**Cheminformatics of Antibiotics**

Sarah Cooper

**Introduction**

Antibiotics have been around since the discovery of penicillin in the 1920’s, but in recent years the number of new antibiotics that are approved by the FDA has come to a screeching halt (Figure 1). At the same time these bacteria are developing evolutionary mechanisms to render our treatments ineffective, resulting in a drastic rise of antibiotic resistance (Figure 2). My research aims to study antibiotic membrane permeability through molecular simulations, while this project intends to build on that work by looking at Quantitative Structure Activity Relationships (QSARs) of antibiotics. To do this various physiochemical properties of a number of antibiotics were compared to experimental data in *E. coli and Pseudomonas aeruginosa*. This will allow for elucidation as to which characteristics are most important for membrane permeability to potentially aid in making new, more effective treatments.

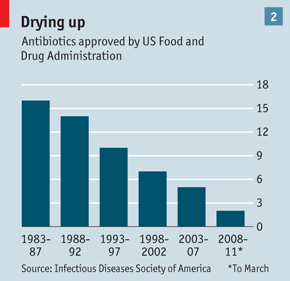


Figure 1: Declining number of antibiotics approved by the FDA. From 2008-2011 only two new drugs were approved to treat bacterial infections.

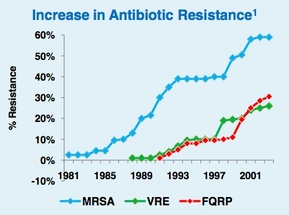


Figure 2: Example of the rise of antibiotic resistance. Note that infections such as MRSA had a resistance rate of almost 60% in 2001 and is likely even higher today.

Gram-negative bacteria have both an outer membrane and an inner membrane. These membranes act like microbial gatekeepers, determining what is allowed into a cell and what is not. Effective antibiotics are able to bypass these gatekeepers either through the membrane itself by diffusion or through channels called porins. Unfortunately, little is known as to which antibiotics prefer which route, but there is evidence behind both mechanisms. If it is understood which ones permeate via each mechanism we can develop novel approaches to designing new drugs.

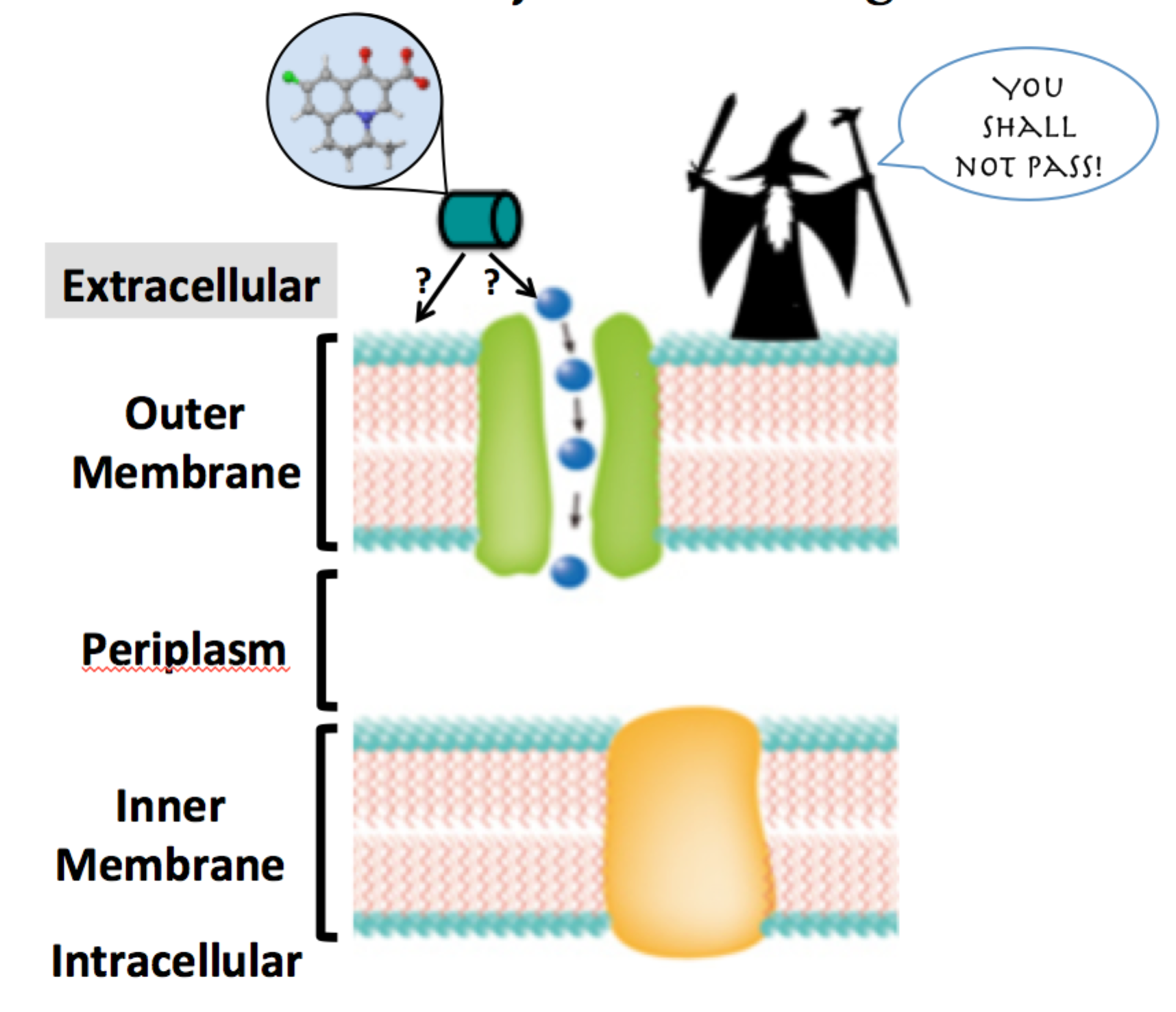


Figure 3: Gram-negative bacteria have “microbial gatekeepers” that control what can and cannot enter the cell. For an antibiotic to be effective it must pass through both the inner and outer membranes and avoid being ejected out by an efflux pump.

**Hypothesis**

A membrane views antibiotics not as a set of atoms connected by bonds as we would draw it, but rather as an ensemble of physiochemical characteristics governed by structure-function relationships. In determining these structure-function relationships we will gain insight into the chemistry behind antibiotic membrane permeability.

**Methods**

Grey indicates steps that were done before this project but are included for completeness of the procedure for the lab notebook portion for reproducibility. These steps are also included because I created and adapted several scripts to automate this workflow. The majority of the commands used can be seen in the scripts folder on github since all of this work was automated.

**Prepare antibiotic structures for molecular dynamics (MD) simulations**

1. Got coordinate files from ZINC database (<http://zinc.docking.org/>) for structures that have MIC data
2. Use openbabel to convert mol2 files to xyz files for optimization
   1. babel – i mol2 \*.mol2 –o xyz \*.xyz
3. Copy coordinate files to Newton
4. Optimize geometries in Gaussian09 (MOL = antibiotic name)
   1. MOL.b3lyp.631gdp.opt.com with coordinates and parameters
   2. MOL.b3lyp.631gdp.opt.sge to submit to Newton
5. Frequency calculations in Gaussian 09 to check for imaginary frequencies
   1. Use optimized geometry from 4.
   2. MOL.b3lyp.631gdp.freq.com with optimized coordinates and parameters
   3. MOL.b3lyp.631gdp.freq.sge to submit to Newton
6. If there are negative frequencies rerun optimization with a lower basis set, then optimize that final structure with 631gdp and run frequency calculations again
7. If there are no negative frequencies calculate charges for atoms in the molecule in Gaussian09

**Use optimized geometries and charges to prepare for MD simulations**

1. On cmbcluster run:

antechamber -i MOL.log -fi gout -o MOL.prep -fo prepi -c resp

parmchk -i MOL.CHG.prep -f prepi -o MOL.frcmod

To take output file from charge calculations to generate prep file and parameter modification file (frcmod)

1. Run pTranslate-G09Output-XYZ.py to convert Gaussian09 output to xyz file format
   1. Then convert to PDB format

babel – i xyz \*.xyz–o pdb \*.pdb

**Run 1 µs MD simulation on each antibiotic**

1. Minimize with restraints on heavy atoms
2. Minimize without restraints
3. Heat the system to 400 K for 2 ns
4. Cool the system to 310 K for 10 ns
5. Run production MD in ten 100 ns runs

**Scripts**

*1\_qm\_opt.py*

This script is used to generate Gaussian09 input files from a list of antibiotics. This file only requires the coordinates to be added and then it is ready for submission to Gaussian on Newton. Gaussian input files only require the header information that contains all of the instructions on how to run the simulation, a filename, charge and multiplicity of the molecule and the coordinates. To check for negative frequencies use the coordinates from the optimization output file and replace the word opt in the header lines with freq.

*2\_prep\_md.py*

This script uses python to take a template file and iterating through a list in a separate file makes the necessary directories for the MD simulations and adds the appropriate files to run the simulations. This code was adapted from an earlier script to set up other MD simulations.

*3\_run\_plumed.py*

This script uses python to copy over the necessary template files to a certain directory similar to (2) and run the program PLUMED. Note this cannot be run on Newton it was run on the clusters at ORNL where the program is installed. This code was adapted from an earlier script to set up other MD simulations.

*download\_pdb.py*

This script allows you to download PDB files to the current directory and requires only the PDB ID’s in the command line when executing the script.

*g09\_freq.py*

This script is used to check for negative frequencies in geometry optimizations. Negative frequencies arise from structures that while considered optimized do not represent a real molecule.

*rename.sh*

This short bash script is used to batch rename files.

*replace.sh*

This short bash script is used to batch replace a string in files.

**PLUMED descriptors**

Within the software used to run the MD simulations (AMBER14) there is a plugin called PLUMED that will use the trajectories from these simulations and compute various molecular descriptors. These descriptors were then correlated with minimum inhibitory concentration (MIC) data generated from our collaborators. In our case MIC data is the lowest concentration of an antibiotic needed to prevent growth. These experiments were run on both *E. Coli* and *Pseudomonas aeruginosa* (PAE) as different significant descriptor-MIC correlations are expected. We saw significant positive correlations (not shown) with kappa2 (Eq. 1.1 and 1.2) and MIC concentration in *E. coli* with an outer membrane pore, which lead to the interest in this work and expansion of the number of descriptors.

Efflux pumps are one way in which bacteria are able to resist our treatments by ejecting our drugs out of the cell once they have entered. Here we’ve knocked out the functionality of these efflux pumps as well as generated a pore in the cell membrane to see the effects the pore has on antibiotic effectiveness. Below are the different experimental conditions for the MIC data collection for bot *E. coli* and PAE. *E. coli* only has a single efflux pump knockout while PAE which has many more efflux pumps has variations of 3 and 6 pump knockouts.

***E. Coli***

1. Wild type inner membrane (**WT MIC**)
2. Wild type inner membrane with a pore (**WT PORE MIC**)
3. TOLC dependent transporter knock out inner membrane (**ΔTOLC MIC**)
4. TOLC dependent transporter knock out with a pore inner membrane (**ΔTOLC PORE MIC**)
5. Wild type inner membrane with outer membrane pore (**WT/WT PORE OM**)
6. TOLC dependent transporter and TOLC efflux pump knock out with outer membrane pore (**ΔTOLC/ ΔTOLC PORE OM**)
7. Wild type inner membrane with TOLC efflux pump outer membrane knock out (**WT/ΔTOLC PORE EFFLUX**)
8. Wild type inner membrane with a pore with TOLC efflux pump outer membrane knock out (**WT PORE/ ΔTOLC PORE EFFLUX**)
9. Wild type inner membrane with a pore with outer membrane barrier (**WT PORE/ ΔTOLC PORE EFFLUX**)

**PAE**

1. Wild type inner membrane (**PAO1 MIC**)
2. Wild type inner membrane with pore (**PAO1 UP MIC**)
3. 3 efflux pump inner membrane knockout (**∆3 MIC**)
4. 3 efflux pump inner membrane knockout with pore (**∆3 up MIC**)
5. 6 efflux pump inner membrane knockout (**∆6 MIC**)
6. 6 efflux pump inner membrane knockout with pore (**∆6 up MIC**)
7. Wild type with outer membrane pore (**PAO1/PAO1 UP OM**)
8. 3 efflux pump inner membrane and 3 efflux pump outer membrane knockout with outer membrane pore (**∆3/∆3 UP OM**)
9. 6 efflux pump inner membrane knockout with 6 efflux pump outer membrane knockout and outer membrane pore (**∆6/∆6 UP OM**)
10. Wild type inner membrane with 3 efflux pump knockout outer membrane (**PAO1/∆3 EFFLUX**)
11. 3 efflux pump inner membrane and 6 efflux pump outer membrane knockout (**∆3/∆6 EFFLUX)**
12. Wild type inner membrane, 6 efflux pump knockout outer membrane (**PAO1/∆6 EFFLUX**)
13. Wild type inner membrane with pore, 3 efflux pump knockout outer membrane with pore (**PAO1 UP/∆3 UP EFFLUX**)
14. Wild type inner membrane with pore, 6 efflux pump knockout outer membrane with pore **(PAO1 UP/∆6 UP EFFLUX)**
15. Wild type inner membrane with 6 efflux pump outer membrane with pore and barrier **(PAO1/∆6 UP BARRIER)**

**MOE Descriptors**

MOE (version 2015.10) has the ability to take coordinate files (xyz, pdb, mol2, etc.) and calculate various molecular descriptors to expand upon the original descriptors from PLUMED. Our goal was to correlate the *E. Coli* and PAE experimental data with these descriptors to calculate structure activity relationships and determine which characteristics affect membrane permeability. Biological activity is thought to be correlated with properties that arise as a result of chemical structure and physiochemical characteristics of molecules. These descriptors include properties such as shape, surface area, volume, density, hydrophillicity, lipophillicity, etc. Approximately 280 different descriptors were used on 62 antibiotics used on *E. coli* and 44 antibiotics used in PAE.

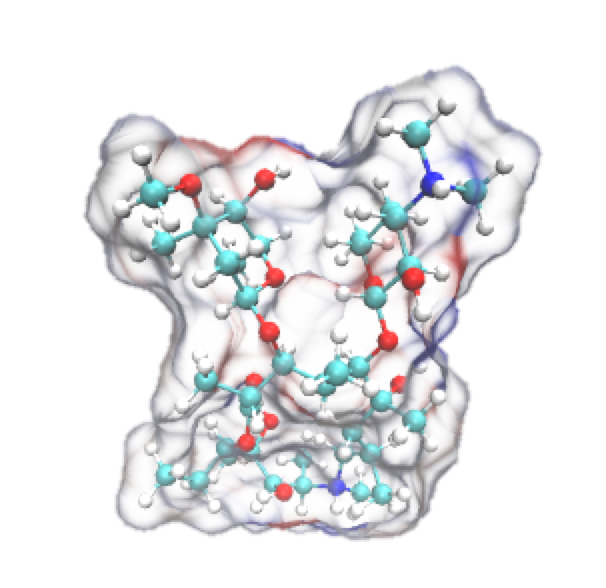
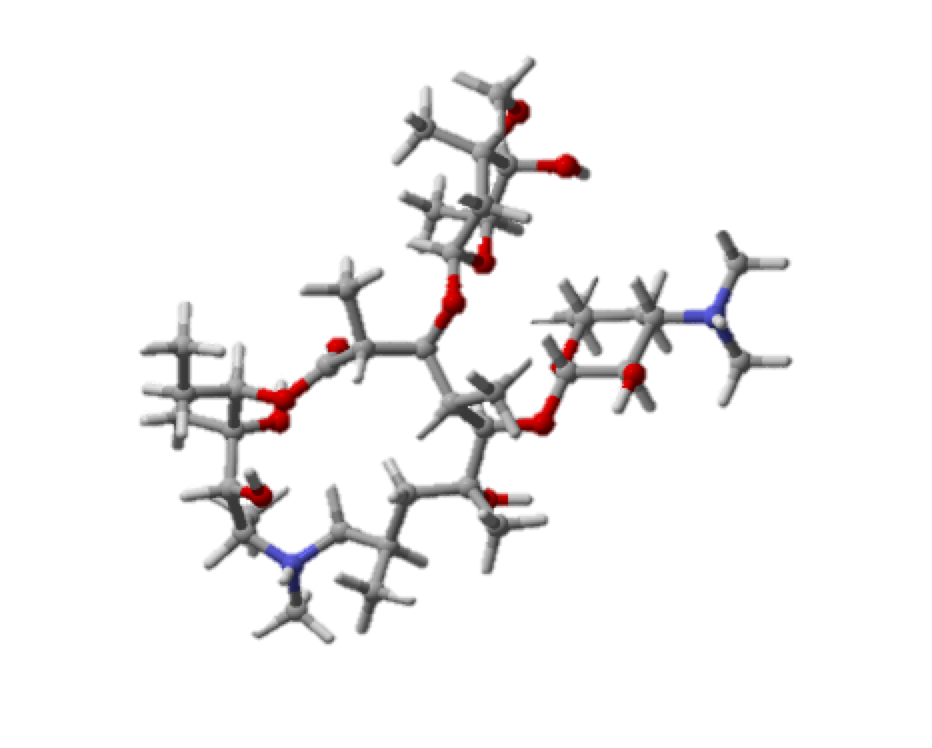
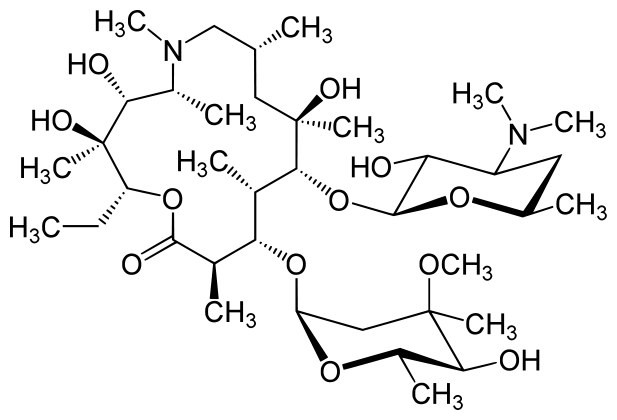


Figure 4: Multiple ways to represent azithromycin. Left is a 2D structure, middle is a 3D structure, and right is a 3D structure taking into account van der Waals surface area of the molecule. These are examples of the many different ways we can look at physiochemical properties of molecules.

**Correlation Matrix**

Rstudio was used to generate a correlation matrix for these descriptors with experimental data based on an online tutorial available at (<http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software>)

The output is a correlation matrix that displays Pearson correlation values, which range from -1 to 1 with 0 being no correlation, 1 being perfect positive correlation, and -1 being perfect negative correlation. The descriptor values were saved as a csv file and imported into R. The command

res <- cor(my\_data)

generates the correlation matrix which is then exported to a csv file. This correlation matrix csv file was then imported into excel. Conditional formatting was used to color code the range of correlation coefficients to allow for easy identification of strong correlations. Individual correlation plots were then linked to the cells of the correlation matrix that had correlation values greater than 0.70 or less than -0.70.

Our confidence in our results about the statistical significance of these correlations lies in the size of the data set. Originally the 12 antibiotics that had completed MD trajectories were used to identify which descriptors might be important. Later this test set was expanded to 20 molecules with MD trajectories. When promising correlations were seen here this was expanded upon to include all antibiotics with MIC data (62 antibiotics for *E. coli* and 44 antibiotics for PAE)

A spreadsheet (**e.coli\_pae\_correlated.xlsx**) in the data analysis folder contains the entire correlation matrix for the original 20 antibiotics. **All\_correlated.xlsx** contains the correlation matrix for all of the antibiotics with MIC data. The different rows are different experimental conditions, while the columns are the MOE descriptors. is color coded as a heat map to make it easy to note strong correlations. There is a key to the right to describe color-coding for the first 20. Extended description of the various descriptors and individual scatterplots are not included in all\_correlated.xlsx since most of the highly correlating trends are lost.

\*\*\*\*\*\*\*\*\*\*\*\* **Since there are so many descriptors please look here for the data \*\*\*\*\*\*\*\*\*\*\*\*\***

\* Red triangles in the corner for descriptors have more information if you hover your mouse over

the cell. Correlation matrix cells have an image of the scatterplot for highest R values to visualize trends.

\*\* The MIC data is only collected at certain concentrations so there are gaps in the scatterplots with no datapoints

**Grouping of Antibiotics**

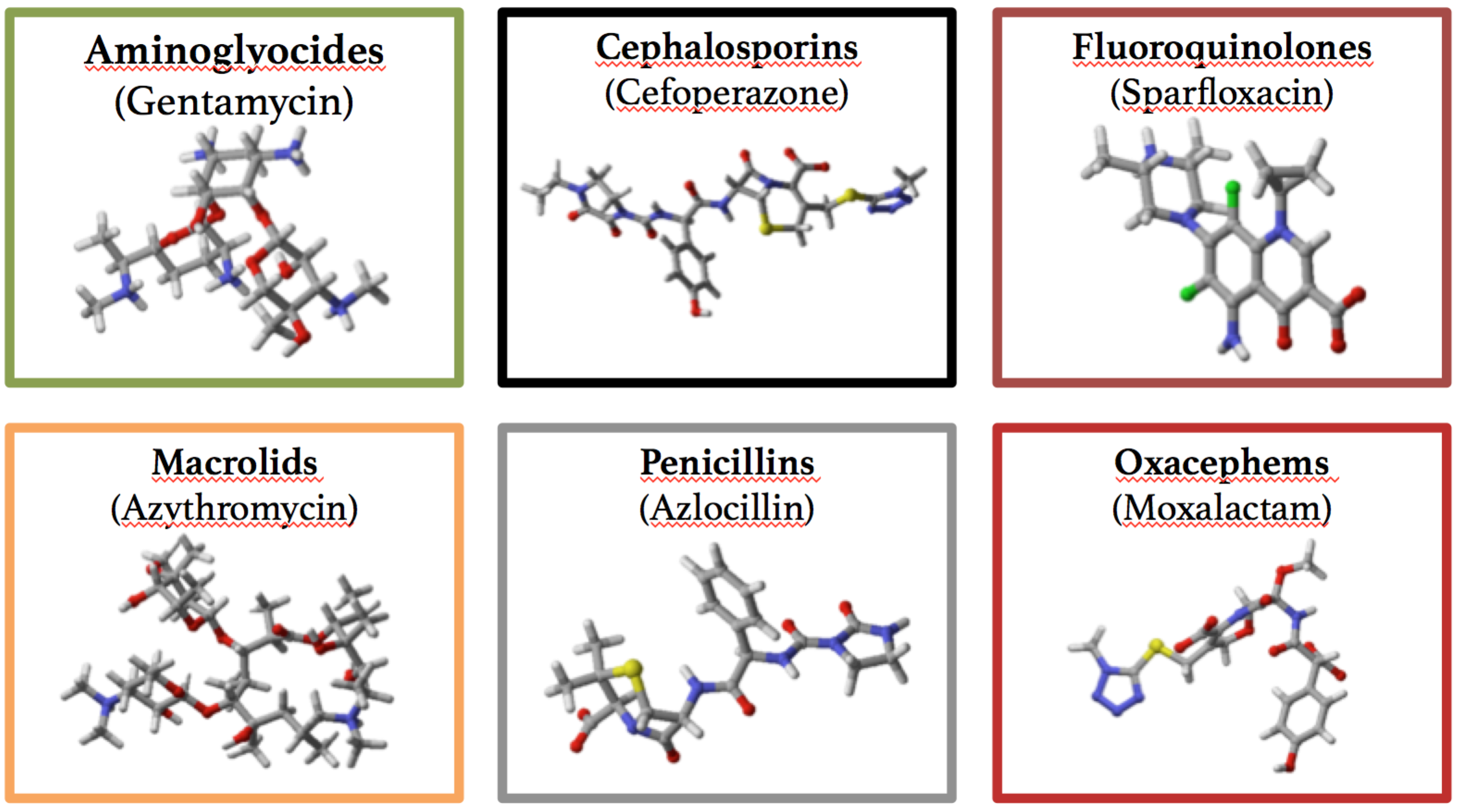
****

Figure 5: Representative structures of the various classes of antibiotics used.

The main challenge to this project was figuring out how to group the antibiotics. Since there is such a wide range of antibiotic classes as you can see in Figure 5. These structures range from large bulky molecules such as macrolids and aminoglyocsides to more linear molecules such as cephalosporins, penicillins, and oxacephens, to smaller, planar molecules such as fluoroquinolones. These all have different functional groups, charges, atoms, etc. and the available MIC data was a limiting factor in grouping since some classes had more data available than others. I grouped the antibiotics by class, but due to the small sample size the results are not shown as I was not confident in the statistical significance.

**Results**

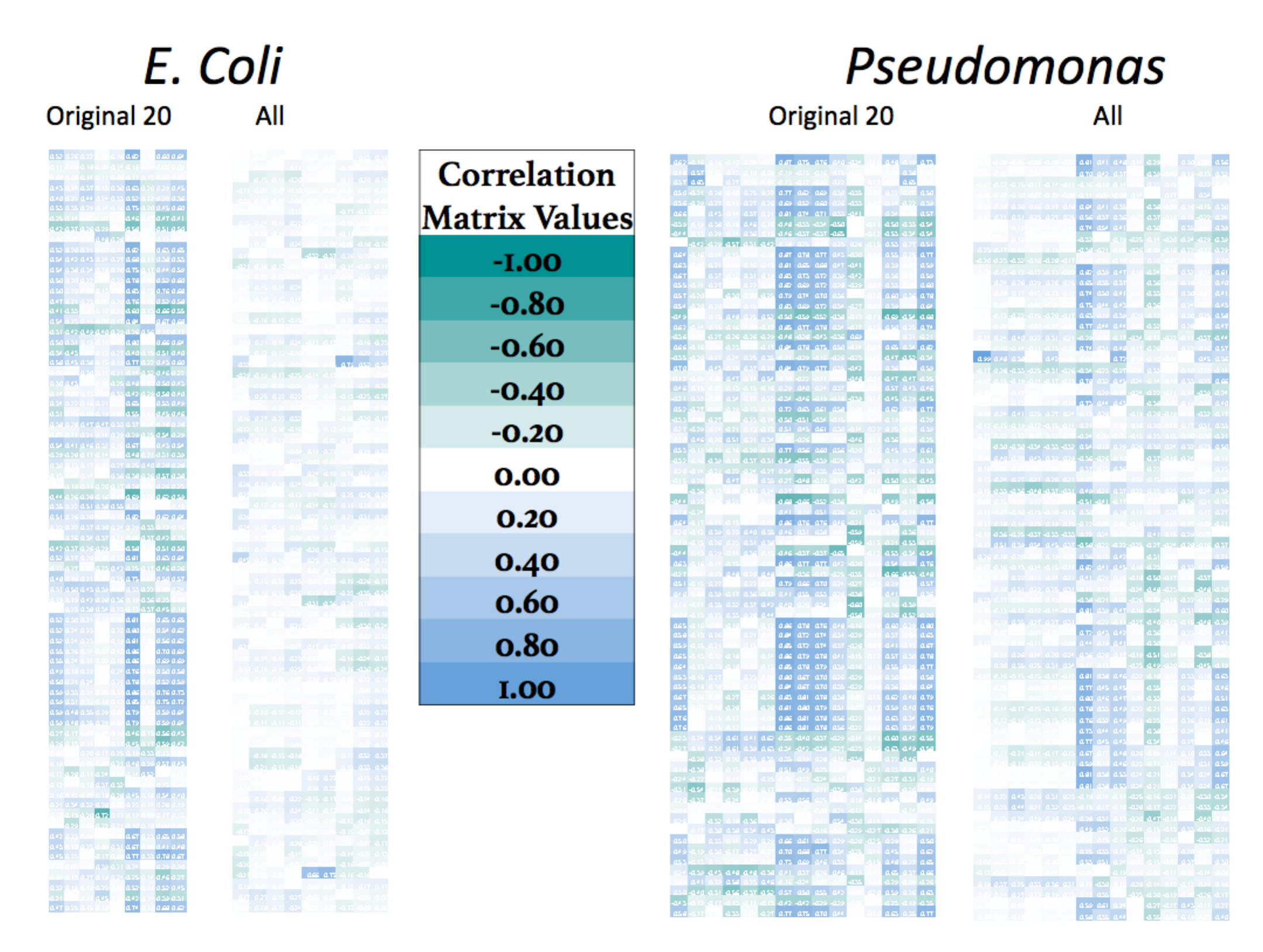


Figure 6: Zoomed out view of heat maps for correlations. To see R values and column/row labels see the excel spreadsheet.

In Figure 4 you can see that there are definite trends in correlation. For both *E. coli* and PAE these trends are similar for the original 20 antibiotics. Most correlations are positive, indicating a relationship such that as a property such as size or polarizibility increases, the MIC also increases. Overall the wild type showed fewer highly correlating values, while efflux pump knockouts and hyperporinated conditions showed more highly correlating values. When all of the antibiotics are considered for the comparisons most of the highly correlating values are lost. This stresses the importance of grouping so I plan to group them in various ways such as charge, antibiotic class, size, etc.

Example Scatterplots:



Figure 7: *E.* coli efflux pump knockout with outer membrane pore correlated with the first principal moment of inertia shows a positive correlation (R = 0.92).



Figure 8: *E. coli* efflux pump knockout with outer membrane pore correlated with the molar refractivity shows a positive correlation. (R = 0.79)



Figure 9: PAE outer membrane pore correlated with the hydrophobic van der Waals surface area shows a positive correlation. (R = 0.85).



Figure 10: PAE outer membrane pore correlated with the first kappa shape index shows a positive correlation (0.85).

**Conclusions**

* No strong correlations were seen for WT-MIC, WT pore MIC, ∆TOLC-MIC, WT/WT pore OM, and WT/∆TOLC pore efflux.
* ΔTolC/ΔTolC pore OM showed many significant correlations such as number of carbons, molecular weight, van der Waals surface area and volume, all of which indicate a size relationship.
* PAO1-Up MIC, Δ3-Up-MIC, Δ6-Up MIC, PAO1/Δ3-Efflux, Δ3/Δ6-Efflux, PAO1 Up/Δ3 Up Efflux, and PAO1-Up-Δ6-Up Efflux showed no strong correlations.
* Δ3–MIC, Δ6-MIC, and PAO1/Δ6 Efflux showed several moderate and a few strong correlations.
* PAO1/PAO1 Up OM, Δ3/Δ3 Up OM, Δ6/Δ6 Up OM, and PAO1/Δ6 Up Barrier all showed several strong correlations such as number of atoms, molar refractivity, kappa shape indices, atom connectivity indices, sum of atomic polarizability, VDW surface area and VDW volume. Several of these strongly correlating descriptors were also strongly correlated in *E. coli* ΔTOLC/ΔTOLC pore OM.