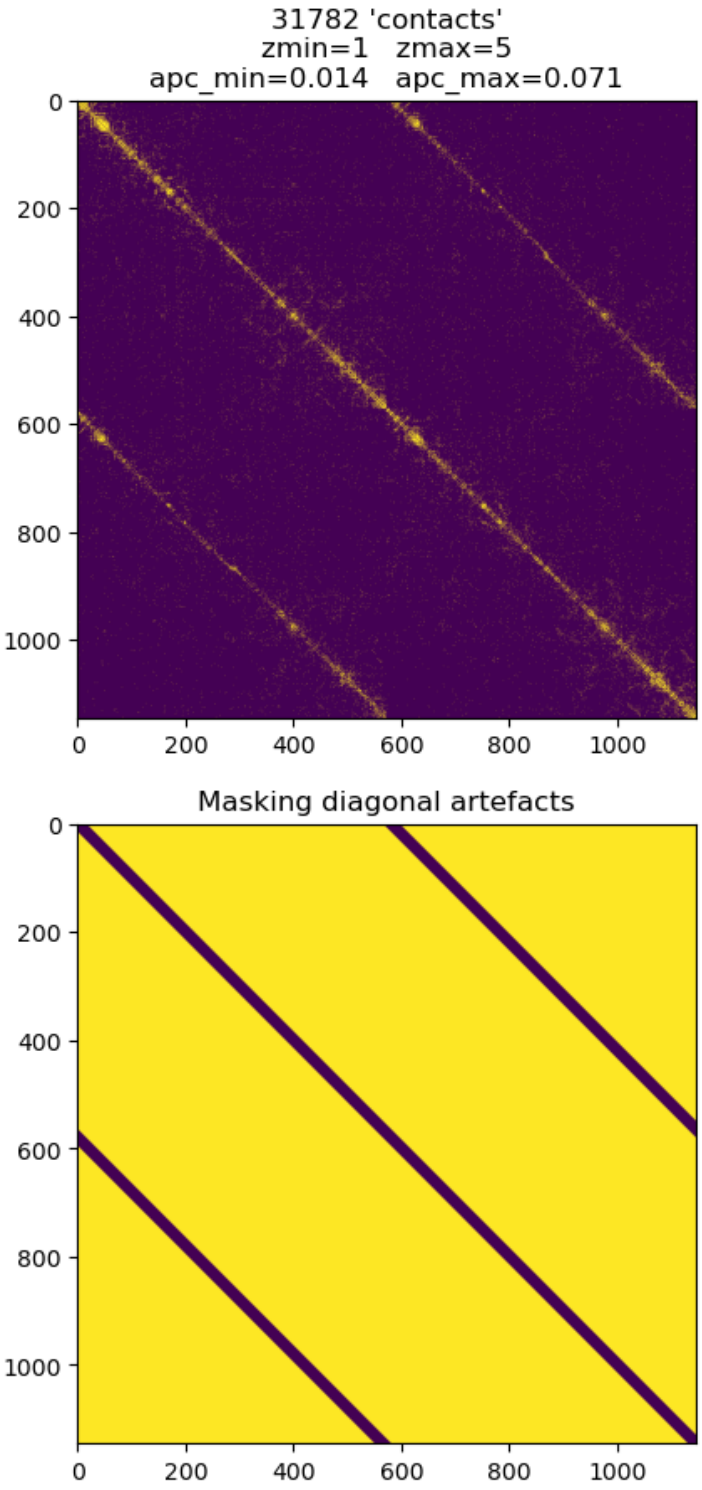


Figure 27: **Top**: boolean matrix of couplings withing the 1 to 5 z-score range (equal to apc range from 0.014 to 0.071). Close to the main diagonal shape similar to experimental contact matrix intra-TMD² region can be seen. **Bottom**: diagonal filters used for eliminating artificial strong signal coming from half-size transporter sequence concatenation. Original work.

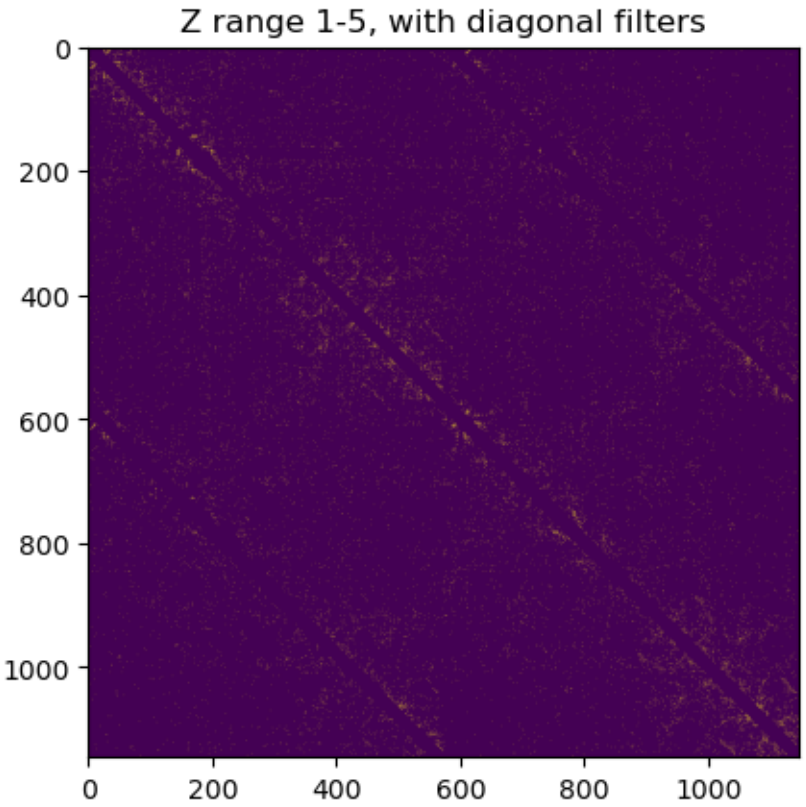


Figure 28: Final coupling matrix with total of 17072 couplings.

3.5. DCA-free feature selection

Figure 29 shows boolean matrix of aligned contact map positions with variance in top 10%. Selecting high variance positions with at least 10% of values above 0.5 yielded **53368** features.

Figure 30 shows 32 randomly selected residue pairs and their C_α distances.

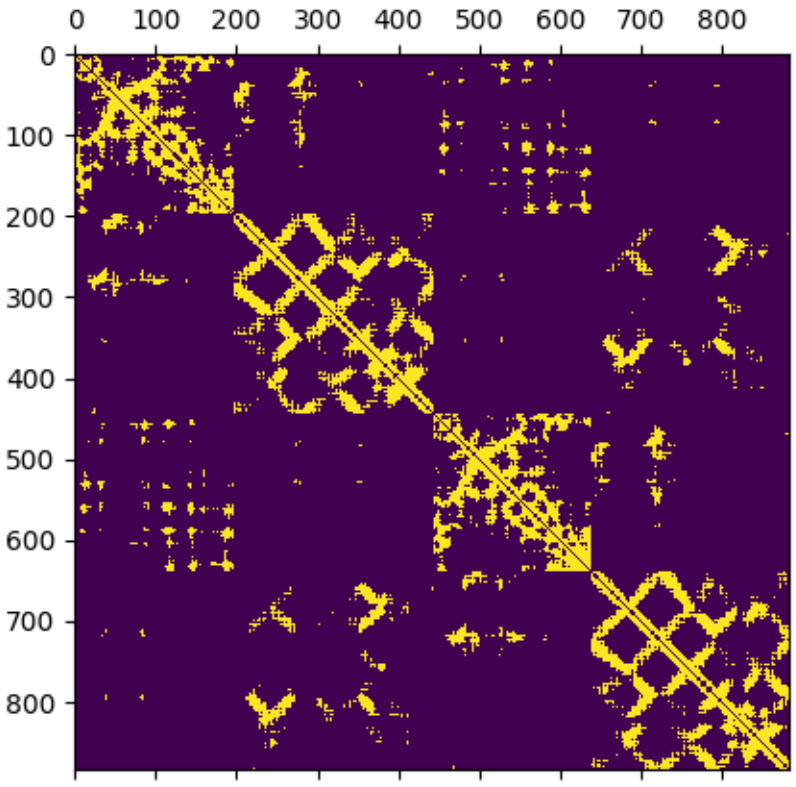


Figure 29: Positions in aligned contact maps with top 10% highest variance. Original work.

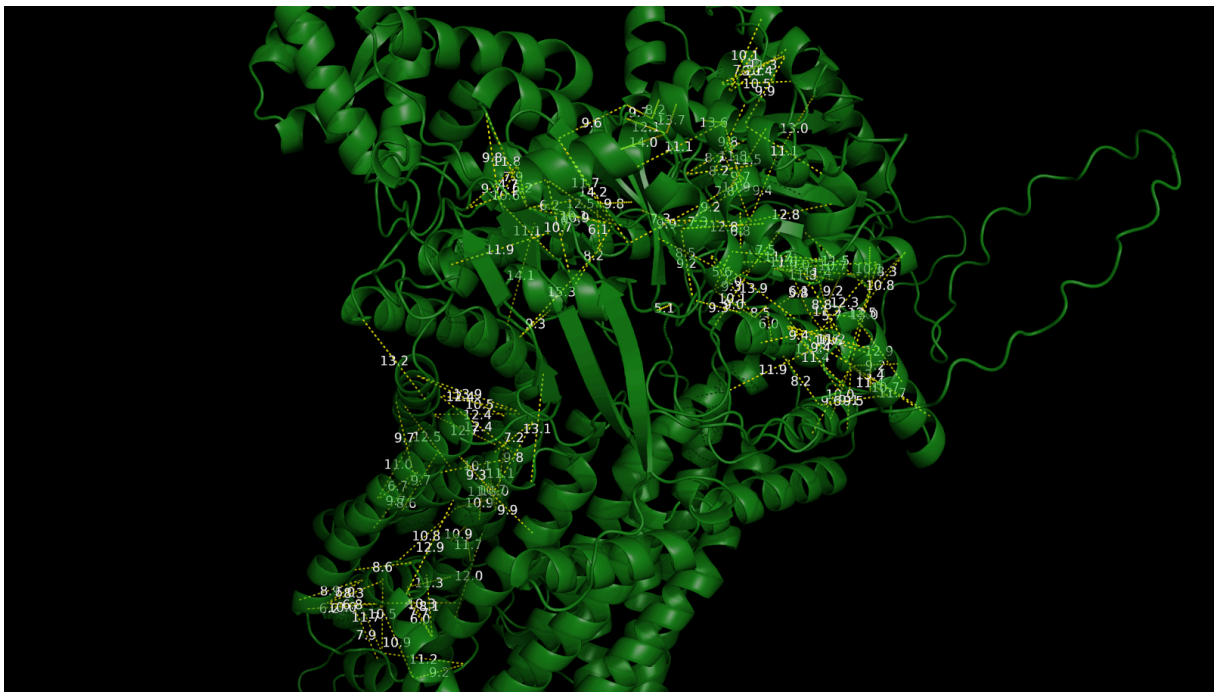


Figure 30: Visualization of 32 randomly selected features mapped to ABCG46 structure. Original work.

3.6. Mapping DCA couplings to ABCG46

Our procedures of mapping DCA predicted contacts to AF predicted ABCG46 structure predominantly selected residue pairs that we never expect to be in contact in ABCG transport cycle.

For this reason we concluded that we don't have a reliable way of using DCA predicted contacts for contact map filtering, and decided to not use it for feature selection for SVC and CNN model input.

3.7. LinearSVC

Figure 31 shows mean recall statistic per tested C hyperparameter values with standard error. Described procedure yielded optimal C value of $1.389 \cdot 10^{-5}$.

Table 6 shows classification metrics for Leave-One-Out (LOO) final model validation trained on contact map data scaled with `StandardScaler()`.

Figure 32 shows precision-recall curve for LOO model validation.

Figure 33 shows mean values of SVC coefficients plotted against their standard error (SE). Correlation between magnitude of coefficient and its SE makes the selection of conformation specific contacts difficult without determining an optimal tradeof between feature importance and uncertainty.

Even the best model was not able to reliably differentiate the classes, and was characterized by low stability of coefficients. In it's present state it can't be used to reliably select state specific residue contacts.

statistic	value
accuracy	0.209
precision	0.048
recall	0.133
f1	0.070

Table 6: Leave-One-Out validation metrics. Original work.

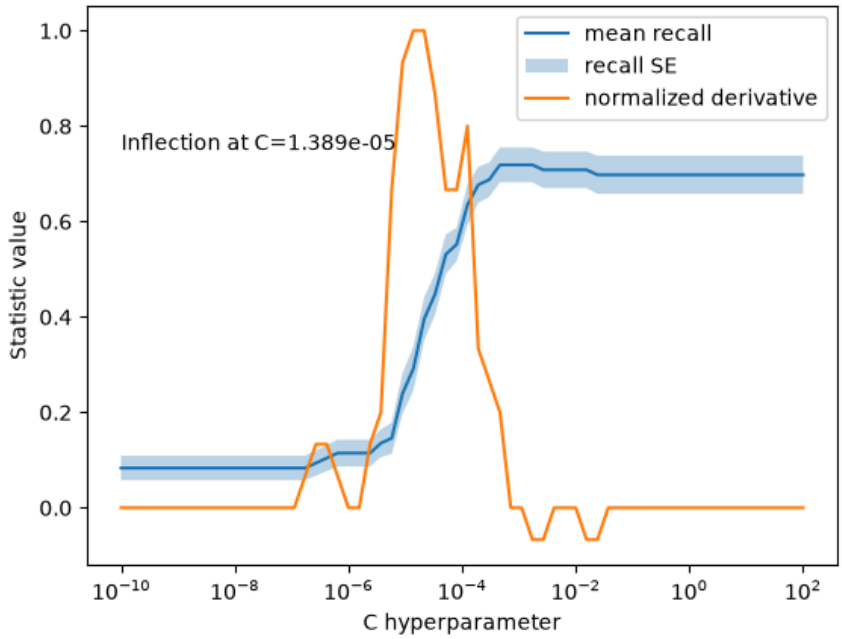


Figure 31: Blue: mean recall statistic calculated for 32 stratified random shuffle split validations per C hyperparameter value. Orange: normalized mean recall derivative with computed mean recall curve inflection point C value. Original work.

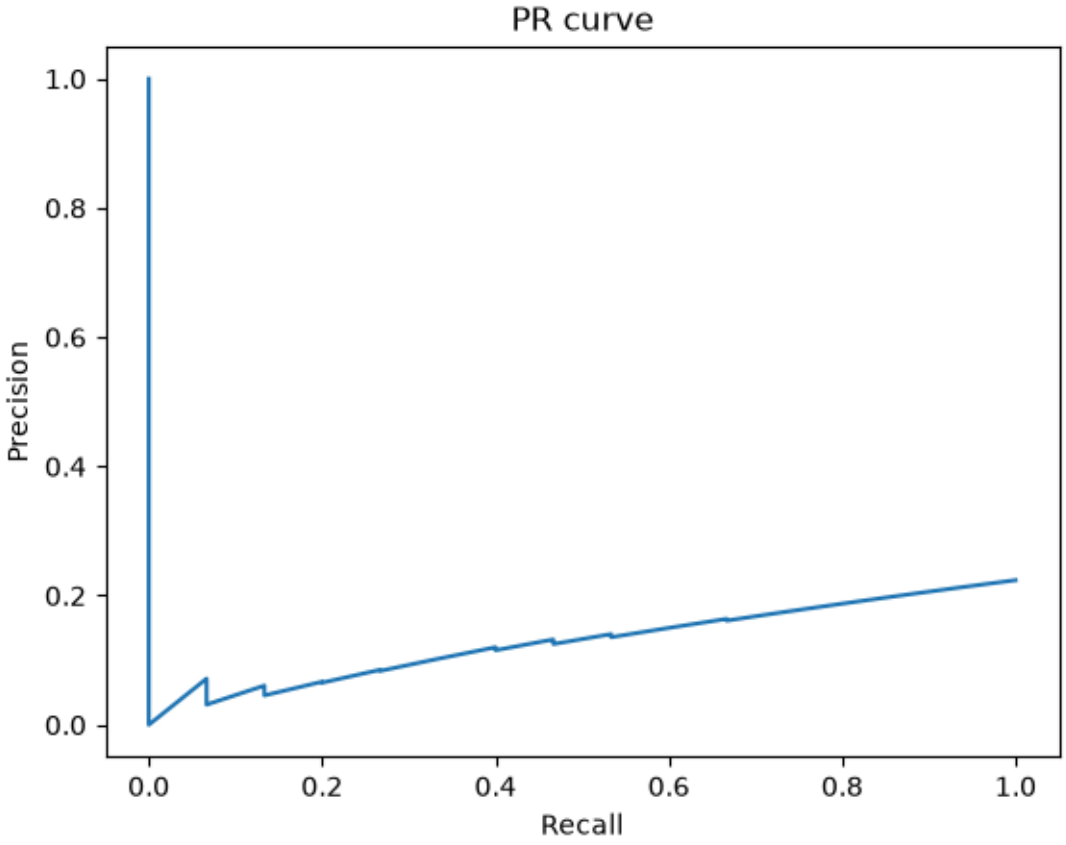


Figure 32: Precision-Recall curve for Leave-One-Out validation of the final SVC model. Original work.

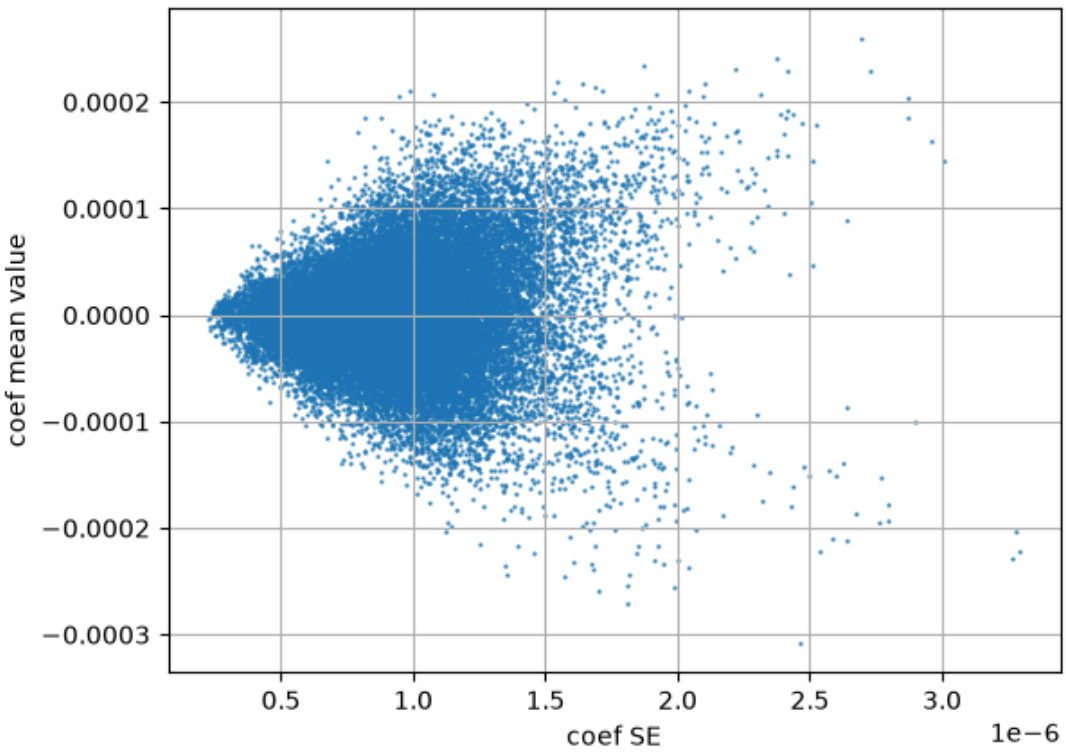


Figure 33: Scatterplot of mean SVC coefficient values in respect to their standard error (SE) over all 67 Leave-One-Out validation training runs. Original work.

3.8. CNN

Across all tested input architecture adjustments, the network consistently converged to predicting the majority class (IF), even when tested on the training set, reflecting the severe class imbalance in the training data (roughly 4:1 ratio of IF:OF; **Figure 17**).

Training on datasets with undersampled majority class did not mitigate this issue.

4. Discussion

4.1. Model performance

All tested models had poor classification performance, and in case of LinearSVC the coefficients intended to provide feature importance were highly unstable, largerly depending on specific training set or even the initial random state of the model.

Very low number of samples coupled with high dimensionality of feature vectors made feature selection and model validation difficult.

Results did not provide a conclusive list of state specific residue contacts, that could be chosen without arbitrary selection.

4.2. Problematic aspects of the data

4.2.1. Experimental structures

Many experimental structures used in this project came from testing drug binding, rather than systematic studies of ABCG transport mechanism. Additionally, the desired conformation

has to be secured or induced, by binding with an antibody, ligand or introducing mutations, possibly introducing heterogeneity in structures labeled with same conformation.

Structure resolution was in the medium to low range, making precise interatomic distances unreliable.

4.2.2. Sequence data

Existance of full-size, homo- and heterodimer forms raises a question of how the input MSA should be prepared. Introducing pseudo-full-size concatenated sequences leads to high artificial couplings along the main and phase-shifted diagonals, which might introduce model bias that cannot be filtered out by removing signals from diagonals. Also, the decision to concatenate half-size sequences with their exact copies is arbitrary. I did not attempt to determine how many half-size sequeces from the dataset belong to heterodimer ABCGs. If heterodimerizing half-size ABCGs were to be paired with their respective biological partners, it's unclear how they should be ordered, eg. as ABCG5-ABCG8 or ABCG8-ABCG5, whether both variants should be included or whether half-size and full-size transporters coupling matrices should be computed individually.

4.3. DCA coupling interpretation

The idea of using DCA for contact map filtering is based on the assumption that strong couplings are interpretable as residue contacts likely to exist in most of the members of a particular protein family. Based on our DCA results, it's clear that in case of our data, both low and high couplings needed to be removed to limit noise and artificial high values. But such procedure is sensitive to biased judgement and selecting for coupling matrices resembling experimental maps, rather than the other way around, defeating the point of DCA based filtering.

4.4. Experimental distance and contact maps

Experimental distance maps show a high degree of variance in regions of interaction between NBDs of transporter halves that is not explained by their conformation. Possibly steming from low structure resolution and differing experimental setups.

4.5. Possible technical improvements

4.5.1. Further contact map computation optimization

Described distance map computation optimization technique relies on computing reasonably small but not minimal enclosing spheres. In larger systems, where finding minimal enclosing spheres could lead to a noticable improvement in runtime, Welzl's algorithm **(Welzl, 1991)** for solving *smallest-circle problem* could be implemented.

4.5.2. CNNs

Two main problems with CNN training were:

1. high class imbalance,
2. low sample size.

The network consistently learned to allways predict the majority class (IF conformation). Weighted sampling with replacement can generate a larger dataset without class imbalance,

but leads to overfitting, in which case LRP pixel importance likely does not reflect their biological importance.

Couple possible solutions could be implemented to mitigate mentioned problems without overfitting:

4.5.2.1. DATA AUGUMENTATION

To prevent overfitting due to exact duplicates in data, sampling with replacement could be paired with transformations that preserve pixel — residue-residue distance relationship. Which includes:

1. shifting (translation),
2. fliping along vertical, horizontal or diagonal axes,
3. rotations by 90°, 180° or 270°,
4. noise injection.

4.5.2.2. CNN PRETRAINING

While the overall ABCG structure is unique, like all proteins it follows a modular architecture with structural motifs present in other families, eg. the NBD. In principle contact maps are able to identify secondary structure, α-helices, β-sheets and their super-secondary structures, eg. greek key motif, but also large conformational changes, eg. NBD dimerization. A CNN learning from scratch only on ABCG data will have to learn both simple and common features and ones specific to ABCGs in different conformations, on a small dataset.

In similar cases in general image recognition problems the approach is to use a model pretrained on large general dataset of a similar kind, and fine-tune with a small problem specific dataset. This allows the model to learn general complex features, and then utilize pretrained filters to learn high level problem specific features on a much smaller dataset without overfitting.

In this case the proposed approach would be to pretrain a CNN on a large protein contact map dataset, to learn general protein structural motifs. Then fine-tune on the small ABCG dataset, to learn ABCG conformation specific features.

5. CONCLUSIONS

The method shows promise for identifying state specific residue contacts in ABCG transporters, but the lack of tested guideliness and best practices makes interpreting and utilizing intermediate results difficult.

I consider a pretrained CNN to be the most promising approach to be tested next. CNNs work well with complex unstructured data composed of composite local features. Provided pretraining on larger diverse protein dataset goes well, the available ABCG experimental structure dataset should be sufficient for fine tuning. For explainability a LRP implementation based on original paper by **(Bach et al., 2015)** would have to be implemented for the current versions of Keras and Tensorlow, which might take considerable time.

Bibliography

Aittoniemi, J., Wet, H. de, Ashcroft, F. M., & Sansom, M. S. P. (2010). Asymmetric Switching in a Homodimeric ABC Transporter: A Simulation Study. *Plos Computational Biology*, *6*(4), e1000762. https://doi.org/10.1371/journal.pcbi.1000762

Alam, A., & Locher, K. P. (2023). Structure and Mechanism of Human ABC Transporters. *Annual Review of Biophysics*, *52*(1), 275–300. https://doi.org/10.1146/annurev-biophys-111622-091232

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., Samek, W., Müller, K.-R., Dähne, S., & Kindermans, P.-J. (2019). iNNvestigate Neural Networks!. *Journal of Machine Learning Research*, *20*(93), 1–8. http://jmlr.org/papers/v20/18-540.html

Almén, M. S., Nordström, K. J., Fredriksson, R., & Schiöth, H. B. (2009). Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biology*, *7*(1). https://doi.org/10.1186/1741-7007-7-50

Altschul, S. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Altschul, S. F., Wootton, J. C., Gertz, E. M., Agarwala, R., Morgulis, A., Schäffer, A. A., & Yu, Y. (2005). Protein database searches using compositionally adjusted substitution matrices. *The FEBS Journal*, *272*(20), 5101–5109. https://doi.org/10.1111/j.1742-4658.2005.04945.x

An, N., Huang, X., Yang, Z., Zhang, M., Ma, M., Yu, F., Jing, L., Du, B., Wang, Y.-F., Zhang, X., & Zhang, P. (2024). Cryo-EM structure and molecular mechanism of the jasmonic acid transporter ABCG16. *Nature Plants*, *10*(12), 2052–2061. https://doi.org/10.1038/s41477-024-01839-0

Arai, N., Furuta, T., & Sakurai, M. (2017). Analysis of an ATP-induced conformational transition of ABC transporter MsbA using a coarse-grained model. *Biophysics and Physicobiology*, *14*(0), 161–171. https://doi.org/10.2142/biophysico.14.0_161

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, *10*(7), e130140. https://doi.org/10.1371/journal.pone.0130140

Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I., & Langmead, C. J. (2011). Learning generative models for protein fold families. *Proteins: Structure, Function, And Bioinformatics*, *79*(4), 1061–1078. https://doi.org/https://doi.org/10.1002/prot.22934

Berman, H. M. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

Chen, M., Schmid, M. F., & Chiu, W. (2023). Improving resolution and resolvability of single-particle cryoEM structures using Gaussian mixture models. *Nature Methods*, *21*(1), 37–40. https://doi.org/10.1038/s41592-023-02082-9

Chollet, F., & others. (2015,). *Keras*.

Daphne Koller, N. F. (2009). *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*.

Developers. (2026,). *TensorFlow*. Zenodo. https://doi.org/10.5281/ZENODO.4724125

Do, T. H. T., Martinoia, E., & Lee, Y. (2018). Functions of ABC transporters in plant growth and development. *Current Opinion in Plant Biology*, *41*, 32–38. https://doi.org/10.1016/j.pbi.2017.08.003

Edgar, R. C. (2022). Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-34630-w

Farhat, D., Rezaei, F., Ristovski, M., Yang, Y., Stancescu, A., Dzimkova, L., Samnani, S., Couture, J.-F., & Lee, J.-Y. (2022). Structural Analysis of Cholesterol Binding and Sterol Selectivity by ABCG5/G8. *Journal of Molecular Biology*, *434*(20), 167795. https://doi.org/10.1016/j.jmb.2022.167795

Google. (n.d.). *Google Colaboratory*. https://colab.research.google.com/

Gudyś, A., Zielezinski, A., Notredame, C., & Deorowicz, S. (2026). Fast and accurate multiple-protein-sequence alignment at scale with FAMSA2. *Nature Biotechnology*. https://doi.org/10.1038/s41587-026-03095-3

Harris, A., Wagner, M., Du, D., Raschka, S., Nentwig, L.-M., Gohlke, H., Smits, S. H. J., Luisi, B. F., & Schmitt, L. (2021). Structure and efflux mechanism of the yeast pleiotropic drug resistance transporter Pdr5. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-25574-8

Harris, C. R., Millman, K. J., Walt, S. J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. van, Brett, M., Haldane, A., Río, J. F. del, Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hasanabady, M. H., & Kalalinia, F. (2016). ABCG2 inhibition as a therapeutic approach for overcoming multidrug resistance in cancer. *Journal of Biosciences*, *41*(2), 313–324. https://doi.org/10.1007/s12038-016-9601-5

Hayashi, T., Chiba, S., Kaneta, Y., Furuta, T., & Sakurai, M. (2014). ATP-Induced Conformational Changes of Nucleotide-Binding Domains in an ABC Transporter. Importance of the Water-Mediated Entropic Force. *The Journal of Physical Chemistry B*, *118*(44), 12612–12620. https://doi.org/10.1021/jp507930e

Huang, X., Zhang, X., An, N., Zhang, M., Ma, M., Yang, Y., Jing, L., Wang, Y., Chen, Z., & Zhang, P. (2023). Cryo-EM structure and molecular mechanism of abscisic acid transporter ABCG25. *Nature Plants*, *9*(10), 1709–1719. https://doi.org/10.1038/s41477-023-01509-7

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Hwang, J.-U., Song, W.-Y., Hong, D., Ko, D., Yamaoka, Y., Jang, S., Yim, S., Lee, E., Khare, D., Kim, K., Palmgren, M., Yoon, H. S., Martinoia, E., & Lee, Y. (2016). Plant ABC Transporters Enable Many Unique Aspects of a Terrestrial Plant's Lifestyle. *Molecular Plant*, *9*(3), 338–355. https://doi.org/10.1016/j.molp.2016.02.003

Irobalieva, R. N., Manolaridis, I., Jackson, S. M., Ni, D., Pardon, E., Stahlberg, H., Steyaert, J., & Locher, K. P. (2023). Structural Basis of the Allosteric Inhibition of Human ABCG2 by Nanobodies. *Journal of Molecular Biology*, *435*(19), 168234. https://doi.org/10.1016/j.jmb.2023.168234

Jackson, S. M., Manolaridis, I., Kowal, J., Zechner, M., Taylor, N. M. I., Bause, M., Bauer, S., Bartholomaeus, R., Bernhardt, G., Koenig, B., Buschauer, A., Stahlberg, H., Altmann, K.-H., & Locher, K. P. (2018). Structural basis of small-molecule inhibition of human multidrug transporter ABCG2. *Nature Structural & Molecular Biology*, *25*(4), 333–340. https://doi.org/10.1038/s41594-018-0049-1

Jones, P. M., & George, A. M. (2023). The Switch and Reciprocating Models for the Function of ABC Multidrug Exporters: Perspectives on Recent Research. *International Journal of Molecular Sciences*, *24*(3), 2624. https://doi.org/10.3390/ijms24032624

Juan, D. de, Pazos, F., & Valencia, A. (2013). Emerging methods in protein co-evolution. *Nature Reviews Genetics*, *14*(4), 249–261. https://doi.org/10.1038/nrg3414

Kamisetty, H., Ovchinnikov, S., & Baker, D. (2013). Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences*, *110*(39), 15674–15679. https://doi.org/10.1073/pnas.1314045110

Kang, J., Yim, S., Choi, H., Kim, A., Lee, K. P., Lopez-Molina, L., Martinoia, E., & Lee, Y. (2015). Abscisic acid transporters cooperate to control seed germination. *Nature Communications*, *6*(1). https://doi.org/10.1038/ncomms9113

Khunweeraphong, N., & Kuchler, K. (2025). The human ABCG2 transporter engages three gates to control multidrug extrusion. *Iscience*, *28*(3), 112125. https://doi.org/10.1016/j.isci.2025.112125

Kowal, J., Ni, D., Jackson, S. M., Manolaridis, I., Stahlberg, H., & Locher, K. P. (2021). Structural Basis of Drug Recognition by the Multidrug Transporter ABCG2. *Journal of Molecular Biology*, *433*(13), 166980. https://doi.org/10.1016/j.jmb.2021.166980

Kuromori, T., & Shinozaki, K. (2010). ABA transport factors found in Arabidopsis ABC transporters. *Plant Signaling & Behavior*, *5*(9), 1124–1126. https://doi.org/10.4161/psb.5.9.12566

Kuromori, T., Miyaji, T., Yabuuchi, H., Shimizu, H., Sugimoto, E., Kamiya, A., Moriyama, Y., & Shinozaki, K. (2010). ABC transporter AtABCG25 is involved in abscisic acid transport and responses. *Proceedings of the National Academy of Sciences*, *107*(5), 2361–2366. https://doi.org/10.1073/pnas.0912516107

Lam, S. K., Pitrou, A., & Seibert, S. (2015). Numba: a LLVM-based Python JIT compiler. *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, 1–6. https://doi.org/10.1145/2833157.2833162

Laskowski, R. A., Hutchinson, E., Michie, A. D., Wallace, A. C., Jones, M. L., & Thornton, J. M. (1997). PDBsum: a web-based database of summaries and analyses of all PDB structures. *Trends in Biochemical Sciences*, *22*(12), 488–490. https://doi.org/10.1016/s0968-0004(97)01140-7

Lee, J.-Y., Kinch, L. N., Borek, D. M., Wang, J., Wang, J., Urbatsch, I. L., Xie, X.-S., Grishin, N. V., Cohen, J. C., Otwinowski, Z., Hobbs, H. H., & Rosenbaum, D. M. (2016). Crystal structure of the human sterol transporter ABCG5/ABCG8. *Nature*, *533*(7604), 561–564. https://doi.org/10.1038/nature17666

Li, Z., Jaroszewski, L., Iyer, M., Sedova, M., & Godzik, A. (2020). FATCAT 2.0: towards a better understanding of the structural diversity of proteins. *Nucleic Acids Research*, *48*(W1), W60–W64. https://doi.org/10.1093/nar/gkaa443

Locher, K. P. (2016). Mechanistic diversity in ATP-binding cassette (ABC) transporters. *Nature Structural & Molecular Biology*, *23*(6), 487–493. https://doi.org/10.1038/nsmb.3216

Lomize, A. L., Todd, S. C., & Pogozheva, I. D. (2021). Spatial arrangement of proteins in planar and curved membranes by PPM 3.0. *Protein Science*, *31*(1), 209–220. https://doi.org/10.1002/pro.4219

Lomize, M. A., Pogozheva, I. D., Joo, H., Mosberg, H. I., & Lomize, A. L. (2011). OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Research*, *40*(D1), D370–D376. https://doi.org/10.1093/nar/gkr703

Manolaridis, I., Jackson, S. M., Taylor, N. M. I., Kowal, J., Stahlberg, H., & Locher, K. P. (2018). Cryo-EM structures of a human ABCG2 mutant trapped in ATP-bound and substrate-bound states. *Nature*, *563*(7731), 426–430. https://doi.org/10.1038/s41586-018-0680-3

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., & Sander, C. (2011). Protein 3D Structure Computed from Evolutionary Sequence Variation. *Plos ONE*, *6*(12), e28766. https://doi.org/10.1371/journal.pone.0028766

Martín~Abadi, Ashish~Agarwal, Paul~Barham, Eugene~Brevdo, Zhifeng~Chen, Craig~Citro, Greg~S.~Corrado, Andy~Davis, Jeffrey~Dean, Matthieu~Devin, Sanjay~Ghemawat, Ian~Goodfellow, Andrew~Harp, Geoffrey~Irving, Michael~Isard, Jia, Y., Rafal~Jozefowicz, Lukasz~Kaiser, Manjunath~Kudlur, … Xiaoqiang~Zheng. (2015,). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. https://www.tensorflow.org/

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference: Proceedings of the 9th Python in Science Conference*. https://doi.org/10.25080/Majora-92bf1922-00a

Mitrovic, D., McComas, S. E., Alleva, C., Bonaccorsi, M., Drew, D., & Delemotte, L. (2023). Reconstructing the transport cycle in the sugar porter superfamily using coevolution-powered machine learning. *Elife*, *12*. https://doi.org/10.7554/elife.84805

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., & Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, *108*(49). https://doi.org/10.1073/pnas.1111471108

Natarajan, K., Xie, Y., Baer, M. R., & Ross, D. D. (2012). Role of breast cancer resistance protein (BCRP/ABCG2) in cancer drug resistance. *Biochemical Pharmacology*, *83*(8), 1084–1103. https://doi.org/10.1016/j.bcp.2012.01.002

Nugent, T., & Jones, D. T. (2012). Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proceedings of the National Academy of Sciences*, *109*(24). https://doi.org/10.1073/pnas.1120036109

Orlando, B. J., & Liao, M. (2020). ABCG2 transports anticancer drugs via a closed-to-open switch. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-16155-2

Pakuła, K. (2023,). *Restriction of access to the central cavity is a major contributor to substrate selectivity in plant ABCG transporters*. https://doi.org/.org/10.1007/s00018-023-04751-6

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peng, Y., Lu, Y., Sun, H., Ma, J., Li, X., Han, X., Fang, Z., Tan, J., Qiu, Y., Qu, T., Yin, M., & Yan, Z. (2024). Cryo-EM structures of Candida albicans Cdr1 reveal azole-substrate recognition and inhibitor blocking mechanisms. *Nature Communications*, *15*(1). https://doi.org/10.1038/s41467-024-52107-w

R Core Team. (2026). *R: A Language and Environment for Statistical Computing*. https://doi.org/10.32614/R.manuals

Rasouli, A., Yu, Q., Dehghani-Ghahnaviyeh, S., Wen, P.-C., Kowal, J., Locher, K. P., & Tajkhorshid, E. (2022). Differential dynamics and direct interaction of bound ligands with lipids in multidrug transporter ABCG2. *Proceedings of the National Academy of Sciences*, *120*(1). https://doi.org/10.1073/pnas.2213437120

Rosset, L., Netti, R., Muntoni, A. P., Weigt, M., & Zamponi, F. (2025). adabmDCA 2.0—A Flexible but Easy-to-Use Package for Direct Coupling Analysis. In *Protein Evolution* (pp. 83–104). Springer US. https://doi.org/10.1007/978-1-0716-4828-5_6

Saier, M. H. (2006). TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Research*, *34*(90001), D181–D186. https://doi.org/10.1093/nar/gkj001

Saier, M. H., Reddy, V. S., Moreno-Hagelsieb, G., Hendargo, K. J., Zhang, Y., Iddamsetty, V., Lam, K. J. K., Tian, N., Russum, S., Wang, J., & Medrano-Soto, A. (2020). The Transporter Classification Database (TCDB): 2021 update. *Nucleic Acids Research*, *49*(D1), D461–D467. https://doi.org/10.1093/nar/gkaa1004

Saier, M. H., Reddy, V. S., Tsu, B. V., Ahmed, M. S., Li, C., & Moreno-Hagelsieb, G. (2015). The Transporter Classification Database (TCDB): recent advances. *Nucleic Acids Research*, *44*(D1), D372–D379. https://doi.org/10.1093/nar/gkv1103

Saier, M. H., Yen, M. R., Noto, K., Tamang, D. G., & Elkan, C. (2009). The Transporter Classification Database: recent advances. *Nucleic Acids Research*, *37*(Database), D274–D278. https://doi.org/10.1093/nar/gkn862

Schneider, E., & Hunke, S. (1998). ATP-binding-cassette (ABC) transport systems: Functional and structural aspects of the ATP-hydrolyzing subunits/domains. *FEMS Microbiology Reviews*, *22*(1), 1–20. https://doi.org/10.1111/j.1574-6976.1998.tb00358.x

Schrödinger, LLC. (2015). *The PyMOL Molecular Graphics System, Version~1.8.*

Sigrist, C. J. A., Castro, E. de, Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., Bougueleret, L., & Xenarios, I. (2012). New and continuing developments at PROSITE. *Nucleic Acids Research*, *41*(D1), D344–D347. https://doi.org/10.1093/nar/gks1067

Skarda, L., Kowal, J., & Locher, K. P. (2021). Structure of the Human Cholesterol Transporter ABCG1. *Journal of Molecular Biology*, *433*(21), 167218. https://doi.org/10.1016/j.jmb.2021.167218

Sun, Y., Wang, J., Long, T., Qi, X., Donnelly, L., Elghobashi-Meinhardt, N., Esparza, L., Cohen, J. C., Xie, X.-S., Hobbs, H. H., & Li, X. (2021). Molecular basis of cholesterol efflux via ABCG subfamily transporters. *Proceedings of the National Academy of Sciences*, *118*(34). https://doi.org/10.1073/pnas.2110483118

Sułkowska, J. I., Morcos, F., Weigt, M., Hwa, T., & Onuchic, J. N. (2012). Genomics-aided structure prediction. *Proceedings of the National Academy of Sciences*, *109*(26), 10340–10345. https://doi.org/10.1073/pnas.1207864109

Taylor, N. M. I., Manolaridis, I., Jackson, S. M., Kowal, J., Stahlberg, H., & Locher, K. P. (2017). Structure of the human multidrug transporter ABCG2. *Nature*, *546*(7659), 504–509. https://doi.org/10.1038/nature22345

team, T. pandas development. (2020, February). *pandas-dev/pandas: Pandas.* Zenodo. https://doi.org/10.5281/zenodo.3509134

Thomas, C., & Tampé, R. (2020). Structural and Mechanistic Principles of ABC Transporters. *Annual Review of Biochemistry*, *89*(1), 605–636. https://doi.org/10.1146/annurev-biochem-011520-105201

Thomas, C., Aller, S. G., Beis, K., Carpenter, E. P., Chang, G., Chen, L., Dassa, E., Dean, M., Duong Van Hoa, F., Ekiert, D., Ford, R., Gaudet, R., Gong, X., Holland, I. B., Huang, Y., Kahne, D. K., Kato, H., Koronakis, V., Koth, C. M., … Tampé, R. (2020). Structural and functional diversity calls for a new classification of ABC transporters. *FEBS Letters*, *594*(23), 3767–3775. https://doi.org/10.1002/1873-3468.13935

Thomas, K., Benjamin, R.-K., Fernando, P., Brian, G., Matthias, B., Jonathan, F., Kyle, K., Jessica, H., Jason, G., Sylvain, C., Paul, I., Damián, A., Safia, A., Carol, W., & Team, J. D. (2016). Jupyter Notebooks - a publishing format for reproducible computational workflows. In

Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press. https://doi.org/10.3233/978-1-61499-649-1-87

Turner, J., Abbott, S., Fonseca, N., Pye, R., Carrijo, L., Duraisamy, A. K., Salih, O., Wang, Z., Kleywegt, G. J., Morris, K. L., Patwardhan, A., Burley, S. K., Crichlow, G., Feng, Z., Flatt, J. W., Ghosh, S., Hudson, B. P., Lawson, C. L., Liang, Y., … Ma, X. (2023). EMDB—the Electron Microscopy Data Bank. *Nucleic Acids Research*, *52*(D1), D456–D465. https://doi.org/10.1093/nar/gkad1019

Vasiliou, V., Vasiliou, K., & Nebert, D. W. (2008). Human ATP-binding cassette (ABC) transporter family. *Human Genomics*, *3*(3), 281. https://doi.org/10.1186/1479-7364-3-3-281

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, *25*(9), 1189–1191. https://doi.org/10.1093/bioinformatics/btp033

Welzl, E. (1991). Smallest enclosing disks (balls and ellipsoids). In *New Results and New Trends in Computer Science* (pp. 359–370). Springer-Verlag. https://doi.org/10.1007/bfb0038202

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2026). *dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org/

Xin, J., Zhou, Y., Qiu, Y., Geng, H., Wang, Y., Song, Y., Liang, J., & Yan, K. (2024). Structural insights into AtABCG25, an angiosperm-specific abscisic acid exporter. *Plant Communications*, *5*(1), 100776. https://doi.org/10.1016/j.xplc.2023.100776

Xu, D., Li, Y., Yang, F., Sun, C.-R., Pan, J., Wang, L., Chen, Z.-P., Fang, S.-C., Yao, X., Hou, W.-T., Zhou, C.-Z., & Chen, Y. (2022). Structure and transport mechanism of the human cholesterol transporter ABCG1. *Cell Reports*, *38*(4), 110298. https://doi.org/10.1016/j.celrep.2022.110298

Ying, W., Liao, L., Wei, H., Gao, Y., Liu, X., & Sun, L. (2023). Structural basis for abscisic acid efflux mediated by ABCG25 in Arabidopsis thaliana. *Nature Plants*, *9*(10), 1697–1708. https://doi.org/10.1038/s41477-023-01510-0

Yu, Q., Dehghani-Ghahnaviyeh, S., Rasouli, A., Sadurni, A., Kowal, J., Bang-Soerensen, R., Wen, P.-C., Tinzl-Zechner, M., Irobalieva, R. N., Ni, D., Stahlberg, H., Altmann, K.-H., Tajkhorshid, E., & Locher, K. P. (2024). Modulation of ABCG2 Transporter Activity by Ko143 Derivatives. *ACS Chemical Biology*, *19*(11), 2304–2313. https://doi.org/10.1021/acschembio.4c00353

Yu, Q., Ni, D., Kowal, J., Manolaridis, I., Jackson, S. M., Stahlberg, H., & Locher, K. P. (2021). Structures of ABCG2 under turnover conditions reveal a key step in the drug transport mechanism. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-24651-2

Zhang, H., Huang, C.-S., Yu, X., Lee, J., Vaish, A., Chen, Q., Zhou, M., Wang, Z., & Min, X. (2021). Cryo-EM structure of ABCG5/G8 in complex with modulating antibodies. *Communications Biology*, *4*(1). https://doi.org/10.1038/s42003-021-02039-8