Imperial College
London

# *In silico* Pharmacological Target Characterisation of indole-3-acetate

Arinbjörn Kolbeinsson

Supervised by Dr Marc-Emmanuel Dumas and Professor Robert Glen

Department of Surgery and Cancer
Faculty of Medicine
Imperial College London
24 August 2016

# Statement of originality

I certify that this thesis, and the research to which it refers, are the product of my own work, conducted during the current year of the MRes in Biomedical Research at Imperial College London. Any ideas or quotations from the work of other people, published or otherwise, or from my own previous work are fully acknowledged in accordance with the standard referencing practices of the discipline.

The I3A kinase screening and GC-MS experiments were performed by other members of Dumas' lab. I was involved only in the data and network analysis of the latter.

The PCA, OPLS-DA, validation and normalisation MATLAB scripts that were used are part of the Imperial CSM software toolbox. Other MATLAB code was written by me.

The molecular dynamics python scripts were written by Dr Mark Williamson at the University of Cambridge. He also wrote the plotting functions which I used to generate the simulation figures and plot.

# Abstract

Changes in the human gut microbiota have been shown to impact human health and disease. The gut microbiota produces a number of metabolites that act as chemical messengers by binding to human targets and impacting metabolic and signalling pathways. Identifying these metabolites' targets and pathways is vital for interpreting disease mechanisms and valuable for identifying potential drug candidates. Preliminary data from Dumas' lab suggested a potential role for a specific gut microbial metabolite, indole-3-acetate, as an inhibitor for VEGFR-2. My research hypothesis is that I3A will dock to VEGFR-2, inhibit normal ATP binding and consequently influence VEGFR-2 activity and its downstream signalling and metabolic pathways. I demonstrate that indole-3-acetate can dock to the ATP binding cavity of VEGFR-2, blocking phosphorylation by ATP and therefore inhibiting the activity of VEGFR-2. The network analysis reveals an effect in two metabolites – stearic acid and putrescine – that are, respectively, involved in lipid metabolism and nitric oxide and AKT activity. Inhibiting VEGFR-2 using I3A could prove an effective strategy in suppressing angiogenesis in tumours.

# Introduction

## Microbiota

The human microbiota is the population of non-pathogenic bacteria and other microorganisms that share body space with their human host (Costello *et al.* 2012). These microorganisms hold a symbiotic relationship with the host and are indirectly involved in a number of vital human metabolic and signalling pathways. The mutually beneficial relationship involves the host providing energy and nutrients for the residents of the microbiota, which then process them into useful metabolites for the host. A wide range of factors have been shown to affect the microbial ecology. Dietary supplements can influence gene function and contribute to food intolerances (Suez *et al.* 2014) and antibiotics during early life development can have permanent health effects including increased potency of high fat diet induced obesity (Cox *et al.* 2014).

The metabolites that are produced by the microbiota have a variety of different human targets. Imbalances to the microbiota have been linked to a number of conditions including obesity (Cani *et al.* 2008), diabetes (Wen *et al.* 2008) and Crohn's disease (Sokol *et al.* 2008). Its role in cardiometabolic diseases has also been studied (Neves *et al.* 2015). The gut microbiota produces metabolites, including deoxycholic acid (DCA) and lithocholic (LCA) acids, which can impact cardiovascular disease outcomes by regulating properties such as appetite and glucose homeostasis. The microbiota has also been shown to educate the host’s immune system by inducing the production of specialised T cells, which give a higher tolerance to the microbiota (Lathrop *et al.* 2011).

## Key microbial metabolite families in health and disease

Indoles are a family of small molecules that have been heavily linked with the gut microbiota (Neves *et al.* 2015), and are produced by bacteria that convert tryptophan into indole (Jensen *et al.* 1995). Indoles have been shown to bind to human receptors. Pregnane X receptor (PXR) is a target for indole-3-propionic acid (IPA) (Venkatesh *et al.* 2014). PXR regulates intestinal barrier function and therefore IPA has a direct effect on mucosal immunity. 3-indoxylsulphate modulates aryl hydrocarbon receptor (AhR). Current literature suggest that AhR is a factor in cardiometabolic disease by mediating inflammation (Neves *et al.* 2015). It is apparent that indoles play an important role in the human – gut microbiome relationship.

Indole-3-acetate (I3A) is a naturally occurring molecule with a molecular weight of 175.19 $g \cdot mol^{-1}$. Produced by plants as a way to regulate their growth, this molecule is abundant in people who have a vegetable-stem rich diet (Simon & Petrášek 2011) (Holmes *et al.* 2007). I3A has been shown to interact with human signalling networks by binding to receptors such as the AhR (Jin *et al.* 2014), but the understanding of I3A's effects on humans is not complete. Is I3A limited to receptor binding, or are there other potential targets?

Preliminary results from a recent screening of 456 kinases by Dumas' lab identified vascular endothelial growth factor receptor 2 (VEGFR-2) as the only target for I3A. I3A was shown to inhibit VEGFR-2 with a dissociation constant, $K_d$, of 930 nM.

## VEGFR-2 and Angiogenesis

VEGFR-2 is one of three subtypes of receptors for vascular endothelial growth factors (VEGF). The signalling mode of action for all three receptors is through cell surface

tyrosine kinases. This causes them to dimerise and activate as a result of transphosphorylation (Swain *et al.* 2003). VEGFR-2 is the most studied of the three subtypes and is known to be directly involved in angiogenesis (Holmes *et al.* 2007).

Angiogenesis is the natural process of proliferation of new blood vessels from existing ones. It is different to vasculogenesis, which is *de novo* generation of new blood vessels during embryonic development. Angiogenesis plays a number of key roles in normal processes such as wound repair and recruiting new blood vessels after prolonged exercise (Carmeliet & Jain 2000). Over the last decades, the role of angiogenesis in tumour growth has been demonstrated. In order for a tumour to reach a critical size, additional blood vessels are required to provide essential nutrients. Regulating pro-angiogenesis molecules and angiogenesis signalling pathways has therefore been studied as a strategy to arrest tumour growth (Folkman 1995).

## Cheminformatics

The docking mechanism of ligands to receptors has become an essential part of drug discovery studies. They can be modelled using computer simulations and given a quantitative score depending on the predicted favourability and likelihood of a docking (Korb *et al.* 2009). One of the prominent docking programs is GOLD, produced by the Cambridge Crystallographic Data Centre. Beyond docking, molecular dynamics provides a way to simulate the entire set of atoms in a molecular system. Recent computer optimisations to both software algorithms and hardware have led to great improvements in molecular dynamics' accuracy and simulation durations (Solomon-Ferrer 2013). These simulations can provide a perspective on the conformational and energy evolutions that the system undergoes (Karplus & Kuriyan 2005).

## Project aims

My research hypothesis is that I3A will dock to VEGFR-2, inhibit normal ATP binding and consequently influence VEGFR-2 activity and its downstream signalling and metabolic pathways. One process which will be suppressed by VEGFR-2 inhibition is angiogenesis. Investigating the molecular interactions is vital for understanding this mechanism.

Therefore, I aim to complete the following objectives:

1. Model the docking process of I3A and VEGFR-2
2. Investigate the mode of I3A inhibition of VEGFR-2
3. Simulate VEGFR-2 using molecular dynamics to study its properties, stability and conformation
4. Investigate the metabolic signature associated with I3A-induced inhibition of VEGFR-2 in endothelial cells
5. Explore and visualise the signalling and metabolic networks impacted by VEGFR-2 inhibition

# Methods

## Molecular docking

The VEGFR-2 protein file was obtained from RCSB PDB as a .pdb file (PDB id: 5EW3) and was pre-processed using the SYBYL software platform (Tripos, St. Louis, MO). The first pre-processing step was to extract monomer A from the PDB file. The X-ray structure contained two symmetry related molecules, but only one is necessary for the subsequent simulations. The waters, identified as lone oxygen molecules (hydrogens are not resolved at this resolution), were then removed. This was done as the desolvation of the ligand and the protein surface are taken into account in the scoring function. Next, the two histidine residues were fixed, but they were not in the vicinity of the binding site. The protein was then protonated at pH = 7 and minimised using 100 iterations.

The co-crystallized structure of VEGFR-2 in complex with a ligand, AAL993, had been established using X-ray crystallography in a previous study (Bold *et al.* 2015). The docking parameters were optimised by binding the AAL993 to VEGFR-2, and validating that the computational model matched experimental observations.

Both the receptor, VEGFR-2, and ligand, I3A, files were loaded into GOLD in the format .pdb and .mol2, respectively. Using results from the validation, the binding site was defined as a sphere with a radius of 8 Å around the centre carbon atom in VAL848. The option to automatically detect the receptor cavity was selected. This limits atom binding events to solvent-accessible surfaces. Additionally, all hydrogen-bond donors and acceptors are forced to be considered as solvent accessible. The built-in

"chemscore_kinase" configuration template was chosen as it matched the requirements of the docking, and early termination was disabled to constrain the scoring of exactly 10 solutions. The internal ligand energy offset was used in each one. GOLD gives the option to prioritise either accuracy or speed of the docking calculation. In all cases, the highest accuracy setting was selected.

The CHEMPLP scoring function was chosen because it has been shown to provide the most consistently accurate docking predictions of those provided by GOLD (Korb *et al.* 2009). CHEMPLP is a *piecewise linear potential* (PLP) empirical scoring function. It is designed to be optimal for protein-ligand interactions by modelling the steric interactions between the two molecules and is a function of heavy-atom clash potential, torsional potential and hydrogen donors and acceptors. (Korb *et al.* 2009). The same method was used for docking ATP and the program parameters were kept consistent with the I3A dock to ensure compatibility.

To investigate the effects of I3A being bound to VEGFR-2 on ATP docking, a molecule of the merged structures of VEGFR-2 and I3A bound in its optimal site was created. This was done using PyMOL and the entire molecule minimised using PyMOL's sculpting tool. Visualisations were also done in PyMOL.

## Molecular dynamics

*The original python scripts for this procedure were written by Dr Mark Williamson of Cambridge University.*

I used the OpenMM toolkit (Eastman *et al.* 2012) to run molecular dynamics simulations on VEGFR-2. The VEGFR-2 pdb file was pre-processed using PDBFixer, which is part of the OpenMM suite. PDBFixer automatically adds missing atoms, repairs erroneous

residues and performs the following preparation steps. The force fields were assigned using Amber FF99SB parameters. Next the protein was solvated in water using the `modeller.addSolvent` function. The chosen water model was tip3p with a 10 Å padding around VEGFR-2. I then minimised the protein energy to using 1000 iterations. This was done to relax the structure and correct any potentially unstable hydrogen bonds. The whole system was then heated under constant NVT conditions to 300 K for a total of 70 ps. SHAKE was enabled with a non-bonded cut-off of 8 Å. After the heating cycle, I corrected the density of the system under NPT conditions for another 70 ps at 1 bar and 300 K. The final production run was a 10 ns simulation under constant NPT conditions. The time step was 2 ps, giving a total of 5,000,000 steps and the structure and atom velocities were saved to a file every 5,000 steps.

## GCMS dataset

*The following wet lab work was performed by other members of Dumas' lab.*

Human umbilical vein endothelial cells (HUVECs) were seeded into 96-well plates. When they reached 90% confluence, they were divided into four groups. All groups were treated with 20 ng/ml of vascular endothelial growth factor (VEGF-A). One group was treated with VEGF-A only while the other three groups were additionally treated with 0.01 μM, 1 μM and 100 μM of I3A, respectively. After 24 hours, the supernatants of the HUVECs were collected and centrifuged at 17,000g and 4°C for 5 minutes and then stored at -20°C. The standard GC-MS profiling protocol for cell supernatants was then followed.

The data I was provided were the binary logarithm of the raw values from the machine. I converted these into the common logarithm of the raw values using MATLAB. The coefficient of variation (CV) = standard deviation(x) / mean(x) was calculated for the

quality control samples of each identified feature. I then used the CVs to filter the features to those with CV < 15% to include only those that showed consistency over the quality control nine samples. I analysed the filtered data using principal component analysis (PCA) to visualise the clustering of the groups and gain a better understanding of the dataset. The MATLAB PCA function that was used is available as part of the Imperial CSM toolbox, and originally written by Dr Jake Pearce.

Using the built-in `fitlm` function in MATLAB I fitted a linear regression model for each feature. The model was constructed in *log10* space in order to obtain reduce the skewness of the independent variable, concentration of I3A. This gave sample concentrations of -8, -6 and -4. The samples with no added I3A were approximated to have a concentration of $10^{-10}$, equivalent to -10 in log10 space. Features with *p*-values < 0.05 were selected and studied further using network analysis. The $r^2$ of each fit was calculated and adjusted for the number of coefficients, which were two in this case.

I constructed a network using MetaboSignal (Rodriguez-Martinez, A. 2016). The network consisted of a number of common human metabolic pathways from KEGG: [**hsa00010, hsa00020, hsa00030, hsa00040, hsa00051, hsa00052, hsa00053, hsa00061, hsa00062, hsa00071, hsa00072, hsa00100, hsa00120, hsa00130, hsa00140, hsa00220, hsa00230, hsa00240, hsa00250, hsa00260, hsa00270, hsa00280, hsa00290, hsa00300, hsa00310, hsa00330, hsa00340, hsa00350, hsa00360, hsa00380, hsa00400, hsa00410, hsa00430, hsa00450, hsa00460, hsa00471, hsa00472, hsa00480, hsa00500, hsa00520, hsa00561, hsa00562, hsa00565, hsa00590, hsa00591, hsa00592, hsa00600, hsa00620, hsa00630, hsa00640, hsa00650, hsa00670, hsa00730, hsa00740, hsa00750, hsa00760, hsa00770, hsa00780, hsa00785, hsa00790, hsa00830, hsa00860, hsa00900, hsa00910, hsa00920, hsa00982, hsa00524, hsa00232, hsa00564, hsa00563, hsa00980, hsa00510, hsa00512, hsa00532, hsa00534, hsa00531, hsa00563, hsa00601, hsa00603, hsa00604, hsa00970** ]. And the signalling pathways: [**hsa04014, hsa04015, hsa04010, hsa04012, hsa04310, hsa04330, hsa04340,**

**hsa04350, hsa04390, hsa04370, hsa04630, hsa04064, hsa04668, hsa04066, hsa04068, hsa04020, hsa04072, hsa04071, hsa04024, hsa04022, hsa04151, hsa04152, hsa04150, hsa04910, hsa04922, hsa04920, hsa03320, hsa04912, hsa04976, hsa04930, hsa04932**]. I then selected added all the KEGG metabolic pathways which contain the previously identified features and the VEGFR-2 signalling pathway to the network. MetaboSignal was used to calculate the shortest paths (Davidovic *et al.* 2011) between VEGFR-2 and each identified feature. The resulting subnetwork was then exported and plotted using Cytoscape (Shannon et al., 2003).

# Results

To study the effects of I3A on VEGFR-2 and metabolic and signalling pathways in detail, I implemented two complimentary strategies. Firstly, the biophysical aspect was studied by incorporating molecular docking and molecular dynamics to characterise the interaction of VEGFR-2, ATP and I3A. Then, these results were integrated with signalling, metabolomic and molecular networks to form a biochemical approach. The initial step was to study the molecular docking properties using the GOLD software program.

## Validation and ATP docking of VEGFR-2

In order to validate the approach, we first checked that the co-crystallised ligand AAL993 could be docked into the predicted crystallographic position.

The AAL993 ligand was shown to bind to the VEGFR-2 ATP pocket in the same conformation as the benchmark X-ray co-crystallised structure indicated in Figure 1A. The root-mean-square deviation (RMSD) was 0.14 Å, meaning that the GOLD docking simulation was successful and accurate.

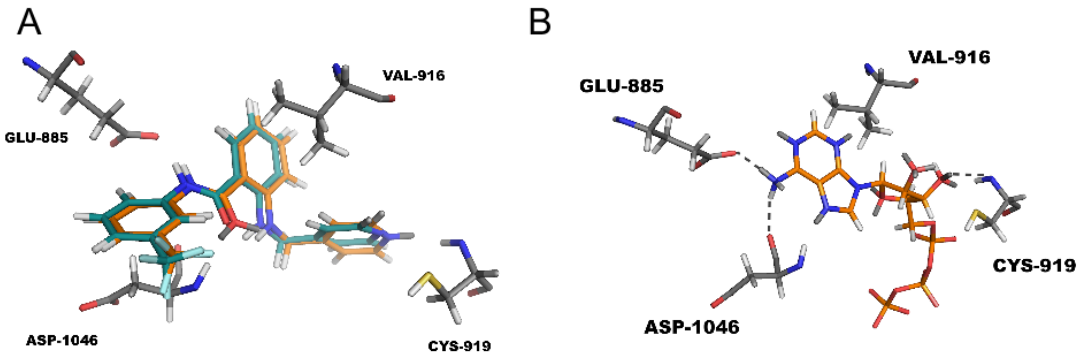
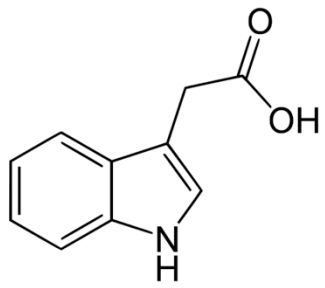



**Figure 1.** Validation of GOLD docking **A)** Comparison of the crystallographic position of AAL993 (in blue) and the position predicted by GOLD (in orange). The surrounding VEGFR-2 residues are illustrated in grey. **B)** The best GOLD solution for ATP (orange). The VEGFR-2 residues of interest are in grey and labelled. Hydrogen bonds are represented as dashed lines.

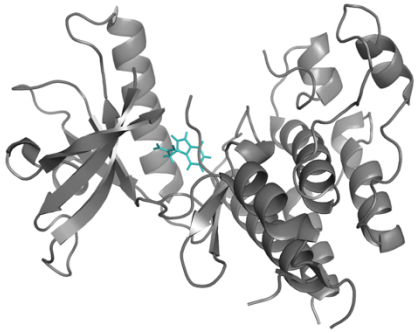
The ATP binding mode was determined by a GOLD docking run with VEGFR-2 as the protein and ATP as the ligand (Figure 1B). The highest scoring solution formed hydrogen bonds with ASP-1046, GLU-885, CYS-919, GLU-917, LYS-868 and ANS-923. Out of these six residues, four correspond with the previously identified ATP binding cavity of VEGFR-2 (Bold *et al.* 2015). These results re-demonstrate this binding site. Having benchmarked the docking models, I then moved on to I3A docking.

## Identification of an I3A docking site by Molecular Docking

A




B



C

| Solution number | ChemPLP score | Interacting VEGFR-2 residue |
|---|---|---|
| 1 | 45.58 | CYS-919 |
| 2 | 46.40 | CYS-919 |
| 3 | 45.86 | ASP-1046 |
| 4 | 47.09 | CYS-919 |
| 5 | 45.96 | CYS-919 |
| 6 | 45.93 | ASP-1046 |
| 7 | 45.88 | ASP-1046 |
| 8 | 49.89 | ASP-1046 |
| 9 | 46.34 | CYS-919 |
| 10 | 49.82 | ASP-1046 |

D

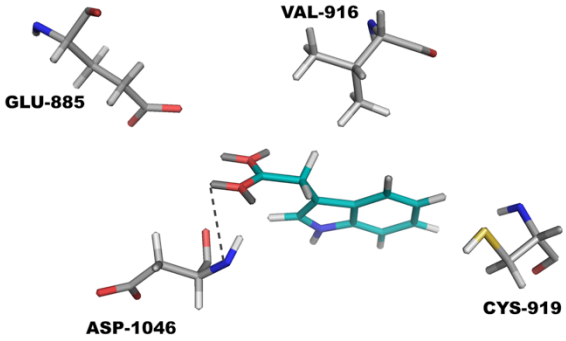



**Figure 2.** I3A can be docked into the ATP binding cavity of VEGFR-2 **A)** Structure of indole-3-acetate (I3A) **B)** Overview of the best binding site of I3A (cyan) on VEGFR-2 (grey) **C)** Table of docking solutions of I3A calculated by GOLD. **D)** The highest scoring I3A docking solution (number eight) in cyan and the surrounding VEGFR-2 residues of interest (grey). The hydrogen bond is represented with a dashed line and has an angle of 117.35°.

The I3A docking in GOLD produced ten solutions. Every solution formed a hydrogen bond with either the ASP-1046 or CYS-919 residue on VEGFR-2. The highest scoring docking solution forms a hydrogen bond between one of I3A's acetate oxygen atoms and the backbone nitrogen atom in the ASP-1046 residue. This is shown in Figures 2B and 2D. The bond length is 2.74 Å, shorter than the sum of van der Waals radii of the two atoms, indicating a hydrogen bond. The angle of this bond is 117.35°. The alternative docking mode of I3A is to form a hydrogen bond with the CYS-919 residue. The hydrogen of the indole nitrogen bonds (donor) with the OH group on CYS-919 with a bond length of 2.66

Å. Both ASP-1046 and CYS-919 have been identified as components in the ATP binding pocket of VEGFR-2 (Bold *et al.* 2015). A more detailed evaluation of the results reveals that the CHEMPLP score of the best docking solution was 49.89. The piecewise-linear-potential (PLP) score is reported as -43.47. Hydrogen bond score was 1.00 and C–H···O interaction score was 0.61.

The scores of the 10 docking solutions calculated using the GOLD run are shown in Figure 2C. Solution 8 provides the highest score and therefore the highest binding affinity. Five of the solutions form a bond with CYS-919 and the other five form a bond with ASP-1046. Interestingly, the five ASP-1046 bonded solutions are not clustered together using a complete linkage algorithm of the RMSD (see supplementary information). This suggests that an I3A molecule will not have to move far to form a hydrogen bond with a different residue.

## I3A inhibition of ATP

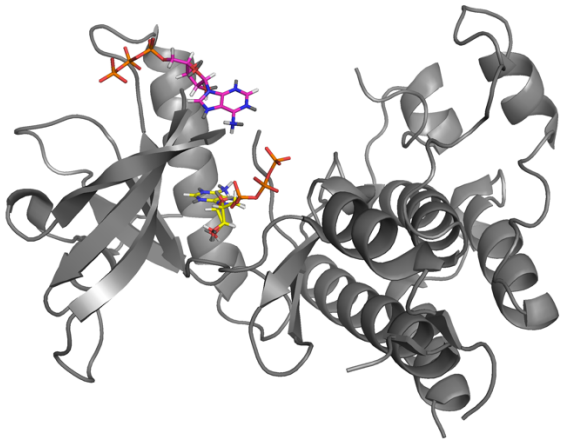


**Figure 3.** I3A inhibits normal ATP binding of VEGFR-2. Superposition of two ATP docking models: in yellow is the best ATP position in the absence of I3A. If I3A is docked in its highest scoring binding site, then the optimal ATP binding site moves out of the binding cavity (ATP molecule in magenta).

I3A was shown to inhibit VEGFR-2 by preventing the binding of ATP. Docking the merged structure of I3A and VEGFR-2 and ATP in GOLD revealed that the best docking solution

of ATP is outside the binding pocket. This altered docking conformation suggests that ATP is not able to interact with any of the residues normally involved in ATP binding. Figure 3 shows that ATP will adopt a different binding site in the event of I3A docking. In order to gain a better understanding of the properties of VEGFR-2, molecular dynamics simulations were run on the entire molecule.

## Molecular dynamics simulation of solvated VEGFR-2

Molecular dynamics was used to simulate the entire set of atoms in the VEGFR-2 protein. This gives a comprehensive representation of the system including the statistical distribution of the protein's conformations.

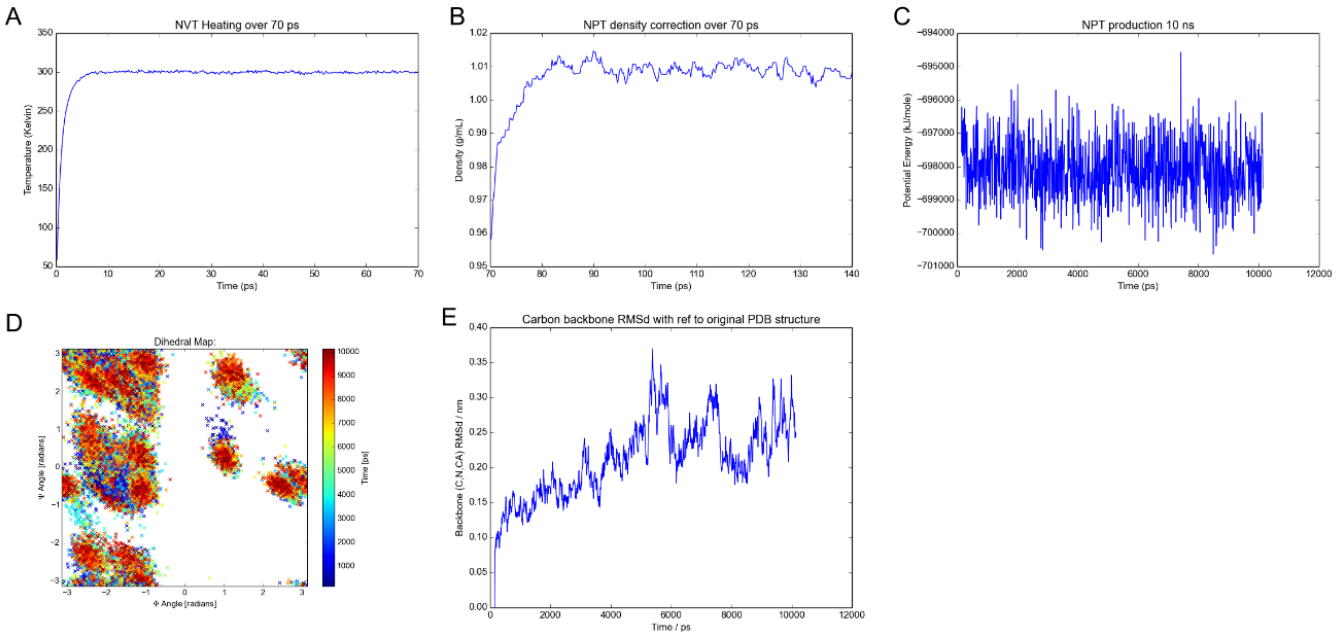



**Figure 4. A)** NVT heating to 300 K over 70 ps. The molecule stabilises after 7 ps **B)** NPT density correction over 70 ps. Note the axis range from 70-140 ps, with T = 0 representing the start of the heating **C)** Potential energy of the NPT production simulation over 10 ns. The potential energy range is 10,000 kJ/mole. **D)** Dihedral map, or Ramachandran plot, of the VEGFR-2 backbone angles. **E)** RMSD plot of the carbon backbone of VEGFR-2 over a 10 ns production run.

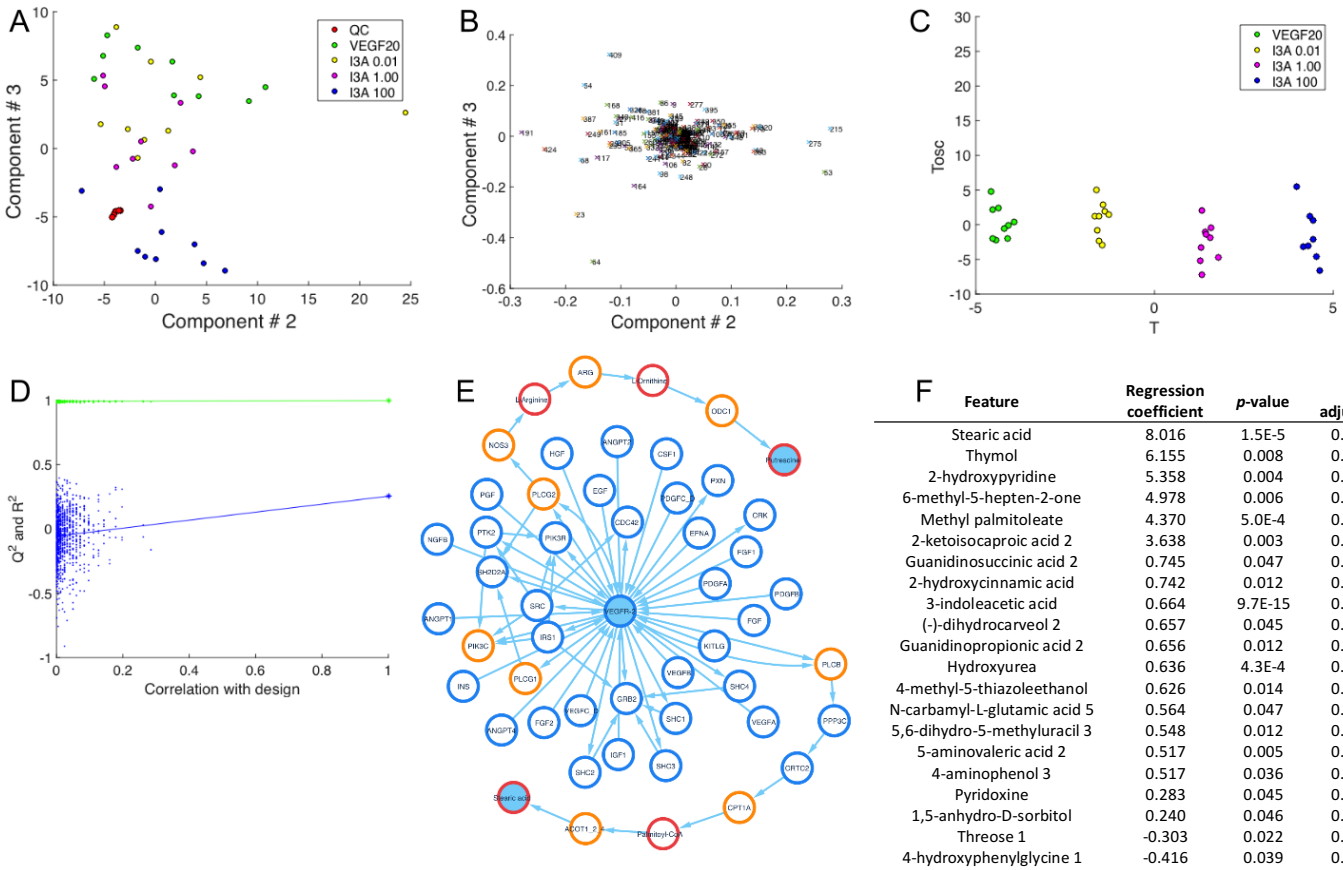
The solvated VEGFR-2 molecule was minimised using 1000 iterations and then heated to 300K, as shown in figure 4A. It showed stable density after 14 ps under NPT density correction (Figure 4B). The potential energy evolution of the entire 10 ns production run ranges from -705,000 to -695,000 (Figure 4C). The Ramachandran plot (Figure 4D)

highlights the distribution of possible $\Psi$ and $\Phi$ angles. There are a few clusters of stability, particularly the region $-\pi$ rads < $\Phi$ < -0.25 $\pi$ rads. Figure 4E shows RMSD from the original carbon and nitrogen backbone. The peaks at around 5.5 ns and 7.5 ns correspond with the two isolated clusters seen in the Ramachandran plot. This could suggest conformational flips in the backbone of the molecule and could have an effect on docking. These simulations are necessary for further simulations involving the VEGFR-2 and I3A which can reveal the spatial probability distribution of I3A and give more insight into the molecular docking properties of I3A. The downstream effects of VEGFR-2 inhibition by I3A was investigated using GC-MS metabolomics.

## GC-MS data analysis

The high sensitivity of GC-MS (Kanani *et al.* 2008) was valuable when studying the effects of I3A docking on VEGFR-2 pathways. Affected pathways would result in a change in the concentration of metabolite products, which is detectable with GC-MS technology.

The initial GC-MS dataset that was provided contained 436 features. These were filtered to include only features whose quality control samples had a coefficient of variation of less than 15%. A limit of 30% was also evaluated, but a 15% threshold provided preferable clustering after PCA. Pre-processing using 5% dynamic range offset addition, probabilistic quotient normalisation, centring and unit variance scaling gave poorer clustering than limiting by coefficient of variation to 15% (see supplementary information).




| Feature | Regression coefficient | *p*-value | $R^2$ adjusted |
|---|---|---|---|
| Stearic acid | 8.016 | 1.5E-5 | 0.411 |
| Thymol | 6.155 | 0.008 | 0.163 |
| 2-hydroxypyridine | 5.358 | 0.004 | 0.198 |
| 6-methyl-5-hepten-2-one | 4.978 | 0.006 | 0.178 |
| Methyl palmitoleate | 4.370 | 5.0E-4 | 0.283 |
| 2-ketoisocaproic acid 2 | 3.638 | 0.003 | 0.208 |
| Guanidinosuccinic acid 2 | 0.745 | 0.047 | 0.085 |
| 2-hydroxycinnamic acid | 0.742 | 0.012 | 0.148 |
| 3-indoleacetic acid | 0.664 | 9.7E-15 | 0.827 |
| (-)-dihydrocarveol 2 | 0.657 | 0.045 | 0.087 |
| Guanidinopropionic acid 2 | 0.656 | 0.012 | 0.148 |
| Hydroxyurea | 0.636 | 4.3E-4 | 0.289 |
| 4-methyl-5-thiazoleethanol | 0.626 | 0.014 | 0.141 |
| N-carbamyl-L-glutamic acid 5 | 0.564 | 0.047 | 0.085 |
| 5,6-dihydro-5-methyluracil 3 | 0.548 | 0.012 | 0.146 |
| 5-aminovaleric acid 2 | 0.517 | 0.005 | 0.190 |
| 4-aminophenol 3 | 0.517 | 0.036 | 0.097 |
| Pyridoxine | 0.283 | 0.045 | 0.087 |
| 1,5-anhydro-D-sorbitol | 0.240 | 0.046 | 0.086 |
| Threose 1 | -0.303 | 0.022 | 0.120 |
| 4-hydroxyphenylglycine 1 | -0.416 | 0.039 | 0.094 |
| Putrescine | -0.455 | 0.019 | 0.126 |

**Figure 5.** Two pathways were found to be affected by I3A inhibition of VEGFR-2 **A)** PCA scores plot of GC-MS data of HUVECs treated with different concentrations of I3A **B)** PCA loadings plot of the same analysis. **C)** OPLS-DA scores plot of the GC-MS dataset. There is clear separation of the four groups. **D)** Cross validation of the OPLS-DA model over 1000 permutations. $Q^2$ in blue and $R^2$ in green. The *p*-value is 0.038 **E)** Signalling and metabolic network around VEGFR-2. The network contains signalling genes (blue outline) and metabolic genes (orange outline) that can be connected to VEGFR-2 with a distance of one. The pathways connecting VEGFR-2 to putrescine and stearic acid are also shown in the network with metabolites in red. Nodes of interest are filled in blue. **F)** Table of statistically significant features (*p*-value < 0.05) identified in the GC-MS experiment. Regression coefficient represents the slope of the linear fit.

The PCA scores are visualised in figure 5A. The quality control samples are grouped together indicating favourable clustering. Most of the separation between the four classes of varying I3A concentration is seen along the third component (y-axis). The loadings plot (figure 5B) illustrates the features which explain the variations on the second and third components. The third component is primarily controlled by features number 409 (threo-3-hydroxy-L-aspartate), 54 (2,3-butanediol), 64 (I3A) and 23 (1,4-dihydroxy-2-naphthoic acid). I3A's effect on component three is negative, meaning higher concentrations of I3A correspond to lower component three values on the PCA scores. OPLS-DA was carried out on the four groups. Figure 5C illustrates these scores

and a clear separation of the four experimental groups is observed. The OPLS-DA model was constructed using three components as this provided good prediction accuracy while not overfitting (see Figure 5D and supplementary data for $R^2$ and $Q^2$). The data were mean centred and scaled to unit variance prior to the fitting. The model was validated using 1000 permutations. The significance of the model was a *p*-value of 0.038.

Linear regression analysis on the features revealed 21 features that correlated with I3A concentrations to a statistically significant amount (*p*-value < 0.05). 10 of these features are not part of any published pathways on KEGG database. I3A was also identified as a highly correlated feature, supporting the validity of the analysis. The features and their regression coefficient (slope), *p*-value and adjusted $r^2$ are displayed in figure 5F.

The network analysis of each feature identified two metabolites, putrescine and stearic acid, which are connected to VEGFR-2 through human metabolic and signalling pathways (Figure 5E). These metabolites' correlation with inhibited VEGFR-2 activity establishes that the pathways are affected by I3A docking to VEGFR-2. Putrescine showed negative correlation with increased I3A concentration, as higher concentrations of I3A resulted in lower putrescine concentration. Preliminary experimental results from Dumas' lab match these findings. These in vitro results suggest that NOS and AKT are downregulated by VEGFR-2 inhibition. The current theory is that I3A inhibits AKT phosphorylation and therefore suppresses NO production. As putrescine is a component in this pathway, it is also suppressed. Stearic acid was the other VEGFR-2 connected metabolite. It is connected through a gene, CPT1A, which controls mitochondrial oxidation. Downregulating this pathway might be contributing to the build-up of fatty acids including stearic acid outside the mitochondrial membrane.

# Discussion

My objectives were to study the molecular interaction I3A and VEGFR-2, the docking properties and the effect of I3A-induced VEGFR-2 inhibition on metabolic and signalling pathways. I have shown that *in silico* docking simulations of I3A and VEGFR-2 identify a binding cavity for I3A in the ATP binding pocket of VEGFR-2, specifically with residues ASP-1046 and CYS-919. I3A inhibits VEGFR-2 activity through an ATP-competing effect. I simulated a solvated VEGFR-2 molecule using molecular dynamics, and showed that the system was stable after 5 ns and exhibited small conformation changes. Network analysis of GC-MS data of HUVECs treated with I3A identified two pathways which are affected by I3A binding to VEGFR-2.

The binding cavity I determined for I3A had been identified as a docking site for another molecule in a previous study (Bold *et al.* 2015). The docking of ATP to the residues in this binding cavity is required for transphosphorylation and subsequent dimerisation of the protein to take place (Holmes *et al.* 2007). My results also show that an I3A molecule docked to the ATP binding cavity of VEGFR-2 will inhibit these processes from occurring by inhibiting ATP binding and therefore affect downstream VEGFR-2 signalling events.

### Therapeutic impact

One of the major signalling pathways that VEGFR-2 is involved in is angiogenesis. Inhibiting this pathway can suppress angiogenesis and therefore suppress additional tumour growth (Dias *et al.* 2001) (Folkman 1995), which is a promising cancer treatment strategy. VEGFs have been shown to be upregulated in tumours (Plate *et al.*

1993), compounding the role of VEGFR-2 in tumour growth. Anti-angiogenesis treatments are slowly managing to match their initial predicted potential. Recently, a small number of anti-angiogenesis drugs have entered the market (fda.gov, 2004). These include bevacizumab (trade name Avastin®). Bevacizumab works by inhibiting vascular endothelial growth factor and therefore reducing VEGFR-2 activity. Currently, the United States Food and Drug Administration (FDA) has approved bevacizumab for use, in combination with other drugs, in treatment of colorectal cancer (fda.gov, 2004) and it remains actively studied (Chinot *et al.* 2014). In 2011, the FDA revoked the approval of bevacizumab use in the treatment of breast cancer, as the evidence of increased survival was outweighed by the potential risks of bevacizumab (National Cancer Institute, 2016). Since I3A is a naturally occurring molecule and found in healthy humans, it is conceivable, albeit undocumented, that the side effects of I3A are relatively minor.

In the past decade, the importance of the gut microbiota on human health has become increasingly apparent. Achievements have been made in both understanding and modulating the microbiota with the use of targeted drugs (Jia *et al.* 2008). Computational techniques have vastly improved drug design and efficacy prediction accuracy (Kubinyi 1998). I3A has many desirable drug properties. The drug properties were explored by running the ligand through the drug toxicity prediction program FAF-Drugs3 (Free ADME-Tox Filtering Tool) (Lagorce *et al.* 2015). It passed Lipinski's rule of five (Lipinski *et al.* 2015), the Veber rule (Veber *et al.* 2002), Egan rule (Egan *et al.* 2000) and the Bayer oral Physchem Score (Lobell *et al.* 2006) criteria. It also passed GSK's 4/400 rule (Gleeson 2008), which is a key component for drug safety evaluation. I3A scored high in the vast majority of other filter categories. Notable exceptions include

a low $Fsp^3$ score, where $Fsp^3$ is the fraction of the total number of carbon bonds which are $sp^3$ hybridised (Lovering *et al.* 2009). Additionally, given I3A's low molecular weight (175.19 $g \cdot mol^{-1}$) relative to its octanol:buffer distribution coefficients it falls narrowly outside the Golden Triangle (Johnson *et al.* 2009) of optimal permeability and metabolic stability.

Additional exploration of VEGFR-2 was accomplished using molecular dynamics. The simulations revealed that the backbone of VEGFR-2 slowly shifts from its original structure to a total RSMD of around 0.2 nm. The other backbone shifts that are observed later in the simulation suddenly change the backbone structure and could in theory interfere with I3A binding, although this is unlikely. Simulations by which VEGFR-2 is solvated in an I3A solution would demonstrate the probability density of I3A molecules. It would also reveal if the probability density changes depending on the backbone structure of VEGFR-2.

The identification of two affected pathways describe the effect of I3A-induced inhibition of VEGFR-2 on signalling and metabolic networks. Putrescine and stearic acid were negatively and positively, respectively, correlated with increased I3A concentration. Stearic acid is connected via a pathway containing the CPT1A gene. This gene regulates the transportation across the inner mitochondrial membrane. It is possible that inhibiting VEGFR-2 downregulates this process and promotes build-up of fatty acids, including stearic acid, outside the membrane.

When viewed together, these results demonstrate effects of I3A on human cells. I3A is able to dock to VEGFR-2 and inhibit ATP binding through a competing effect. Molecular

dynamics of VEGFR-2 revealed the stability of the backbone structure and set up a framework for further simulations involving VEGFR-2, I3A and ATP. The docking of I3A to VEGFR-2 and subsequent inhibition means that downstream signalling and metabolic pathways are affected. Using experimental data, I showed that inhibition of VEGFR-2 downregulates signalling pathways related to mitochondrial fatty acid transportation, and NO and AKT activity leading to reduced putrescine production.

## Future work

This project has brought to light new areas of interest that can be explored further. These include more complex simulations of I3A and VEGFR-2. By solvating VEGFR-2 in a solution of I3A, it might be possible to investigate the possibility of allosteric binding sites on VEGFR-2. The statistical distribution of I3A molecules around the protein would reveal areas of high affinity that could be potential binding sites. A binding in one of these hypothetical areas could lead to a conformational change in the protein, which would be demonstrated by the molecular dynamics simulation. The molecular dynamics of VEGFR-2 can be expanded to include both I3A and ATP simultaneously and therefore model the competitive behaviour of the two molecules.

An alternative next step would be to further analyse I3A as a drug candidate. This would involve a more comprehensive analysis of the toxicity and efficacy of the compound; as well as investigating the prospects of preclinical research. A complementary long term project could be to investigate alternative modes of drug delivery to specifically target a tumour. This enables the drug to be delivered at a higher dosage owing to the fact that the molecule will not bind to secondary targets that could cause negative side effects. Potential platforms include perfluorocarbon microbubbles (Choi *et al.* 2011) and genetically engineered bacteria that release I3A only when in a cancerous environment.

This concept is supported by the fact that bacteria that produce I3A have already been identified (Tsavkelova *et al.* 2007). However, these alternate delivery avenues might prove redundant in the case of I3A as no secondary targets have been found through screening by Dumas' group.

## Conclusion

My research hypothesis was that I3A would inhibit ATP docking and affect downstream pathways. My results show that indole-3-acetate docks to VEGFR-2 in the ATP binding cavity. This event inhibits an ATP molecule from binding in its normal way and affects signalling and metabolic pathways downstream of VEGFR-2. VEGFR-2 inhibition is a classic mechanism for angiogenesis impairment. Consistent with its role as a VEGFR-2 inhibitor, I3A inhibited angiogenesis *in vitro* in preliminary results from our lab. My network analysis reveals that I3A-induced inhibition is able to impact metabolic and signalling pathways downstream of VEGFR-2. Two pathways are particularly affected, involving putrescine and stearic acid. Together, these results demonstrate that I3A can modulate both angiogenesis and metabolism, highlighting its therapeutic role, particularly for the treatment of cancer.

# Acknowledgements

I would like to thank Dr Marc Dumas and Professor Robert Glen for their guidance throughout the project, and for introducing me to the many disciplines that were required for this project.

Special thanks must be given to Ana Luisa Neves for her willingness to put aside her own work to discuss ideas and support this project, and for extensively sharing her experimental results.

I would like to recognise Andrea Rodriguez-Martinez for her assistance with metabolic and signalling networks. She showed me her MetaboSignal package, and helped me use it with my network.

I thank Dr Mark Williamson, at the University of Cambridge, for his discussions on molecular dynamics and for giving me permission to use both his python scripts and the server which I used to run the simulations.

Leen Kalash and Dr Nitin Sharma, both at the University of Cambridge, must be acknowledged for their help in molecular dynamics and docking.

# References


Bold, G., Schnell, C., Furet, P., McSheehy, P., Brüggen, J., Mestan, J., Manley, P.W., Drückes, P., Burglin, M., Dürler, U. and Loretan, J., 2015. A Novel Potent Oral Series of VEGFR2 Inhibitors Abrogate Tumor Growth by Inhibiting Angiogenesis. *Journal of medicinal chemistry*, *59*(1), pp.132-146.

Cani, P.D., Bibiloni, R., Knauf, C., Waget, A., Neyrinck, A.M., Delzenne, N.M. and Burcelin, R., 2008. Changes in gut microbiota control metabolic endotoxemia-induced inflammation in high-fat diet–induced obesity and diabetes in mice. *Diabetes*, *57*(6), pp.1470-1481.

Carmeliet, P. and Jain, R.K., 2000. Angiogenesis in cancer and other diseases. *Nature*, *407*(6801), pp.249-257.

Chinot, O.L., Wick, W., Mason, W., Henriksson, R., Saran, F., Nishikawa, R., Carpentier, A.F., Hoang-Xuan, K., Kavan, P., Cernea, D. and Brandes, A.A., 2014. Bevacizumab plus radiotherapy–temozolomide for newly diagnosed glioblastoma. *New England Journal of Medicine*, *370*(8), pp.709-722.

Choi, J.J., Selert, K., Vlachos, F., Wong, A. and Konofagou, E.E., 2011. Noninvasive and localized neuronal delivery using short ultrasonic pulses and microbubbles. *Proceedings of the National Academy of Sciences*, *108*(40), pp.16539-16544.

Costello, E.K., Stagaman, K., Dethlefsen, L., Bohannan, B.J. and Relman, D.A., 2012. The application of ecological theory toward an understanding of the human microbiome. Science, 336(6086), pp.1255-1262.

Cox, L.M., Yamanishi, S., Sohn, J., Alekseyenko, A.V., Leung, J.M., Cho, I., Kim, S.G., Li, H., Gao, Z., Mahana, D. and Rodriguez, J.G.Z., 2014. Altering the intestinal microbiota during a critical developmental window has lasting metabolic consequences. *Cell*, *158*(4), pp.705-721.

Davidovic, L., Navratil, V., Bonaccorso, C.M., Catania, M.V., Bardoni, B. and Dumas, M.E., 2011. A metabolomic and systems biology perspective on the brain of the fragile X syndrome mouse model. *Genome research*, *21*(12), pp.2190-2202.

Dias, S., Hattori, K., Heissig, B., Zhu, Z., Wu, Y., Witte, L., Hicklin, D.J., Tateno, M., Bohlen, P., Moore, M.A. and Rafii, S., 2001. Inhibition of both paracrine and autocrine VEGF/VEGFR-2 signaling pathways is essential to induce long-term remission of xenotransplanted human leukemias.*Proceedings of the National Academy of Sciences*, *98*(19), pp.10857-10862.

Eastman, P., Friedrichs, M.S., Chodera, J.D., Radmer, R.J., Bruns, C.M., Ku, J.P., Beauchamp, K.A., Lane, T.J., Wang, L.P., Shukla, D. and Tye, T., 2012. OpenMM 4: a reusable, extensible, hardware independent library for high performance molecular simulation. *Journal of chemical theory and computation*, *9*(1), pp.461-469.

Egan, W.J., Merz, K.M. and Baldwin, J.J., 2000. Prediction of drug absorption using multivariate statistics. *Journal of medicinal chemistry*, *43*(21), pp.3867-3877.

FDA, Highlights of Prescribing information, 2004. ONLINE. http://www.accessdata.fda.gov/drugsatfda_docs/label/2014/125085s305lbl.pdf

Folkman, J., 1995. Angiogenesis in cancer, vascular, rheumatoid and other disease. *Nature medicine*, *1*(1), pp.27-30.

Furukawa, S., Usuda, K., Abe, M. and Ogawa, I., 2005. Effect of indole-3-acetic acid derivatives on neuroepithelium in rat embryos. *The Journal of toxicological sciences*, *30*(3), pp.165-174.

Gleeson, M.P., 2008. Generation of a set of simple, interpretable ADMET rules of thumb. *Journal of medicinal chemistry*, *51*(4), pp.817-834.

Holmes, K., Roberts, O.L., Thomas, A.M. and Cross, M.J., 2007. Vascular endothelial growth factor receptor-2: structure, function, intracellular signalling and therapeutic inhibition. *Cellular signalling*, *19*(10), pp.2003-2012.

Jensen, M.T., Cox, R.P. and Jensen, B.B., 1995. 3-Methylindole (skatole) and indole production by mixed populations of pig fecal bacteria. *Applied and Environmental Microbiology*, *61*(8), pp.3180-3184.

Jia, W., Li, H., Zhao, L. and Nicholson, J.K., 2008. Gut microbiota: a potential new territory for drug targeting. *Nature Reviews Drug Discovery*, *7*(2), pp.123-129.

Jin, U.H., Lee, S.O., Sridharan, G., Lee, K., Davidson, L.A., Jayaraman, A., Chapkin, R.S., Alaniz, R. and Safe, S., 2014. Microbiome-derived tryptophan metabolites and their aryl hydrocarbon receptor-dependent agonist and antagonist activities. *Molecular pharmacology*, *85*(5), pp.777-788.

Johnson, T.W., Dress, K.R. and Edwards, M., 2009. Using the Golden Triangle to optimize clearance and oral absorption. *Bioorganic & medicinal chemistry letters*, *19*(19), pp.5560-5564.

Kanani, H., Chrysanthopoulos, P.K. and Klapa, M.I., 2008. Standardizing GC–MS metabolomics. *Journal of Chromatography B*, *871*(2), pp.191-201.

Karplus, M. and Kuriyan, J., 2005. Molecular dynamics and protein function.*Proceedings of the National Academy of Sciences of the United States of America*, *102*(19), pp.6679-6685.

Korb, Oliver, Thomas Stutzle, and Thomas E. Exner. "Empirical scoring functions for advanced protein− ligand docking with PLANTS." *Journal of chemical information and modeling* 49.1 (2009): 84-96.

Kubinyi, H., 1998. Combinatorial and computational approaches in structure-based drug design. *Current Opinion in Drug Discovery and Development*,*1*(1), pp.16-27.

Lagorce, D., Sperandio, O., Baell, J.B., Miteva, M.A. and Villoutreix, B.O., 2015. FAF-Drugs3: a web server for compound property calculation and chemical library design. *Nucleic acids research*, *43*(W1), pp.W200-W207.

Lathrop, S.K., Bloom, S.M., Rao, S.M., Nutsch, K., Lio, C.W., Santacruz, N., Peterson, D.A., Stappenbeck, T.S. and Hsieh, C.S., 2011. Peripheral education of the immune system by colonic commensal microbiota. *Nature*,*478*(7368), pp.250-254.

Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J., 2012. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, *64*, pp.4-17.

Lobell, M., Hendrix, M., Hinzen, B., Keldenich, J., Meier, H., Schmeck, C., Schohe-Loop, R., Wunberg, T. and Hillisch, A., 2006. In silico ADMET traffic lights as a tool for the prioritization of HTS hits. *ChemMedChem*, *1*(11), pp.1229-1236.

Lovering, F., Bikker, J. and Humblet, C., 2009. Escape from flatland: increasing saturation as an approach to improving clinical success. *Journal of medicinal chemistry*, *52*(21), pp.6752-6756.

Rodriguez Martinez, A. MetaboSignal a network based approach to overlay metabolic and signalling pathways to study metabolic profiles. *Submitted to Bioinformatics,* 2016*.*

National Cancer Institute. 2016. ONLINE. Available at: http://www.cancer.gov/about-cancer/treatment/drugs/fda-bevacizumab

Neves, A.L., Chilloux, J., Sarafian, M.H., Rahim, M.B.A., Boulange, C.L. and Dumas, M.E., 2015. The microbiome and its pharmacological targets: therapeutic avenues in cardiometabolic diseases. *Current opinion in pharmacology*, *25*, pp.36-44.

Plate, K.H., Breier, G., Millauer, B., Ullrich, A. and Risau, W., 1993. Up-regulation of vascular endothelial growth factor and its cognate receptors in a rat glioma model of tumor angiogenesis. *Cancer Research*, *53*(23), pp.5822-5827.

Salomon-Ferrer, R., Götz, A.W., Poole, D., Le Grand, S. and Walker, R.C., 2013. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *Journal of chemical theory and computation*, *9*(9), pp.3878-3888.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks.*Genome research*, *13*(11), pp.2498-2504.

Simon, S. and Petrášek, J., 2011. Why plants need more than one type of auxin. *Plant Science*, *180*(3), pp.454-460.

Sokol, H., Pigneur, B., Watterlot, L., Lakhdari, O., Bermúdez-Humarán, L.G., Gratadoux, J.J., Blugeon, S., Bridonneau, C., Furet, J.P., Corthier, G. and Grangette, C., 2008. Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proceedings of the National Academy of Sciences*, *105*(43), pp.16731-16736.

Suez, J., Korem, T., Zeevi, D., Zilberman-Schapira, G., Thaiss, C.A., Maza, O., Israeli, D., Zmora, N., Gilad, S., Weinberger, A. and Kuperman, Y., 2014. Artificial sweeteners induce glucose intolerance by altering the gut microbiota. *Nature*, *514*(7521), pp.181-186.

Swain, R.A., Harris, A.B., Wiener, E.C., Dutka, M.V., Morris, H.D., Theien, B.E., Konda, S., Engberg, K., Lauterbur, P.C. and Greenough, W.T., 2003. Prolonged exercise induces angiogenesis and increases cerebral blood volume in primary motor cortex of the rat. *Neuroscience*, *117*(4), pp.1037-1046.

Tsavkelova, E.A., Cherdyntseva, T.A., Klimova, S.Y., Shestakov, A.I., Botina, S.G. and Netrusov, A.I., 2007. Orchid-associated bacteria produce indole-3-acetic acid, promote seed germination, and increase their microbial yield in response to exogenous auxin. *Archives of Microbiology*, *188*(6), pp.655-664.

Veber, D.F., Johnson, S.R., Cheng, H.Y., Smith, B.R., Ward, K.W. and Kopple, K.D., 2002. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of medicinal chemistry*, *45*(12), pp.2615-2623.

Venkatesh, M., Mukherjee, S., Wang, H., Li, H., Sun, K., Benechet, A.P., Qiu, Z., Maher, L., Redinbo, M.R., Phillips, R.S., Fleet, J.C. and Dumas, M.E., 2014. Symbiotic bacterial metabolites regulate gastrointestinal barrier function via the xenobiotic sensor PXR and Toll-like receptor 4. *Immunity*, *41*(2), pp.296-310.

Wen, L., Ley, R.E., Volchkov, P.Y., Stranges, P.B., Avanesyan, L., Stonebraker, A.C., Hu, C., Wong, F.S., Szot, G.L., Bluestone, J.A. and Gordon, J.I., 2008. Innate immunity and intestinal microbiota in the development of Type 1 diabetes. *Nature*, *455*(7216), pp.1109-1113.

# Supplementary information

## Docking solutions comparison

**Table 1.** RSMD comparison of I3A solutions.

| Solution number | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.0 | 7.3 | 7.4 | 7.4 | 5.1 | 2.3 | 2.3 | 4.7 | 5.1 |
| 2 | | 7.3 | 7.4 | 7.5 | 5.2 | 2.3 | 2.3 | 4.7 | 5.1 |
| 3 | | | 0.1 | 0.4 | 4.5 | 6.5 | 6.4 | 7.2 | 4.5 |
| 4 | | | | 0.4 | 4.5 | 6.6 | 6.5 | 7.2 | 4.6 |
| 5 | | | | | 4.6 | 6.6 | 6.5 | 7.2 | 4.7 |
| 6 | | | | | | 4 | 3.8 | 7.1 | 0.1 |
| 7 | | | | | | | 0.1 | 5 | 3.9 |
| 8 | | | | | | | | 5 | 3.8 |
| 9 | | | | | | | | | 7.1 |

## Clustering of I3A docking solutions

Clustering of docking solutions by RSMD distance in Å. Clusters are separated by vertical lines and the distance is in the leftmost column.

Distance | Clusters

0.04 | 1 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

0.07 | 1 2 | 3 | 4 | 5 | 6 10 | 7 | 8 | 9 |

0.14 | 1 2 | 3 4 | 5 | 6 10 | 7 | 8 | 9 |

0.15 | 1 2 | 3 4 | 5 | 6 10 | 7 8 | 9 |

0.37 | 1 2 | 3 4 5 | 6 10 | 7 8 | 9 |

2.32 | 1 2 7 8 | 3 4 5 | 6 10 | 9 |

4.65 | 1 2 7 8 | 3 4 5 6 10 | 9 |

5.04 | 1 2 7 8 9 | 3 4 5 6 10 |

7.45 | 1 2 3 4 5 6 7 8 9 10 |

## PCA pre-processing

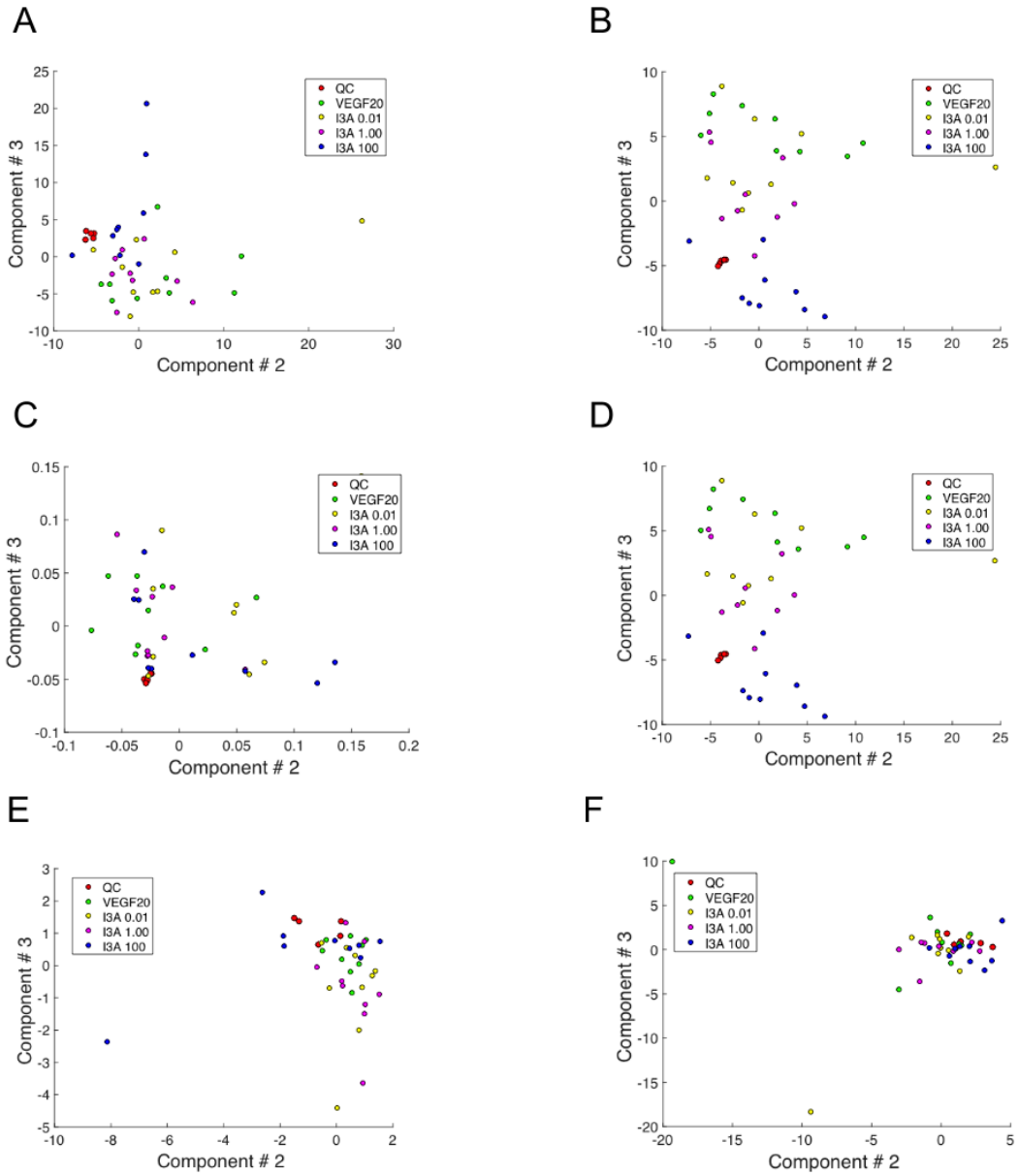



**Figure 6.** Comparison of different GC-MS pre-processing techniques attempted. Filtering by CV<15% provided the best separation **A)** Log10 transformed and filtered by coefficient of variation of QCs < 30% **B)** Log10 transformed and filtered by coefficient of variation of QCs < 15%. This performed better than the 30% and was chosen **C)** Log10 transformed, filtered by CV<15% and probabilistic quotient normalisation applied **D)** Log10 transform, filtered by CV<15% and 5% dynamic range of each feature added **E)** 5% dynamic range of each feature added then Log10 transformed and filtered by CV<15% **F)** Log10 transform, filtered by CV<15%, 5% dynamic range of each feature added and centered and scaled

**Table 2.** Attributes of the OPLS-DA model of GC-MS data.

| Component number | $Q^2$ | $R^2$ |
|---|---|---|
| 1 | 0.2593 | 0.7368 |
| 2 | 0.1414 | 0.9331 |
| 3 | 0.2546 | 0.9817 |
| 4 | 0.2424 | 0.9969 |
| 5 | 0.2328 | 0.9993 |
| 6 | 0.2341 | 0.9998 |