



# **DISCOVERY OF GAMMA SECRETASE INHIBITORS FOR BREAST CANCER THERAPY THROUGH CHEMOGENOMIC METHODS**

**By**

**Ngceboyakwethu P. Zinyama  
(R0537389)**

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy of Chemistry in the department of Chemical Sciences, Faculty of Science and Technology at Midlands State University, Zimbabwe

Supervisors:

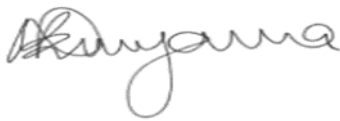
Professor Upenyu Guyo  
Professor Grace Mugumbate

**October 2023**

## Declaration

I **Ngceboyakwethu Primrose Zinyama** hereby declare that this thesis submitted for the degree of Doctor of Philosophy at the Midlands State University, Gweru, Zimbabwe, entitled, "*Discovery of gamma-secretase inhibitors for breast cancer therapy through chemogenomic methods*" is my own and it has not been previously submitted to this or any other institution for examination; secondary sources used in this thesis was duly acknowledged.

**Ngceboyakwethu Primrose Zinyama (R0537389)**




Signature

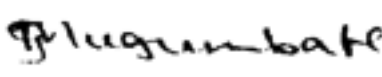
Date 18/10/23

## Supervisors

**Prof. Upenyu Guyo**

Signature .....  ..... Date.....19/10/2023.....

**Prof. Grace Mugumbate**

Signature ...  ...Date.....19/10/23

## **Dedication**

To my loving husband, Kevin, and children, Kebo and Kevin N.

## **Acknowledgements**

I am deeply indebted to the following:

My supervisors, Professors Grace Mugumbate and Upenyu Guyo, for their insight, guidance, patience and support.

The Department of Chemical Sciences for their encouragement and provision of resources I needed for research.

Professor Fanuel Lampiao for the bioassays.

The Drug Discovery and Informatics Research Group for the motivation and helpful discussions.

My Parents, Canaan and Sindiso Masuku and Thabo for spurring me on and providing the shoulder to lean on.

My husband Kevin and children for the love, and support.

I am grateful to the Midlands State University through the Research and Innovation Division, for funding this work.

## **Abstract**

Breast cancer recurrence is often treated with hormonal and targeted chemotherapy. To this end, nicastrin, a protein involved in Notch signaling, has been associated with breast cancer recurrence. In this work, binding sites in nicastrin were identified. Binding interactions, and modes, of known nicastrin inhibitors were investigated using structure-based techniques. A binding site, termed the DYIGS binding site, (named after the conserved hydrophilic residues Asp336, Tyr337, Iso338, Gly339, and Ser340 found in the site) was identified. The binding mechanisms, and interactions, of known nicastrin inhibitors were investigated in the identified binding sites. This was done using binding free energy calculations, and per residue decomposition analysis. Residues such as Val138, Gln139, Asp143, Arg105 and Glu174 were discovered to be important in the interactions. The physicochemical properties and scaffold space of nicastrin inhibitors were investigated. Scaffold analysis and machine learning models identified specific connectivity containing a sulfon, sulfonamide, or sulfonamide connected to cyclic structures; and a halide or a halide connected to a benzene ring as being associated with high activity for nicastrin inhibition. Seven nicastrin inhibitors were discovered using this information. A preliminary antitumour bioassay confirmed the activity of six of the seven compounds, which inhibited tumour growth by more than 20%. However, three of these compounds demonstrated acceptable physicochemical and pharmacokinetic properties. The identification of these nicastrin actives opens new avenues for the development of breast cancer treatments.

## **Keywords**

Breast cancer; Gamma-secretase; Nicastrin; Chemogenomic; Chemical space;  
Machine learning; Docking; Molecular dynamics.

# Table of Contents

## Contents

<b>Declaration</b> .....	i
<b>Dedication</b> .....	ii
<b>Acknowledgements</b> .....	iii
<b>Abstract</b> .....	iv
<b>Table of Contents</b> .....	vi
<b>Acronyms</b> .....	x
<b>Research Outputs</b> .....	xii
<b>Publications</b> .....	xii
<b>List of Figures</b> .....	xiii
<b>List of Tables</b> .....	xiv
1. Breast Cancer and Nicastrin.....	1
1.1 Introduction.....	1
1.2 Background.....	1
1.3 Genesis of breast cancer.....	4
1.4 Treatment of breast cancer.....	5
1.5 The Gamma-secretase complex.....	6
1.6 Gamma-secretase inhibitors against breast cancer.....	8
1.7 Nicastrin subunit.....	10
1.8 Advances in the development of inhibitors towards nicastrin.....	12
1.9 Problem statement.....	13
1.10 Justification of the study.....	14
1.11 Aims and objectives.....	15
1.11.1 Aims.....	15
1.11.2 Objectives.....	15
1.12 References.....	16
2. Literature review: Experimental Methods and Techniques.....	21
2.1 Introduction.....	21
2.2 Binding site identification.....	21
2.3 Binding site identification methods.....	22
2.3.1 Energy based methods for binding site identification.....	23
2.4 Binding site identification methods used for nicastrin.....	27
2.4.1 AutoDock Vina.....	27

2.4.2	Internal Coordinate Mechanism .....	29
2.4.3	Druggability analysis .....	30
2.4.4	Molecular Dynamic Simulations .....	31
2.5	Ligand based methods .....	34
2.5.1	Quantitative Structure Activity Relationships.....	35
2.5.2	Dataset collection and curation for Machine Learning.....	35
2.5.3	Descriptor calculation .....	38
2.5.4	Removal of redundant and correlating descriptors.....	40
2.5.5	Descriptor subset generation .....	40
2.5.6	Other Machine learning Algorithms .....	43
2.5.7	Model Validation.....	46
2.5.8	Applicability domain .....	48
2.6	Virtual Screening.....	48
2.6.1	Database filters .....	49
2.6.2	Database screening.....	50
2.7	Pharmacokinetic analysis of orally available drugs.....	50
2.7.1	Barriers to drug delivery .....	50
2.8	Antitumor Tests .....	52
2.9	References .....	54
3.	Analysis of binding sites and ligand induced conformation of nicastrin and binding modes of its inhibitors.....	65
3.1	Introduction .....	65
3.2	Methods.....	65
3.2.1	Protein preparation.....	65
3.2.2	Ligand collection and curation.....	66
3.2.3	Ligand preparation for docking.....	66
3.2.4	Blind docking calculations.....	68
3.2.5	Assessing druggability of the identified binding sites in nicastrin .....	68
3.2.6	Molecular dynamic simulations to optimize the docked protein-ligand complexes....	69
3.2.7	Calculating the free energy of binding.....	69
3.2.8	Characterizing binding modes and interactions of known inhibitors in nicastrin .....	70
3.3	Results and Discussion .....	70
3.3.1	Binding sites in nicastrin .....	70
3.3.2	Binding site characterization and druggability assessment.....	73
3.3.3	Mechanism of nicastrin ligand binding .....	77



3.3.4	Binding free energy calculations and per residue free energy decomposition analysis of the complex .....	80
3.3.5	Characterization of binding modes and interactions of known inhibitors .....	82
3.4	Conclusion.....	87
3.5	References .....	89
4.	Chemical Space Analysis and Virtual screening for Nicastrin Inhibitors.....	92
4.1	Introduction .....	92
4.2	Methods.....	92
4.2.1	Preparation of compound datasets .....	92
4.2.2	Navigating the chemical space.....	93
4.2.3	Quantitative Structural Activity Relationship (QSAR) prediction using Machine Learning approaches.....	96
4.2.4	Ligand based virtual screening using generated QSAR models.....	97
4.2.5	Applicability Domain .....	97
4.2.6	Structure based virtual screening of the Maybridge screening set.....	98
4.2.7	Diversity selection.....	99
4.3	Results and discussion .....	99
4.3.1	Physicochemical property space.....	99
4.4	Profiling of drug-likeness .....	105
4.4.1	Scaffold chemical analysis.....	106
4.4.2	Structure activity relationships and Activity cliff analysis of gamma-secretase inhibitors 109	
4.4.3	Quantitative Structural Activity Relationship (QSAR) prediction using Machine Learning 112	
4.4.4	Interpretation of J48 and NB models.....	117
4.4.5	Structure based virtual screening of the Maybridge dataset .....	120
4.4.6	Interactions of identified hits in the binding site.....	121
4.5	Conclusion.....	130
4.6	References .....	132
5.	Biological evaluation, and physicochemical and pharmacokinetic property profiling of hit compounds .....	135
5.1	Introduction .....	135
5.2	Methods.....	136
5.2.1	Culturing of the Agrobacteria tumefaciens .....	136
5.2.2	Test for antibiotic resistance of the hits .....	136
5.2.3	Carrot disc assay .....	137
5.2.4	Assessment of oral availability and pharmacokinetic property evaluation of hits.....	138

5.3	Results and discussion .....	139
5.3.1	Biological evaluation of antitumor properties.....	139
5.3.2	Physicochemical property prediction .....	144
5.3.3	Prediction of Absorption, Distribution, Metabolism, Excretion and Toxicity .....	145
5.4	Conclusion.....	150
5.5	References .....	151
6.	Summary and Conclusions.....	154
6.1	Introduction .....	154
6.1.1	Nicastrin binding sites and binding modes and interactions of known inhibitors.....	154
6.1.2	Nicastrin hits identified from virtual screening .....	155
6.2	Conclusion.....	157
6.3	Future Work.....	158
APPENDIX	.....	159

## Acronyms

### Amino acid residues

Amino acid	Three letter code	One letter code
Alanine	Ala	A
Aspartic acid	Asp	D
Asparagine	Asn	N
Arginine	Arg	R
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S



## Research Outputs

### Publications

1. **Zinyama NP**, Guyo U, Mugumbate G, Identification of binding sites in nicastrin and binding modes of its inhibitors, *F1000Research* 2023, 12:150, DOI 10.12688/f1000research.130518.1.

### Manuscripts under review

1. Jonathan T. Bvunzawabaya, **Ngceboyakwethu P. Zinyama**, Brilliant Nyathi, Upenyu Guyo, Fanuel Lampiao and Grace Mugumbate, *Exploring the chemical diversity of gamma secretase inhibitors using artificial intelligence and cheminformatics for breast cancer drug discovery*.
2. **Zinyama NP**, Guyo U, Mugumbate G, *Machine learning modelling, virtual screening and biological evaluation of nicastrin hits for breast cancer therapy*.

### Skills

1. Participation in the hands-on international workshop on Bridging data science and AI/ML tools to infectious disease research at SunSquare Cape Town City Bowl, Cape Town, South Africa from 27-30 September 2022.
2. Participation in the Hands-on training workshop in Computational Chemistry at the Zimbabwe Centre High Performance Computing from 18-22 July 2016.

## List of Figures

Figure 1.1 2020 Global breast cancer age standardised.....	3
Figure 1.2 Cartoon representation of the gamma-secretase complex.....	8
Figure 1.3 Gamma-secretase inhibitors used in breast cancer therapy.....	9
Figure 1.4 The structure of nicastrin.....	11
Figure 2.1 The docking process.....	25
Figure 2.2 The Lennard-Jones potential.....	27
Figure 2.3 The Wrapper subset evaluation illustration .....	44
Figure 2.4 Illustration of part of a decision tree. ....	46
Figure 2.5 Confusion Matrix.....	47
Figure 3.1 A set of 30 known gamma secretase inhibitors.....	76
Figure 3.2 Predicted binding sites in nicastrin. ....	83
Figure 3.3 Correlation between druggability of binding sites in nicastrin and hydrophobicity .....	86
Figure 3.4 Surface of nicastrin (yellow) with compound CID44433923 .....	87
Figure 3.5 Change in carbon alpha RMSD of nicastrin, ligand CID44433923 and complex.....	88
Figure 3.6 Change in RMSF for the carbon alpha residues .....	89
Figure 3.7 Radius of gyration of nicastrin carbon alpha residues of nicastrin.....	90
Figure 3.8 Binding mode of CID44433923 .....	92
Figure 3.9 Maximum common sub-structures.....	93
Figure 3.10 Docked pose of compound CID 15953832 in the DYIGS binding site. ....	97
Figure 4.1 Physicochemical property distribution of FDA approved breast cancer drugs .....	114
Figure 4.2 Property-based chemical space of FDA approved breast cancer drugs .....	115
Figure 4.3 Drug-like, extended-drug like, lead-like, fragment-like, PPI like profiles .....	117
Figure 4.4 The most common scaffolds .....	119
Figure 4.5 A scatter plot of activity as pIC50 against core fragment.....	121
Figure 4.6 A structure-activity landscape index (SALI) plot of nicastrin inhibitors.....	123
Figure 4.7 Representation of the J48 Wrapper model. ....	131
Figure 4.8 The DYIGS site in nicastrin. ....	133
Figure 4.9 Nicastrin inhibitors selected for experimental validation .....	135
Figure 4.10 Molecular docking 2D mode of interaction of compound 3255.....	136
Figure 4.11 Molecular docking 2D mode of interaction of compound 8361.....	137
Figure 4.12 Molecular docking 2D mode of interaction of compound 8796.....	138
Figure 4.13 Molecular docking 2D mode of interaction of compound 6218.....	139

Figure 4.14 Molecular docking 2D mode of interaction of compound 6197.....	140
Figure 4.15 Molecular docking 2D mode of interaction of compound 2251.....	141
Figure 4.16 Molecular docking 2D mode of interaction of compound 8250.....	142
Figure 5.1 Hit compounds with antibacterial susceptibility .....	150
Figure 5.2 The three active hits .....	151
Figure 5.3 Bioavailability radar plots of hit compounds .....	155
Figure 5.4 The BOILED-Egg model of the selected hit compounds. ....	158

## List of Tables

Table 2.1 Validation Parameters.....	48
Table 3.1 Predicted binding site residues .....	80
Table 3.2 Druggability assessment of nicastrin binding sites in two different conformers .....	85
Table 3.3 Energy contributions .....	91
Table 4.1. Molecular properties and conditions used to categorise compounds in datasets for drug discovery .....	105
Table 4.2 Eigen values and explained variance of Principal components .....	115
Table 4.3. Subsets of descriptors selected for the development of the different Machine Learning models.....	125
Table 4.4. Summary of Model performance.....	128
Table 4.5. Important binding site residues in nicastrin DYIGS binding site .....	132
Table 5.1 Resistance to antibiotics and percentage inhibition of <i>A. tumefaciens</i> by the identified hits .....	152
Table 6.1 Physicochemical, pharmacokinetic and Antitumour properties of the hits .....	166
Table A 1 Absorption.....	171
Table A 2 Distribution .....	171
Table A 3 Metabolism .....	172
Table A 4 Excretion .....	173

# **1 Breast Cancer and Nicastrin**

## **1.1 Introduction**

A meticulous treatment plan based on parameters such as tumour subtype, cancer stage, genetic markers, and the patient's overall health, determines the outcome of breast cancer treatment. However, the existence of signaling pathways such as Notch, which promote the proliferation of breast cancer stem cells, leads to resistance and recurrence even after an initial effective treatment. Furthermore, due to a lack of screening techniques, and limited access to diagnostic facilities, late-stage cancers that are aggressive, and resistant to treatment are common. Most breast cancer patients in developing countries are in this situation.

This chapter gives the study's background and emphasises the significance of developing compounds for estrogen-resistant breast cancer. The gamma-secretase enzyme complex, particularly the nicastrin subunit, is linked to estrogen-resistant breast cancer. The gap that remains, and serves, as the study's foundation was identified after reviewing studies that are related to the development of therapies for nicastrin.

## **1.2 Background**

Breast cancer is one of the leading causes of death in women worldwide. Global incidence and mortality data highlight the devastation caused by breast cancer. According to GLOBOCAN 2018,<sup>[1]</sup> breast cancer is the most commonly diagnosed cancer in women accounting for 11.6% of all cases in women, with nearly 60% of these

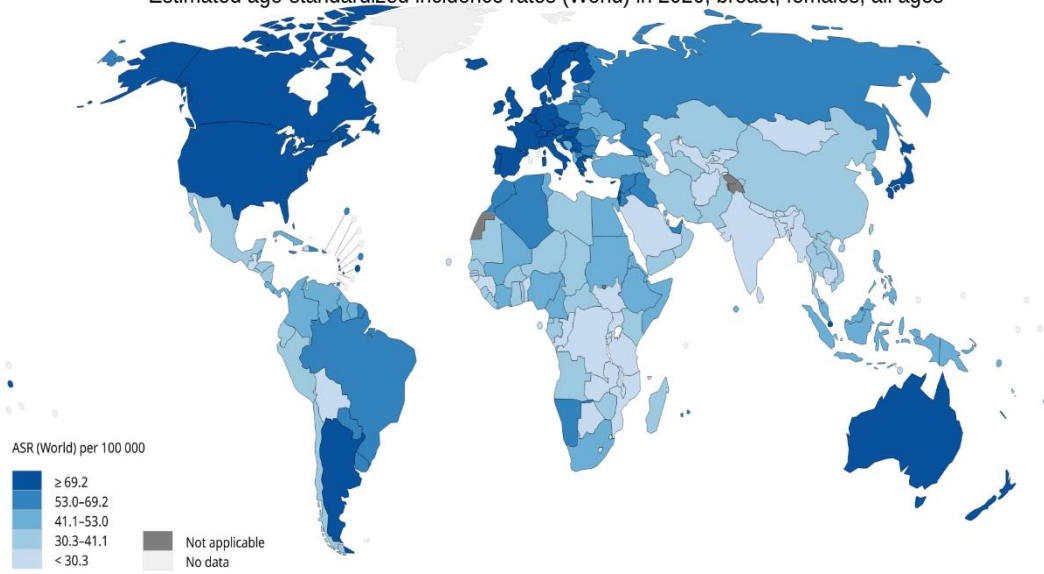


women coming from low-income countries.<sup>[2, 3, 4]</sup> Despite the low incidence rates in sub-Saharan Africa, particularly Zimbabwe (39.2 per 100 000) (**Figure 1.1a**), the region has higher fatality rates, with Zimbabwe having 20.2 per 100 000 (**Figure 1.1b**) due to diagnostic difficulties, late presentation, and unaffordable care. <sup>[1, 5]</sup>

---

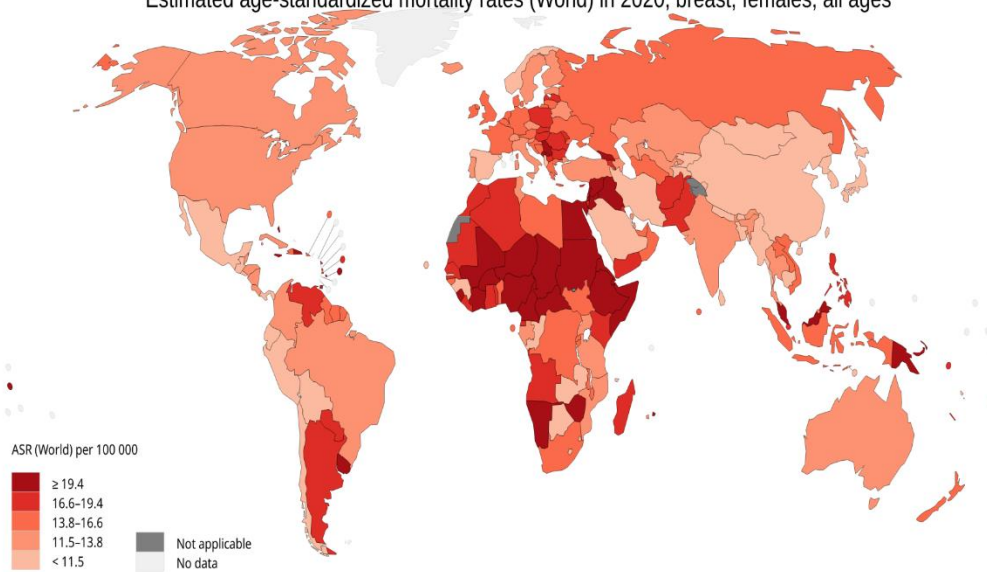
A.

Estimated age-standardized incidence rates (World) in 2020, breast, females, all ages



B.

Estimated age-standardized mortality rates (World) in 2020, breast, females, all ages



**Figure 1.1** 2020 Global breast cancer age standardised.

**A.** incidence rates and **B.** mortality rates. <sup>[1]</sup> *Data source:* GLOBOCAN 2020 at <http://gco.iarc.fr/today>.

---

Breast cancer affects women of all ages, not just postmenopausal women, and this is a significant source of concern. <sup>[2]</sup> According to the Global Cancer Registry, approximately 104 females, aged 25 to 44, are diagnosed with breast cancer in Zimbabwe annually. <sup>[1]</sup> Tumours that appear at such a young age are aggressive and have a terrible prognosis. <sup>[5,6]</sup> Unfortunately, because of the failure to detect hormone and gene status, most patients have a low survival making tailored treatment difficult. <sup>[7]</sup> An overview of the progression of breast cancer is discussed in the following section in order to account for the therapeutic alternatives that can be established in such circumstances.

### **1.3 Genesis of breast cancer**

Breast cancer originates from the epithelial cell lining milk ducts. <sup>[8]</sup> Although the underlying aetiology of breast cancer is unknown, accumulating evidence suggests that mammary stem cells in the breast proliferate and differentiate to give rise to all other epithelial cell types inside the breast. These stem cells become tumour targets due to their ability to specialise and self-renew. <sup>[9,10]</sup> Breast tumours are extremely diverse due to the highly dynamic cellular development of these stem cells. <sup>[11]</sup> This encompasses a variety of histopathological, biochemical, and clinical characteristics that exhibit varying responses to treatment approaches, resulting in resistance and recurrence. <sup>[12]</sup>

Breast cancer recurrence and proliferation are induced when signaling pathways necessary for normal stem cell self-renewal, proliferation, and differentiation are disrupted, resulting in the formation of cancer stem cells (CSC). CSCs are rare or minority cell populations in malignancies <sup>[9,13]</sup> that are capable of differentiating

progeny and play a role in cancer initiation, development, metastasis, recurrence, and drug resistance. [9,14]

#### **1.4 Treatment of breast cancer**

Breast tumours have been classified into molecular subtypes using hormone receptor types such as estrogen receptor alpha (ER $\alpha$ ), triple-negative breast cancer (TNBC), and human epidermal growth factor receptor amplified (HER2<sup>+</sup>). These molecular subtypes as well as the patient's pathological traits, and tumour stage, determine the therapeutic options. [15] For example, the ER $\alpha$  sub-type accounts for 75% of breast cancer cases, and standard therapies restricts the function of the ER $\alpha$ . Tamoxifen, a common endocrine treatment, has a high success rate in blocking E2/ER $\alpha$  interactions. [16]

However, tamoxifen's efficacy is obscured by 40-50% resistance due to other signaling pathways that are not dependent on the E2/ER $\alpha$  interaction that fuels the growth of breast cancer cells. [17] Between 10–30% of breast cancer patients with endocrine resistant cancer are at high risk of developing brain metastases. [18–20] To this effect, therapeutic techniques that target breast cancer stem cells as well as the tumour microenvironment to improve the performance of traditional medications like tamoxifen are being developed to greatly improve breast cancer treatment. Signaling pathways have an impact on tumour microenvironments. [21]

Breast cancer recurrence and progression have been linked to signaling pathways involved in stem cell survival and development, as well as cell homeostasis. Several studies have discovered that specific breast cancer subtypes have increased expression of these signaling pathways, which include notch, hedgehog, and Wnt. [9,13]

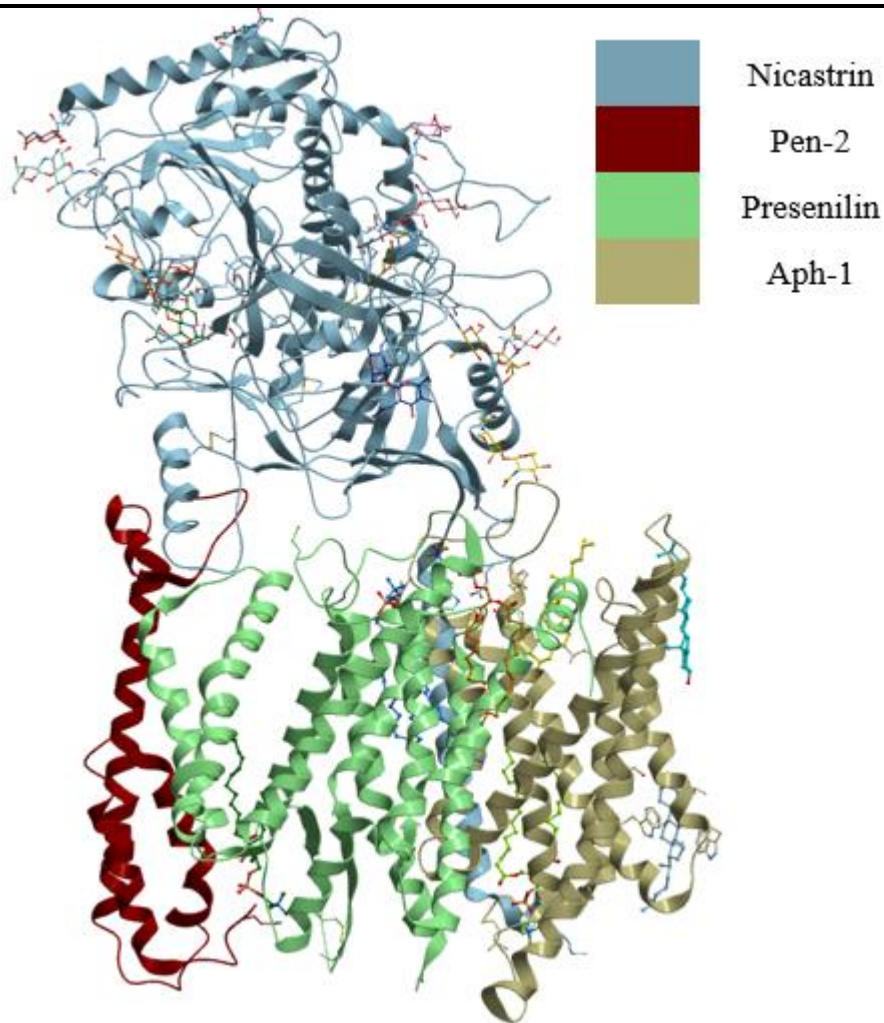
The notch pathway is particularly important because it interacts with other pathways. Physiologically, notch signaling is involved in embryonic development and appropriate mammary development. [22] The notch is also involved in cell homeostasis processes such as stemness, differentiation, and death [22, 23]. Pathologically, on the other hand, it plays a role in tumorigenesis through a variety of mechanisms that promote a wide range of cancer phenotypes. [10, 24, 25] Changes in notch signaling promote the survival, renewal, and differentiation of cancer stem cells. For example, its interaction with the estrogen pathway promotes resistance to standard anti-estrogen therapies. [26,27] Modulation of the notch pathway is critical for the success of these anti-estrogen therapies.

Several proteases, including the gamma-secretase complex, are required for notch signaling. [28] Notch signaling can be modulated by targeting receptors involved in notch proteolysis, such as gamma-secretase complex subunits. [23] The gamma-secretase complex processes notch to produce Notch Intracellular Domain (NICD), which when translocated to the nucleus [29] activates either homeostasis processes or altered signaling, resulting in stem cell expansion. [30] To develop drugs against the gamma-secretase complex for breast cancer treatment, it is necessary to first understand the architecture of the complex. The gamma-secretase complex is discussed in detail in the sections that follow.

## **1.5 The Gamma-secretase complex**

Structural genomics has provided near-atomic resolution structures of the gamma-secretase complex using cryo-electron microscopy (cryoEM) analysis, and the structure has been detailed as a high molecular weight complex that is minimally composed of four subunits, presenilin, nicastrin, anterior pharynx defective (1) (Aph-

1) and presenilin enhancer (2) (Pen-2) <sup>[31]</sup> as shown in **Figure 1.2**. These units work together to activate the gamma-secretase. Pen-2 binds to presenilin and participates in complex maturation. Presenilin is a nine-pass transmembrane protein that forms the complex's catalytic core. Nicastrin, on the other hand, has a large glycosylated ectodomain with a single transmembrane domain and is involved in substrate recognition and recruitment, whereas Aph-1 has seven transmembrane domains and is involved in complex assembly. <sup>[32,33]</sup> The focus of this research is designing compounds that target nicastrin. The following section discusses notch and breast cancer gamma-secretase inhibitors.



**Figure 1.2** Cartoon representation of the gamma-secretase complex.

The four subunits are displayed: nicastrin, presenilin, aph-1 and pen-2. Glycan residues on nicastrin are displayed in stick representation.

---

## 1.6 Gamma-secretase inhibitors against breast cancer

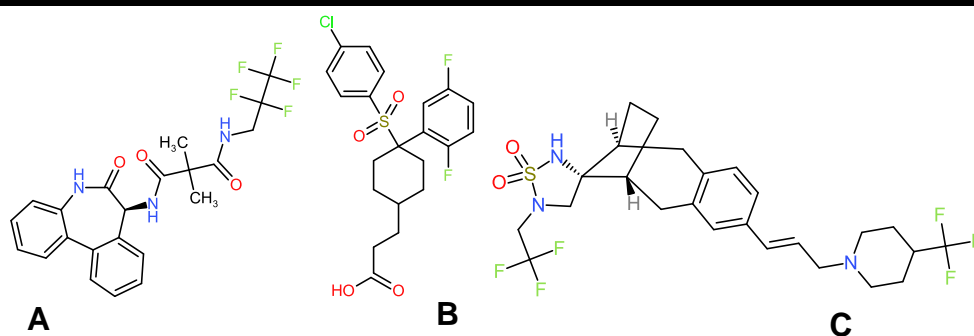
Gamma-secretase inhibitors (GSI) have been extensively studied, and several of them are promising in cancer therapy targeting presenilin, the catalytic core of the gamma-secretase complex. <sup>[34]</sup> The compound, R04929097, a GSI, has been studied in a variety of cancers, including cervical, colon, refractory, ovarian, tracheal, pancreatic, and breast, and has been shown to be effective as a single agent or in combination with other standard cancer therapies in suppressing cancer stem cells and advanced

solid tumours. <sup>[23]</sup> When R04929097 was applied to inflammatory breast cancer cell lines (**Figure 1.3**) notch target genes were downregulated, making the cells more sensitive to ionizing radiation. <sup>[35]</sup> MK-0752 (**Figure 1.3**), a gamma-secretase inhibitor, was also investigated on human breast tumour grafts and found to be effective in reducing breast cancer stem cells when combined with docetaxel treatment. Another gamma-secretase inhibitor, MRK-003 (**Figure 1.3**), was discovered to inhibit tumour recurrence, and when combined with trastuzumab, it induced tumour regression in ErbB-2 positive breast xenographs. <sup>[36]</sup>

Trastuzumab, a standard cancer drug, inhibits ErbB-2 signaling in breast cancer; however, there is over-amplification of the notch-1 receptor, which is activated in response to the inhibitor <sup>[37]</sup> resulting in drug resistance. *In vitro*, the use of the gamma-secretase inhibitor MRK-003 reduced trastuzumab resistance.

The modulation of the notch can be done by targeting the gamma-secretase complex that mediates its proteolysis, however, inhibition of the gamma-secretase complex by targeting the catalytic site has confounding effects due to its functional link to critical signaling processes. Various studies have looked at modulation of the gamma-secretase complex rather than its inhibition to avoid these confounding effects on normal cell processes. Since nicastrin is involved in substrate recognition and recruitment and not in the catalytic events of the complex, targeting nicastrin could modulate the functions of the complex without completely inhibiting it.





**Figure 1.3** Gamma-secretase inhibitors used in breast cancer therapy.

**A** R04929097 has been used in inflammatory breast cancer cell lines. **B** MK-0752 has been assessed on human breast tumour grafts. **C** MRK-003 has been exposed to ErbB-2 positive breast xenografts.

---

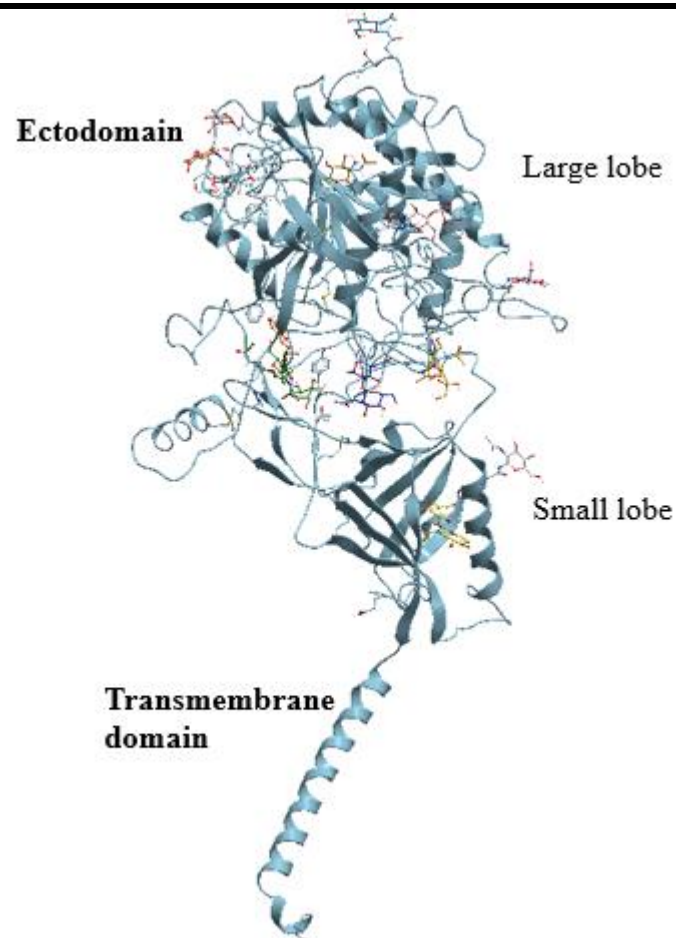
### 1.7 Nicastrin subunit

Nicastrin is composed of a large glycosylated ectodomain and a transmembrane that connects the subunit to the rest of the complex via hydrophobic interactions [38] with the transmembrane base. Nicastrin contains conserved domains that aid in its function. The nicastrin ectodomain is structurally similar to aminopeptidases, particularly the bacterial aminopeptidase (BAP) in that it has a large and small lobe in the ectodomain. The 3D structure of nicastrin is depicted in **Figure 1.4**.

While the active site in BAP is in the large lobe and contains two zinc ions required for protease activity, a similar region in nicastrin is covered by a loop (residues Ser137 to Gly168) that extends from the small to the large lobe and acts as a lid [39]. In nicastrin, the area covered by the lid bears the conserved hydrophilic DYIGS motif (Asp336, Tyr337, Iso338, Gly339, and Ser340), and the residues have been found to be essential in modulating gamma-secretase activity [33, 40, 41] as well as substrate recognition and recruitment.

Another domain found in the large lobe is the tetratricopeptide repeat like domain (TPR-like) which is homologous to the tetratricopeptide repeat 2A (TPR2A) domain, and functions as an interaction module, and a multiprotein complex mediator. [42] According to Zhang et al. [43] the TPR-like domain in nicastrin is involved in substrate recognition and binding, with disruption of its helical structure inhibiting notch processing by the gamma-secretase complex.

---



**Figure 1.4** The structure of nicastrin.

Illustrated with the two domains, the ectodomain and a single transmembrane domain. The Ectodomain contains the large and small lobe.

---

## 1.8 Advances in the development of inhibitors towards nicastrin

The importance of developing nicastrin inhibitors was highlighted in a study by Lombardo et al. <sup>[44]</sup> demonstrating that nicastrin, in conjunction with the notch 4 intracellular domain, and induced resistance of breast cancer cells to Tamoxifen, causes recurrence. As a result of the large ectodomain in nicastrin, advances in identifying nicastrin ligands are focused on the development of monoclonal antibodies. <sup>[45]</sup> Monoclonal antibodies with anti-tumour activity has been developed to modulate nicastrin binding to notch.

Hayashi et al. <sup>[46]</sup> reported the use of a single chain variable fragment (scFv) as an intrabody against nicastrin, which reduced the production of the notch intracellular domain to modulate the production of the notch intracellular domain. This influences breast cancer recurrence. Zhang et al. <sup>[47]</sup> also developed scFVG9, a synthetic antibody fragment that inhibited nicastrin maturation, and prevented the formation of the gamma-secretase complex, thereby affecting notch processing but sparing APP processing.

Filipovic et al. <sup>[48]</sup> developed an anti-nicastrin monoclonal antibody (Anti-NCSTN mAB clone-2H6) that outperformed R04929097, a known small molecule gamma-secretase inhibitor, in terms of anti-tumour efficacy. Similarly, Lombardo et al. <sup>[44]</sup> discovered that an anti-nicastrin monoclonal antibody reduced the population of breast cancer stem cells in endocrine resistant breast cancer cells.

Aside from monoclonal antibodies, Arai et al. <sup>[49]</sup> discovered that cowanin, a natural product, can accelerate the degradation of nicastrin, inhibiting the production of notch intracellular domains. Cowanin reduced nicastrin levels while having no effect on the expression of other gamma-secretase subunits. The discovery of natural products for

nicastrin and the development of monoclonal antibodies for nicastrin demonstrate nicastrin's potential as a target for developing breast cancer drugs. Despite these efforts, there is still a gap in the design of small molecules that specifically target nicastrin. Small molecules that modulate or inhibit notch signaling so far target presenilin, the catalytic centre of the gamma-secretase complex, whereas nicastrin and other components are only thought to be important for the gamma-secretase complex's physiological activity. Small molecules identified specifically for nicastrin are still unavailable.

## **1.9 Problem statement**

Breast cancer is the most common cancer in women worldwide, accounting for 24% of new cancer cases, and 15% of cancer deaths, with less developed countries accounting for 58% of cancer deaths. <sup>[50]</sup> In sub-Saharan Africa, premenopausal women account for a higher proportion of cases, with aggressive tumours with poor prognosis, leading to failure or resistance to standard therapies and high mortality. These aggressive and resistant tumours have a high likelihood of developing into brain metastasis. <sup>[20]</sup> In Zimbabwe, according to the Global Cancer Registry, approximately 104 females between the ages of 25 and 44 are diagnosed with breast cancer, with a low survival rate <sup>[1]</sup>. Patients experience physical, emotional, and social changes that affect their quality of life and productivity and, as a result, national income. <sup>[51]</sup>

Mammography screening, early detection, and early treatment measures have been implemented in developed countries, resulting in improved breast cancer control and mortality rates. In less developed countries, on the other hand, a lack of screening

protocols, awareness, and limited access to diagnostic facilities makes targeted treatment difficult, <sup>[7]</sup> resulting in aggressive tumours resistant to standard therapies.

Since nicastrin is found on the surface of breast cancer cells from resistant cell lines, <sup>[48]</sup> nicastrin-specific inhibitors need to be developed to augment standard breast cancer therapies. Small molecules that target nicastrin and other components of the gamma-secretase complex are available, however binding data does not show small molecules that are specifically designed for nicastrin.

### **1.10 Justification of the study**

Given the challenges presented by breast cancer, therapy that targets specific proteins involved in breast cancer growth and resistance is now being encouraged. According to Francies et al., <sup>[52]</sup> aggressive late-stage cancer with emerging molecular resistance can be managed with systematic and targeted therapies. It is critical to design, and develop, nicastrin-specific inhibitors. Furthermore, breast cancer is the most common type of cancer that spreads to the brain. Because gamma-secretase inhibitors, which are the basis for the computational design of nicastrin inhibitors, were originally created for Alzheimer's disease and can permeate the brain, nicastrin-specific inhibitors may therefore be beneficial in cases of brain metastasis. The success of such therapies could lead to better treatments and overall survival rates, resulting in a higher quality of life and health.

As such, the study aims to develop a long-term approach to addressing people's health and well-being in Africa, and the world at large, with a focus on discovering compounds with activity against breast cancer, particularly resistant subtypes, using chemogenomic methods.

## **1.11 Aims and objectives**

### **1.11.1 Aims**

- Identification of inhibitors of gamma-secretase for breast cancer therapy using chemogenomic methods.

### **1.11.2 Objectives**

- Identify binding sites in the nicastrin subunit and analyse binding modes and interactions of known nicastrin inhibitors through docking calculations and molecular dynamic simulations.
- Analyse and compare the chemical space of known active nicastrin inhibitors and FDA breast cancer drugs.
- Build machine learning models of nicastrin inhibitors.
- Identify inhibitors for nicastrin by virtual screening of compound databases using the machine learning models.
- Determine binding affinities of screened actives and analyse their interactions in the identified binding sites through docking calculations.
- Confirm anti-cancer activity of potential active compounds through biological screening assays.

## 1.12 References

- [1] [1] F. Bray, J. Ferlay, I. Soerjomataram, Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries, *CA Cancer J Clin.* 68 (2018) 394–424. <https://doi.org/10.3322/caac.21492>.
- [2] E. Black, R. Richmond, improving early detection of breast cancer in sub-Saharan Africa: why mammography may not be the way forward, *BMC Glob. Heal.* 15 (2019) 1–11.
- [3] L.A. Torre, F. Bray, R.L. Siegel, J. Ferlay, J. Lortet-Tieulent, A. Jemal, Global cancer statistics, 2012, *CA. Cancer J. Clin.* 65 (2015) 87–108. <https://doi.org/10.3322/caac.21262>.
- [4] D.R. Youlten, S. M. Cramb, N. A. M. Dunn, J.M. Muller, C.M. Pyke, P.D. Baade, The descriptive epidemiology of female breast cancer: An international comparison of screening, incidence, survival and mortality, *Cancer Epidemiol.* 36 (2012) 237–248. <https://doi.org/10.1016/j.canep.2012.02.007>.
- [5] D. Adeloje, O.Y. Sowunmi, W. Jacobs, R.A. David, A. Adeosun, A.O. Amuta, Estimating the incidence of breast cancer in Africa: a systematic review and meta-analysis, *J. Glob. Health.* 8 (2018) 1–21. <https://doi.org/10.7189/jogh.08.010419>.
- [6] Ministry of Health and Child Welfare, The National Cancer Prevention and Control Strategy for Zimbabwe, 2017.
- [7] L.A. Brinton, J.D. Figueroa, B. Awuah, J. Yarney, S. Wiafe, Breast Cancer in Sub-Saharan Africa: Opportunities for Prevention, *Breast Cancer Res. Treat.* 144 (2014) 467–478. <https://doi.org/10.1007/s10549-014-2868-z.Breast>.
- [8] K. B. Sutradhar, L. Amin, Nanotechnology in Cancer Drug Delivery and Selective Targeting, *Nanotechnol.* 2014 (2014) 1–12. <https://doi.org/http://dx.doi.org/10.1155/2014/939378>
- [9] K. Chen, Y. Huang, J. Chen, Understanding and targeting cancer stem cells: therapeutic implications and challenges, *Acta Pharmacol. Sin.* 34 (2013) 732–740. <https://doi.org/10.1038/aps.2013.27>.
- [10] A. Ahmad, Pathways to breast cancer recurrence., *ISRN Oncol.* 2013 (2013) 1–17. <https://doi.org/10.1155/2013/290568>.
- [11] B.G. Cuiffo, A.E. Karnoub, Silencing FOXP2 in breast cancer cells promotes cancer stem cell traits and metastasis, *Mol. Cell. Oncol.* 3 (2016) 1–7. <https://doi.org/10.1080/23723556.2015.1019022>.
- [12] G. Viale, The current state of breast cancer classification, *Ann. Oncol.* 23 (2012) x207–x210. <https://doi.org/10.1093/annonc/mds326>.
- [13] C. Ercan, P.J. van Diest, M. Vooijs and M. V. C. Ercan, P.J. van Diest, Mammary Development and Breast Cancer : *The Role of Stem, Curr. Mol. Med.* 11 (2014) 270–285.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4051995&tool=pmc&rendertype=abstract>.

- [14] H. Al-Hussaini, D. Subramanyam, M. Reedijk, S.S. Sridhar, Notch Signaling Pathway as a Therapeutic Target in Breast Cancer, *Mol. Cancer Ther.* 10 (2011) 9–15. <https://doi.org/10.1158/1535-7163.MCT-10-0677>.
- [15] E.W.J. Mollen, J. Lent, V.C.G. Tjan-heijnen, Moving Breast Cancer Therapy up a Notch, *Front. Oncol.* 8 (2018) 1–25. <https://doi.org/10.3389/fonc.2018.00518>.
- [16] A.K. Shiau, D. Barstad, P.M. Loria, L. Cheng, P.J. Kushner, D.A. Agard, G.L. Greene, H. Hughes, S. Francisco, The Structural Basis of Estrogen Receptor / Coactivator Recognition and the Antagonism of This Interaction by Tamoxifen, *Cell.* 95 (1998) 927–937.
- [17] J.A. Katzenellenbogen, C.G. Mayne, B.S. Katzenellenbogen, G.L. Greene, S. Chandarlapaty, endocrine therapy resistance, *Nat. Rev. Cancer.* (2018). <https://doi.org/10.1038/s41568-018-0001-z>.
- [18] J.P. Leone, B.A. Leone, Breast cancer brain metastases: The last frontier, *Exp. Hematol. Oncol.* 4 (2015) 1–10. <https://doi.org/10.1186/s40164-015-0028-8>.
- [19] B. Cacho-Díaz, H. Spínola-Marño, V.A. Arrieta, M. Granados-García, T. Wegman-Ostrosky, L.G. Mendoza-Olivas, M. Chávez-MacGregor, Diagnosis of brain metastases in breast cancer patients resulting from neurological symptoms, *Clin. Neurol. Neurosurg.* 173 (2018) 61–64. <https://doi.org/10.1016/j.clineuro.2018.08.002>.
- [20] A.D. Hartkopf, E.M. Grischke, S. Y. Brucker, Endocrine-Resistant Breast Cancer: Mechanisms and Treatment, *Breast Care.* 15 (2020) 347–354. <https://doi.org/10.1159/000508675>.
- [21] C.J. Sheeba, G. Marslin, A.M. Revina, G. Franklin, Signaling pathways influencing tumour microenvironment and their exploitation for targeted drug delivery, *Nanotechnol Rev.* 3 (2014) 123–151. <https://doi.org/10.1515/ntrev-2013-0032>.
- [22] B. Purow, Notch Signaling in Embryology and Cancer, *Adv Exp Med Biol.* 727 (2012) 174–315. <https://doi.org/10.1007/978-1-4614-0899-4>.
- [23] X. Yuan, H. Wu, H. Xu, H. Xiong, Q. Chu, S. Yu, G.S. Wu, K. Wu, Notch signaling: An emerging therapeutic target for cancer treatment, *Cancer Lett.* (2015). <https://doi.org/10.1016/j.canlet.2015.07.048>.
- [24] L. Han, S. Shi, T. Gong, Z. Zhang, X. Sun, Cancer stem cells: therapeutic implications and perspectives in cancer therapy, *Acta Pharm. Sin. B.* 3 (2013) 65–75. <https://doi.org/10.1016/j.apsb.2013.02.006>.
- [25] N. Gertsik, D. Chiu, Y.-M. Li, Complex regulation of  $\hat{I}^3$ -secretase: from obligatory to modulatory subunits, *Front. Aging Neurosci.* 6 (2015) 1–10. <https://doi.org/10.3389/fnagi.2014.00342>.
- [26] Y. Lombardo, A. Filipovi, Nicastrin regulates breast cancer stem cell properties and tumour growth in vitro and in vivo, *PNAS.* 109 (2012) 16558–16563. <https://doi.org/10.1073/pnas.1206268109/-DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1206268109>.



- [27] J. Han, M.J. Hendzel, J. Allalunis-turner, Notch signaling as a therapeutic target for breast cancer treatment? *Breast Cancer Res.* 13 (2011) 1–8.
- [28] Y. Li, C. Bohm, R. Dodd, F. Chen, S. Qamar, G. Schmitt-ulms, P.E. Fraser, P.H.S. George-hyslop, Structural biology of presenilin 1 complexes, *Mol. Neurodegener.* 9 (2014) 1–10.
- [29] S.P. McDermott, M.S. Wicha, Targeting breast cancer stem cells, *Mol. Oncol.* 4 (2010) 404–419. <https://doi.org/10.1016/j.molonc.2010.06.005>.
- [30] T. Bouras, B. Pal, F. Vaillant, G. Harburg, M.L. Asselin-Labat, S.R. Oakes, G.J. Lindeman, J.E. Visvader, Notch Signaling Regulates Mammary Stem Cell Function and Luminal Cell-Fate Commitment, *Cell Stem Cell.* 3 (2008) 429–441. <https://doi.org/10.1016/j.stem.2008.08.001>.
- [31] X. Zhang, Y. Li, H. Xu, Y. Zhang, The gamma-secretase complex: from structure to function, *Front. Cell. Neurosci.* 8 (2014) 1–10. <https://doi.org/10.3389/fncel.2014.00427>.
- [32] M.S. Wolfe, Toward the structure of presenilin / gamma-secretase and presenilin homologs Presenilin  $\beta$ , *BBA - Biomembr.* 1828 (2013) 2886–2897. <https://doi.org/10.1016/j.bbamem.2013.04.015>.
- [33] X. Bai, E. Rajendra, G. Yang, Y. Shi, S.H.W. Scheres, Sampling the conformational space of the catalytic subunit of human gamma-secretase, *eLife* (2015) 1–19. <https://doi.org/10.7554/eLife.11182>.
- [34] T.E. Golde, E.H. Koo, K.M. Felsenstein, B. a. Osborne, L. Miele, gamma-Secretase inhibitors and modulators, *Biochim. Biophys. Acta - Biomembr.* 1828 (2013) 2898–2907. <https://doi.org/10.1016/j.bbamem.2013.06.005>.
- [35] B.G. Debeb, E.N. Cohen, K. Boley, E.M. Freiter, L. Li, F.M. Robertson, J.M. Reuben, M. Cristofanilli, T. A., Buchholz, W.A. Woodward, Pre-Clinical studies of Notch signaling inhibitor R04929097 in inflammatory breast cancer cells, *Breast Cancer Targets Ther.* 134 (2012) 495–510. <https://doi.org/10.1007/s10549-012-2075-8.Pre-Clinical>.
- [36] K. Pandya, K. Meeke, A.G. Clementz, A. Rogowski, J. Roberts, L. Miele, K.S. Albain, C. Osipo, Targeting both Notch and ErbB-2 signaling pathways is required for prevention of ErbB-2-positive breast tumour recurrence, *Br. J. Cancer.* 105 (2011) 796–806. <https://doi.org/10.1038/bjc.2011.321>.
- [37] A.F. Schott, M.D. Landis, G. Dontu, K.A. Griffith, M. Rachel, I. Krop, L.A. Paskett, H. Wong, L.E. Dobrolecki, M. Amber, J. Paranilam, D.F. Hayes, M.S. Wicha, J.C. Chang, Preclinical and Clinical Studies of Gamma Secretase Inhibitors with Docetaxel on Human Breast Tumours, *Clin. Cancer Res.* 19 (2013) 1512–1524. <https://doi.org/10.1158/1078-0432.CCR-11-3326.Preclinical>.
- [38] Y. Li, S.H. Lu, C. Tsai, C. Bohm, S. Qamar, R.B. Dodd, W. Meadows, A. Jeon, A. Mcleod, F. Chen, M. Arimon, O. Berezovska, B.T. Hyman, T. Tomita, T. Iwatsubo, C.M. Johnson, L.A. Farrer, G. Schmitt-ulms, P.E. Fraser, P.H.S. George-hyslop, Structural Interactions between Inhibitor and Substrate Docking Sites Give Insight into Mechanisms of Human PS1 Complexes, *Structure.* 22

- (2014) 125–135. <https://doi.org/10.1016/j.str.2013.09.018>.
- [39] A.Y. Kornilova, C. Das, M.S. Wolfe, Differential Effects of Inhibitors on the gamma -Secretase Complex, *J. Biol. Chem.* 278 (2003) 16470–16473. <https://doi.org/10.1074/jbc.C300019200>.
- [40] T. Xie, C. Yan, R. Zhou, Y. Zhao, L. Sun, G. Yang, P. Lu, D. Ma, Y. Shi, Crystal structure of the  $\gamma$ -secretase component nicastrin, *Proc. Natl. Acad. Sci.* 111 (2014) 13349–13354. <https://doi.org/10.1073/pnas.1414837111>.
- [41] G. Yu, M. Nishimura, S. Arawaka, D. Levitan, L. Zhang, A. Tandon, Y.Q. Song, E. Rogaeva, F. Chen, Nicastrin modulates presenilin-mediated notch/glp-1 signal transduction and betaAPP processing., *Nature.* 407 (2000) 48–54. <https://doi.org/10.1038/35024009>.
- [42] S. Shah, S.-F. Lee, K. Tabuchi, Y.-H. Hao, C. Yu, Q. LaPlant, H. Ball, C.E. Dann, T. Südhof, G. Yu, Nicastrin Functions as a  $\gamma$ -Secretase-Substrate Receptor, *Cell.* 122 (2005) 435–447. <https://doi.org/10.1016/j.cell.2005.05.022>.
- [43] N. Zeytuni, R. Zarivach, Structural and functional discussion of the tetra-trico-peptide repeat, a protein interaction module, *Structure.* 20 (2012) 397–405. <https://doi.org/10.1016/j.str.2012.01.006>.
- [44] Y. Lombardo, M. Faronato, A. Filipovic, V. Virillo, L. Magnani, R. Coombes, Nicastrin and Notch4 drive endocrine therapy resistance and epithelial to mesenchymal transition in MCF7 breast cancer cells, *Breast Cancer Res.* 16 (2014) R62. <https://doi.org/10.1186/bcr3675>.
- [45] I. Hayashi, S. Takatori, Y. Urano, Y. Miyake, J. Takagi, M. Sakata-Yanagimoto, H. Iwanari, S. Osawa, Y. Morohashi, T. Li, P.C. Wong, S. Chiba, T. Kodama, T. Hamakubo, T. Tomita, T. Iwatsubo, Neutralization of the  $\gamma$ -secretase activity by monoclonal antibody against extracellular domain of nicastrin, *Oncogene.* 31 (2012) 787–798. <https://doi.org/10.1038/onc.2011.265>.
- [46] I. Hayashi, S. Takatori, Y. Urano, H. Iwanari, N. Isoo, S. Osawa, M.A. Fukuda, T. Kodama, T. Hamakubo, T. Li, P.C. Wong, T. Tomita, T. Iwatsubo, Single chain variable fragment against nicastrin inhibits the gamma secretase activity, *J. Biol. Chem.* 284 (2009) 27838–27847. <https://doi.org/10.1074/jbc.M109.055061>.
- [47] X. Zhang, R. Hoey, A. Koide, G. Dolios, M. Paduch, P. Nguyen, X. Wu, Y. Li, S.L. Wagner, R. Wang, S. Koide, S.S. Sisodia, A synthetic antibody fragment targeting nicastrin affects assembly and trafficking of  $\gamma$ -secretase, *J. Biol. Chem.* 289 (2014) 34851–34861. <https://doi.org/10.1074/jbc.M114.609636>.
- [48] A. Filipović, Y. Lombardo, M. Fronato, J. Abrahams, E. Aboagye, Q.-D. Nguyen, B.B. d'Aqua, A. Ridley, A. Green, E. Rahka, I. Ellis, C. Recchi, N. Przulj, A. Sarajlić, J.-R. Alattia, P. Fraering, M. Deonarain, R.C. Coombes, Anti-nicastrin monoclonal antibodies elicit pleiotropic anti-tumour pharmacological effects in invasive breast cancer cells, *Breast Cancer Res. Treat.* 148 (2014) 455–462. <https://doi.org/10.1007/s10549-014-3119-z>.

- [49] M.A. Arai, R. Akamine, A. Tsuchiya, T. Yoneyama, T. Koyano, The Notch inhibitor cowanin accelerates nicastrin degradation, *Nat. Sci. Reports.* 8 (2018) 1–8. <https://doi.org/10.1038/s41598-018-23698-4>.
- [50] J. Ferlay, H.R. Shin, F. Bray, D. Forman, C. Mathers, D.M. Parkin, Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008, *Int. J. Cancer.* 127 (2010) 2893–2917. <https://doi.org/10.1002/ijc.25516>.
- [51] S. Subramaniam, Y.-C. Kong, K. Chinna, M. Kimman, Y.-Z. Ho, N. Saat, A.R. Malik, A.N. Taib, M.M. Abdullah, G.C.-C. Lim, Y.-L. Tamin, Nor-Saleha Ibrahim Woo, K.-M. Chang, P.-P. Goh, C.-H. Yip, N. Bhoo-Pathy, Health-related quality of life and psychological distress among cancer survivors in Southeast Asia: results from a longitudinal study in eight low-and middle-income countries., *J. Psychol. Soc. Behav. Dimens. Cancer.* 27 (2018) 2172–2179. <https://doi.org/10.1002/pon.4787>.
- [52] F.Z. Francies, R. Hull, R. Khanyile, Z. Dlamini, Breast cancer in low-middle income countries: abnormality in splicing and lack of targeted treatment options., *Am. J. Cancer Res.* 10 (2020) 1568–1591. <http://www.ncbi.nlm.nih.gov/pubmed/32509398><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC7269781>.

## **2 Literature review: Experimental Methods and Techniques**

### **2.1 Introduction**

This chapter provides an overview of the methods used in this study, including binding site identification and validation methods for the gamma secretase units, as well as ligand-based methods for building machine learning QSAR models to investigate structure-activity relationships. Drug delivery barriers, pharmacokinetic analysis, and antitumour tests are also discussed.

### **2.2 Binding site identification**

Throughout an organism's life cycle, proteins are involved in a variety of vital processes and activities that either directly or indirectly carry out significant biological functions. They interact with other biomolecules to carry out specific functions. These interactions take place at locations on the protein known as the binding sites. <sup>[1]</sup> The identification of binding sites is essential in drug design and optimisation.

Binding sites can be identified using substrates, protein structure, and sequence annotation. <sup>[2-4]</sup> The function of the gamma secretase unit, protein nicastrin, has primarily been studied through sequencing, which has resulted in the identification of conserved residues. However, there is a need to expand the functional annotation of nicastrin by identifying binding sites, which are important in drug design. <sup>[5, 6]</sup> Chemogenomic strategies can be used to expand the functional annotation. <sup>[7-9]</sup> Algorithms that use protein structural information to identify binding sites on protein surfaces have proven to be effective, and these can be applied to nicastrin. <sup>[4]</sup>

To predict binding sites, some algorithms search for pockets on the protein surface, while others use binding energies of probes or known ligands placed on a grid to

predict these binding sites. <sup>[10,11]</sup> The ensuing sections provide examples of these techniques.

### **2.3 Binding site identification methods**

In identifying binding sites using structural information, template-based methods <sup>[1,6,12]</sup> geometry-based methods, <sup>[13,14]</sup> and methods based on energy-related calculations <sup>[13]</sup> are commonly used. For example, Goswami et al. <sup>[15]</sup> identified the ThDP binding pocket of the *P. falciparum* DXP synthase using a consensus-based COACH method, whilst Lanka <sup>[16]</sup> used site map to identify binding cavities for inhibitors targeting FAM3B causing type 2 diabetes mellitus.

The principle of similar proteins with similar functions is used in template-based methods to transfer known binding sites from known homologues to the query structure using sequence or structural alignment techniques. <sup>[12]</sup> Unfortunately, template-based methods are limited to templates with high levels of sequence identity, which are required for accurate binding site identification. <sup>[1,6]</sup> Nicastrin's structural similarity to its homologues is less than 50%, with most of the similarities confined to the large lobe. <sup>[17]</sup> As a result, using template-based methods to define nicastrin-binding sites becomes difficult, necessitating the use of other methods.

Methods that use information from the protein structure, such as geometry-based and energy-based methods, are preferred for improving functional annotation. <sup>[18–20]</sup> The atoms of the protein are initially represented as hard van der Waals non-bonded spheres. These spheres represent the distribution of electron clouds around atomic nuclei. The protein is then described as a space-filled shape composed of non-bonded atom balls, with the geometry (area and volume) or physicochemical properties of the protein surface being used to identify protein binding sites. <sup>[21]</sup> Geometry-based

methods identify cavities on the protein's surface by considering the target surface's spatial geometry measurements via a grid system scan, alpha shape, or probe sphere filling. <sup>[14]</sup> The limitation of geometry-based methods is that they are sensitive to protein orientation and grid spacing.

Energy-based methods, on the other hand, estimate the binding site as the location on the protein's surface with the lowest interaction energy between a probe and the protein atoms. <sup>[13]</sup> The Q-SiteFinder is an example of an energy-based method that locates energetically favourable binding sites by utilising the interaction energy between the protein and a simple van der Waals probe. Clusters of energetically favourable probe sites are ranked based on the sum of interaction energies for sites within each cluster. <sup>[11]</sup> Both of these methods have some limitations. A disadvantage of energy-based methods is that they rely on scoring functions for accurate predictions, which can be difficult to establish at times. <sup>[22]</sup>

### **2.3.1 Energy based methods for binding site identification**

Energy-based methods use the distribution of energy values to predict binding sites by calculating non-bonded interactions between atoms within a molecule and those in other molecules. <sup>[13]</sup> Energy-based methods have the advantage of not being affected by protein size, and binding sites can be small but sufficient to accommodate ligands of various sizes. Different chemical probes can also be used to interrogate the protein surface, resulting in the identification of binding sites with different chemical properties within the same site. <sup>[23]</sup> Energy-based methods such as those based on the Lennard Jones potential as well as ligand docking were investigated in this work.

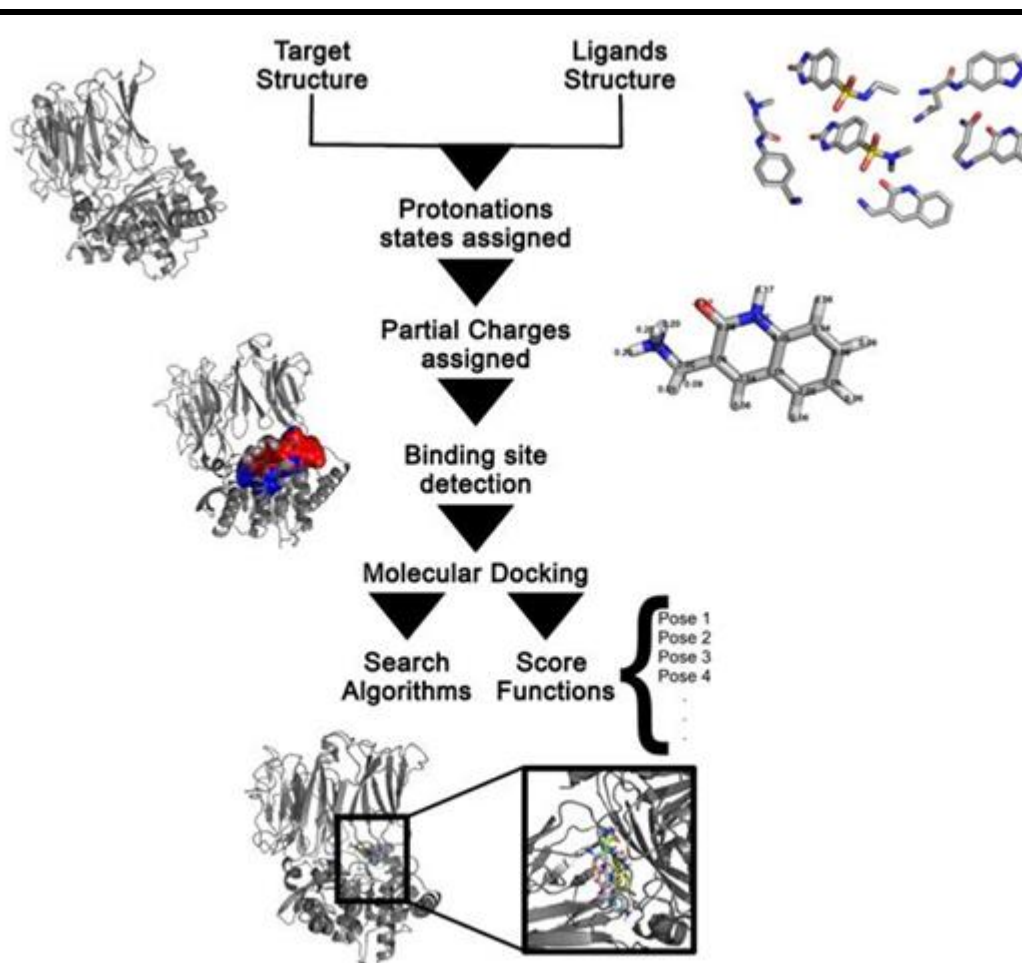
### 2.3.1.1 Ligand docking methods

Non-bonded interactions, as well as other types of energy, such as electrostatic interactions using the coulombic formula, hydrogen bonds, hydrophobic effects, and ionic interactions, are calculated using ligand docking methods. [24] Blind docking is the process of identifying probable binding modes across the entire surface of a protein which is typically done when the binding site on a protein is unknown. [11, 25] The docking method predicts the intermolecular framework between a protein and a ligand, indicating the most energy-efficient binding modes that cause inhibition. [26, 27] To obtain the preferred protein-ligand complex, docking runs and energy calculations are repeatedly performed until a favourable pose is discovered. [11, 28]

The docking process can be divided into two coherent stages as shown in **Figure 2.1**. The first stage is sampling to explore the ligand's poses in the protein binding site. [28] Different algorithms are used to sample the most energy-efficient modes, which distinguishes docking programs. Shape matching, systematic search (exhaustive search, fragmentation, and conformational ensemble), and stochastic search algorithms are the most commonly used sampling algorithms (Monte Carlo simulations, genetic algorithm, Tabu search and swarm optimisation). [26, 29]

The second stage of docking is scoring where results are optimised and ranked based on predicted binding affinities. [28] The most accepted low-energy conformations are found in potential binding sites. Scoring functions include force-field-based, empirical-based, descriptor-based, knowledge-based, and consensus-based methods and some methods combine these functions. [30] For example, the scoring function AutoDock Vina uses a combination of empirical and knowledge-based algorithms. [11,

31]



**Figure 2.1** The docking process. [32]

### 2.3.1.2 Lennard -Jones based methods

Most energy-based methods use the Lennard-Jones potential [33] to approximate binding energies and bond lengths using probes that outline protein surfaces in search of clefts and voids. Different algorithms have different probes. For example, PocketFinder uses an aliphatic probe [34] whereas Q-SiteFinder uses a methyl probe.

[23] The Lennard-Jones potential is a method for calculating the energy for atom pairs

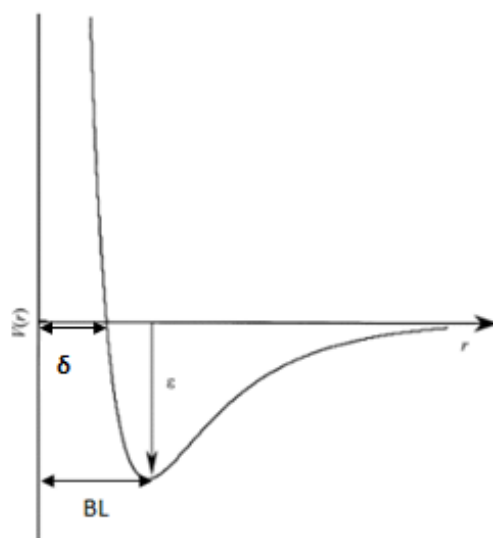


that are not bonded. The binding energies are calculated using the depth of potential wells at preferred interaction distances between the probe and the protein atom. <sup>[35]</sup> These interaction distances represent the bond lengths between atom pairs. **Equation 2.1** can be used to calculate potential energy, which can then be plotted as shown in **Figure 2.2**.

$$PE = 4\varepsilon - \left[\left(\frac{\delta}{r}\right)^{12} - \left(\frac{\delta}{r}\right)^6\right]$$

**Equation 2.1**

Where PE denotes potential energy,  $\varepsilon$  is the depth of the potential energy well and  $\delta$  is the inter atomic distance at zero potential and zero attraction.



**Figure 2.2** The Lennard-Jones potential.

Used to approximate the binding energy as well as the bond length. The depth of the potential well,  $\epsilon$ , represents the binding energy,  $\delta$  is the interatomic distance at zero potential, and zero attraction and BL represents the bond length.

---

## 2.4 Binding site identification methods used for nicastrin

In this work, blind docking with AutoDock Vina <sup>[36]</sup> and PocketFinder <sup>[34]</sup> in the Internal Coordinate Mechanism software (ICM) <sup>[37]</sup> were used to predict and characterise binding sites of nicastrin.

### 2.4.1 AutoDock Vina

The AutoDock Vina <sup>[36]</sup> program was used to identify potential binding sites in nicastrin. AutoDock Vina employs an evolutionary search to determine the conformational space of the ligand. <sup>[11]</sup> An appropriate binding site is mapped by first positioning the ligand (pseudo-random space) used for binding site identification in a 3D space, which is oriented using a rotational vector and a set of rotatable bonds and torsions that are twisted to certain degrees to generate a pose. The protein-ligand conformations are

optimised by determining the best rotation and translation of the ligand within the protein to generate acceptable poses through global optimization via a modified Monte Carlo algorithm and the Metropolis criteria. The Broyden-Fletcher-Goldfarb-Shanno method is used for local optimization after global optimisation. [31, 38]

The generated poses are then scored by using a hybrid algorithm that combines empirical and knowledge-based algorithms. [11, 31] The scoring function takes into account both intermolecular energy (van der Waals and electrostatic interactions between protein and ligand atoms) and intramolecular energy, with the best pose having the least amount of interacting energy, known as binding energy,  $E$ . [36]

**Equations 2.2 to 2.4** describe the Vina scoring function. [36] The binding energy is predicted as the sum of distance dependent atom pair interactions **Equation 2.2**.

$$E = \sum e_{pair}(d)$$

**Equation 2.2**

Where  $d$  is the surface distance calculated using the **Equation 2.3**, where  $r$  is the interatomic distance and  $R_i$  and  $R_j$  are the radii of the atom pair.

$$d = r - R_i - R_j$$

**Equation 2.3**

Every atom pair interacts through a steric interaction given by the first three terms of **Equation 2.4**. There could also be hydrophobic and non-directional H-bonding interactions, given by the last two terms of **Equation 2.4**.

$$e_{pair}(d) = \begin{cases} w_1 * Gauss_1(d) + \\ w_2 * Gauss_2(d) + \\ w_3 * Repulsion(d) + \\ w_4 * Hydrophobic(d) + \\ w_5 * HBond(d) \end{cases}$$

**Equation 2.4**

## 2.4.2 Internal Coordinate Mechanism

The Internal Coordinate Mechanism (ICM) method developed by Molsoft L.L.C [37, 39] was used to identify binding sites, and the sites were compared with those identified during the docking process. The PocketFinder algorithm in ICM creates an orthogonal parallelepiped grid potential map around a protein constructed from a van der Waals force field using an aliphatic probe based on the transformed Lennard-Jones formula (**Equation 2.5**). [34]

$$P_p^o = \sum_{l=1}^N \left( \frac{A_{XlC}}{r_{pl}^{12}} - \frac{B_{XlC}}{r_{pl}^6} \right)$$

**Equation 2.5**

Where  $r_{pl}$  is the distance between the probe  $p$  at the grid node and the protein atom  $X_l$ .  $A_{XC}$  and  $B_{XC}$  are molecular mechanics force fields adapted from the Empirical Energy Program for Peptides. Attractive regions of are maintained by truncating  $P_p^o$  to range of a -0.8.

The 3D grid potential map contains grid points that are used to approximate the shape of the protein [40] with 1.0 Å spacing and an additional 1.0 Å margin beyond the protein's dimensions. [34] The potential energy of the aliphatic probe is stored in these grid points, and the cumulative sum describes the binding site. The generated potential map is then smoothed to highlight van der Waals potential regions and contoured to create potential ligand envelopes.

### 2.4.3 Druggability analysis

After identifying potential protein binding sites, the next step is to determine whether the identified binding sites can bind small drug-like molecules, that is if they are druggable. The druggability analysis connects protein properties to the physicochemical properties of drug-like compounds. <sup>[5,41]</sup> The drug-like properties of a small molecule should be complemented by properties present in the binding site for high affinity binding and modulation or inhibition. During the assessment, druggability indices derived from characteristics of known ligand binding sites are typically used to score identified binding sites. <sup>[41–43]</sup> Ligand binding sites can be characterised by hydrophobicity or polarity, volume, buriedness, and compactness. A metric can be used to determine whether drug-like compounds can target a binding site. An example of such a metric is the DLID score by Sheridan and colleagues <sup>[44]</sup> used to evaluate binding sites. The DLID score is a combination of:

1. Pocket volume.
2. Buriedness, which is a ratio of the solvent accessible surface area covered by its shell to the solvent accessible surface area of the pocket in isolation. The most buried pocket is 1.0 and open pocket, 0.5.
3. Hydrophobicity is the fraction of the pocket surface in contact with hydrophobic atoms of its shell

The DLID is then obtained by combining the above parameters using Equation 2.6.

$$DLID = -8.7 + 1.72 \log(\text{volume}) + 3.94(\text{buriedness}) \\ + 2.27(\text{hydrophobicity})$$

**Equation 2.6**

Where a high DLID (=1) score represents a druggable binding site

To visualise and rank pockets, a druggability landscape based on the calculated DLID score can be plotted. Characterisation of identified binding sites to determine whether they are druggable is critical, given that the majority of known nicastrin-targeting compounds are drug-like molecules. In this study, the deposited 3D structure of gamma-secretase conformers was used to identify possible binding sites in nicastrin. Since ligand binding induced conformational changes were noticed in nicastrin, druggability analysis of the conformers was done to identify nicastrin conformers that would be used for docking and virtual screening processes that followed. The DLID score was calculated within ICM PocketFinder.

#### **2.4.4 Molecular Dynamic Simulations**

A popular method for determining binding sites and modes is molecular docking. The top-scoring poses from docking calculations, which are the best binding modes for inhibitors, can be improved by including structural, dynamic, and entropy effects for the protein ligand. In addition to exploring the conformational aspects of protein-ligand systems, molecular dynamic effects also detail protein-ligand interactions and make binding affinity predictions using binding free energy calculations. <sup>[45]</sup> There are a number of software for molecular dynamic simulations, the GRoningen MAchine for Chemical Simulations (GROMACS) <sup>[46]</sup> Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS), <sup>[47]</sup> Nanoscale Molecular Dynamics (NAMD), <sup>[48]</sup> Chemistry at Harvard Macromolecular Mechanics (CHARMM), <sup>[49]</sup> GENeralised-Ensemble Simulation System (GENESIS) <sup>[50]</sup> and Assisted Molecular Building with Energy Refinement (AMBER) <sup>[51]</sup> amongst others.

In this study, the Molecular dynamic simulations were done through the GRONingen MAchine for Chemical Simulations (GROMACS) 2022.1 software. <sup>[46]</sup> The integration of Newton's equation of motion was used to calculate the movements of atoms over time (**Equation 2.7**).

$$\frac{d^2 r_i(t)}{dt^2} = \frac{F_{i(t)}}{m_i}$$

**Equation 2.7**

Where  $F_{i(t)}$  is force exerted on atom  $i$  at time  $t$ ,  $r_i(t)$  is the vector position of the atom  $i$  at time  $t$  and  $m_i$  is the mass of the atom.

The time is partitioned into time steps ( $\delta t$ ) which is used to propagate the system in time. Integration algorithms are used to derive the Newton's equations using a discrete time numerical approximation. The velocity-Verlet integrator was used to compute the position and velocity of atoms as shown in **Equation 2.8**.

$$r_i(t + \delta t) = r_i(t) + v_i(t) \delta t + \frac{1}{2} a_i(t) \delta t^2$$

$$v_i(t + \delta t) = v_i(t) + \frac{1}{2} [a_i(t) + a_i(t + \delta t)] \delta t$$

**Equation 2.8**

Where  $r_i(t)$  and  $v_i(t)$  are the position, velocity and acceleration of the atom  $i$  at time  $t$  respectively and  $r_i(t + \delta t)$ ,  $v_i(t + \delta t)$  and  $a_i(t + \delta t)$  are position, velocity and acceleration of the atom  $i$  at time  $t$  respectively.

Acceleration is calculated from the forces acting on the atom  $i$  according to Newton's second law (**Equation 2.9**). These forces are computed from the force field. In this study the Charmm36 all atom force field was used. The force field is a function that approximates potential energy, as a sum of bonded (intramolecular) and non-bonded energy terms. [52] The force field estimates bond stretching (harmonic potential), torsional potential (trigonometric function) in the bonded potential. Non-bonded terms consist of van der Waals and Coulomb electrostatic interactions between atoms (**Equation 2.10 and Equation 2.11**)

$$\mathbf{a}_i(t) = \frac{d^2\mathbf{r}_i(t)}{dt^2} = \frac{\mathbf{F}_i(t)}{m_i} = -\frac{dV(\mathbf{r}(t))}{m_i d\mathbf{r}_i(t)}$$

**Equation 2.9**

$$V_{bonded} = \sum_{bonds} K_b(b - b_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} K_X(1 + \cos(nX - \delta))$$

**Equation 2.10**

$$V_{nonbonded} = \sum_{nonbonded\ pairs\ ij} \frac{q_i q_j}{\epsilon r_{ij}} + \sum_{nonbonded\ pairs\ ij} \epsilon_{ij} \left[ \left( \frac{R_{min,ij}}{r_{ij}} \right)^{12} - 2 \left( \frac{R_{min,ij}}{r_{ij}} \right)^6 \right]$$

**Equation 2.11**

Where  $K_b$ ,  $K_\theta$ , and  $K_X$  are the bond, angle and torsion force constants;  $b$ ,  $\theta$  and  $X$  are bond length, bond angle and dihedral angle;  $n$  is the multiplicity and  $\delta$  is the phase of the torsion periodic function;  $r_{ij}$  is the distance between atoms  $i$  and  $j$ ;  $q_i$  and  $q_j$  are the partial charges of atoms  $i$  and  $j$  and  $\epsilon$  is the effective dielectric constant;  $\epsilon_{ij}$  is the Lennard-Jones well depth and  $R_{min,ij}$  is the distance between atoms at Lennard-Jones minimum.



The AnteChamber Python Parser interface (ACPYPE) <sup>[53]</sup> based on ANTECHAMBER <sup>[54]</sup> was used for generating automatic ligand topologies as it creates ligand topology in the format required by GROMACS. The information on the ligand is collected from the molecular coordinate files and the net charge of the ligand is defined using the Gasteiger method. <sup>[55]</sup> The generalized AMBER forcefield, (GAFF) <sup>[56]</sup> was used for ligand parameterisation. Atomic partial charges were determined via ANTECHAMBER.

After parameterisation of both the protein and the ligand, a simulation box is defined to minimise the edge effects in a finite system by application of periodic boundary conditions (PBCs). The use of the PBCs allows for the inclusion of solvent or crystalline environments, whilst containing the molecules so as to preserve the thermodynamic properties such as pressure, temperature and density. The protein and ligand are placed into a space filling box, which is surrounded by translated copies of itself. The space filling boxes are available in different shapes and are suited to a system depending on the shape of the macromolecule. The rhombic dodecahedron and truncated octahedron are more spherical than cubic and hence used in the macromolecule is spherical in shape. The following section describes the ligand-based methods that were used after binding site, binding mode and interaction analysis.

## **2.5 Ligand based methods**

The assumption that similar ligands exhibit similar activity and share similar targets, <sup>[56]</sup> and targets with similar binding sites bind to similar ligands, allows for the investigation of pharmacological links between ligands and their targets. <sup>[57, 58]</sup> In ligand-based drug design, quantitative structure activity relationships (QSAR) and pharmacophore modelling are the most commonly used techniques.

## 2.5.1 Quantitative Structure Activity Relationships

QSARs are used to link chemical structure information to biological activity. <sup>[59]</sup> The primary property of interest in drug design is biological activity, which can be expressed as binding affinity ( $K_i$ ), minimal inhibition concentration (MIC), concentration that causes 50% inhibition ( $IC_{50}$ ), and lethal dose ( $LD_{50}$ ). <sup>[60]</sup> The QSAR model, which relates chemical structures to biological activity, suggests that structurally similar molecules exhibit similar activity. It can be used to expedite and automate the evaluation of biological activity of compound libraries in order to find potential actives towards a target, greatly assisting in the drug discovery process. Following the standard steps required for the creation of a valid model creates these models. These include dataset collection, preparation, and analysis, descriptor calculation, model construction, and model evaluation. <sup>[61]</sup>

## 2.5.2 Dataset collection and curation for Machine Learning

### 2.5.2.1 Dataset collection for QSAR models

Datasets for a QSAR study, are obtained from a variety of sources, including public databases that house annotated biological activity data from high-throughput screening experiments, such as PubChem <sup>[62]</sup> and ChEMBL <sup>[63]</sup> scientific publications, and user-generated data directly from experiments. The collection procedure entails gathering structural data on compounds as well as their biological activity as described by the endpoint for a specific target. When used for QSAR modelling, the chemical structures of the molecules are presented in a machine-readable 2D or 3D format rather than a 2D sketch, as is common in the literature, but rather SMILES (simplified

molecular input line entry system) and SDF (structure data file), which present the molecule in a connection table format. The other formats include MOL2, which is similar to the SDF includes partial charges, and InChI (IUPAC International Chemical Identifier). [64]

Given that the quality of the collected dataset has a significant influence on the quality of the produced QSAR model, it is critical to investigate the types of assays used as well as potential errors in the experimental results. A manual inspection can be performed to ensure quality if the dataset is large enough. [61, 65, 66] The type of QSAR model developed is determined by how the data was collected. For example, predictive categorical models are typically developed when compounds from different assays with different experimental protocols are used for model building. [67]

#### **2.5.2.2 Dataset curation**

Given that the conventional QSAR techniques are limited in the molecular representations of organic structures they can handle; it is critical to clean the chemical structures after collection. The cleaning procedure includes the standardisation, neutralisation, and removal of salts, inorganic, and organometallic compounds. Structures are also canonised to deal with ring aromatisation, neutralisation, and duplicate deletion. [64, 66, 68] Furthermore, the chemical space size, chemical diversity, and activity distribution of the dataset can be used to characterise the dataset's quality. [69] Learning chemical spaces and assuming their activity can help in characterizing the dataset's quality. [70, 71] The chemical space can be defined by common physicochemical properties used in drug-like property analysis of compounds. This includes molecular weights (MW) ranging from 200 to 500 Da, topological polar

surface area (TPSA), hydrogen bond donors (HBD), and hydrogen bond acceptors (HBA); the number of rotatable bonds (nRotB), and solubility predicted by the octanol-water partition coefficient (AlogP).<sup>[72, 73]</sup>

Principal component analysis (PCA), which involves scaling the physicochemical properties to mean zero and variance of one and computing principal components that can be viewed in 2D or 3D, can be used to perform quantitative evaluation and visualisation of the chemical space.<sup>[71, 73]</sup> Chemical diversity of ligand data sets can also be examined using clustering and the calculation of Tanimoto similarity or Euclidean distance by computing pairwise data set diversity of the physicochemical properties and viewing the resulting distance matrix.

### **2.5.2.3 Dataset preparation**

The usefulness of a QSAR model is determined by its ability to accurately predict biological activity, which is accomplished by establishing the predictive power of the model and providing a guarantee of accuracy of the models attached via model validation.<sup>[74]</sup> To validate and establish the predictive power of the models, the dataset should be divided into a training and test set. Model fitting and internal validation are performed on the training set, while prediction and model validation are performed on the test set.<sup>[75, 76]</sup> The dataset is split, either rationally or randomly, to obtain the training and test sets. To avoid bias, the split should ensure that the training set<sup>[77]</sup> covers the chemical space of the test set rather than the actual structure. A rational splitting is based on biological activity or molecular descriptors that quantitatively describe the structure of the compounds.

Some rational splitting methods have the disadvantage of biasing the test set towards the training set, obscuring the model's quality. <sup>[75, 76]</sup> The sizes of the training and test sets also have an impact on the model's quality. The researcher usually chooses different training set-to-test set ratios while keeping in mind the following points raised by Roy and others. <sup>[78]</sup>

1. If the training set contains a small number of compounds, the prediction of the test set is poor because there may not be enough structural information for learning and thus prediction.
2. When a test set contains an insufficient number of compounds, a model with an inflated predictive performance is produced. Typically, the model is built using a training set that comprises between 70% and 90% of the dataset.

### **2.5.3 Descriptor calculation**

The quantitative part of quantitative structure activity relationships represents the independent variables, which are the descriptors that represent the chemical structure. Ligand-based methods are based on the representation of ligands as molecular property descriptors, scaffolds, or fingerprints. <sup>[64, 79, 80]</sup> When computer algorithms are applied to a chemical structure, and its environment, to generate numerical representations of the structure, molecular descriptors are produced.

To calculate descriptors, a variety of tools, and online services are available, including PaDEL, <sup>[81]</sup> CDK, <sup>[82]</sup> DRAGON, <sup>[83]</sup> Corina, <sup>[84]</sup> ChemDes, <sup>[85]</sup> PowerMV, <sup>[86]</sup> and OpenBabel <sup>[87]</sup> to name a few. These tools generate descriptors that fall into different categories based on their information content and dimensionality. The most basic descriptors (1D) account for counts of atom types or fragments and describe the

molecule's global properties such as molecular weight, hydrogen bond donors and acceptors, which are derived from chemical formulae. [88]

These descriptors, also known as constitutional descriptors, describe the chemical composition of the structure while ignoring the compound's connectivity. [88–90] These descriptors are computed quickly and easily, making them ideal for virtual screening campaigns aimed at decoding structural information that affects biological activity. They are useful in the absence of known compounds bound to a biological target because they are not conformation specific in describing the activity of that compound.

Descriptors derived from algorithms applied to the topological representation of the molecule (2D) generate a connectivity table that allows the molecule to be encoded as bit strings or as a graphical representation of the molecule. [65] These are known as topological descriptors, and they describe how the compound is connected by describing the types of bonds and atom interactions that exist [91, 92]. For example, the connectivity of a molecule composed of carbon, oxygen, and nitrogen atoms is clearly described, as is information on their electronegativity, hybridisation, and atomic reactivity. [92] Although they are computationally slower, these are the most commonly used descriptors. Fingerprints can also be used as a 2D representation of molecules. A molecule is decomposed into binary fragments in fingerprint representation, and a hashing algorithm is used to generate possible chemical fragments, each represented by a bit. [93]

Fingerprint-based calculations are much faster to compute than graphical displays. Geometrical (3D) descriptors are derived from algorithms that are applied to the spatial details of a molecule rendered in space as a rigid geometrical object. [79] The incorporation of energy interactions between molecules results in 4D descriptors. [80]

#### **2.5.4 Removal of redundant and correlating descriptors**

Descriptor calculating tools generate a large number of molecular descriptors, which if not reduced, reduces model performance [77] increases the risk of overfitting data due to the presence of redundant descriptors [77, 94, 95] and reduces model interpretability and cost effectiveness of models. [96] Prior to building the model, suitable criteria for selecting descriptor subsets from all calculated descriptors, as well as a measure of the discriminative worth of each descriptor or its redundancy, should be provided in order to reduce the number of descriptors. Dash and Liu [97] outline a step-by-step procedure for descriptor selection, that includes the generation of a descriptor subset, evaluating it, establishing a stopping criterion, and validation.

#### **2.5.5 Descriptor subset generation**

During the generation step, various methods can be used, which are broadly classified as exponential, sequential, and random searches. [97] The difference between these groups is that exponential methods evaluate a number of subsets that grow exponentially with the size of the descriptor space. Exhaustive search and branch and bound search are two examples of exponential methods. [98] In contrast, sequential methods add or remove descriptors sequentially, either one at a time or as few as possible. Several of these methods, such as greedy forward selection or backward elimination, best-first, linear forward selection, floating forward and backward selection, beam search, and race search, are frequently used in descriptor subset selection. [99, 100] Random searches are methods that select random subsets and include simulated annealing, scatter search, random generation, genetic algorithm,

and colony optimisation algorithm. [99, 101] During subset generation, the search direction must be determined, and can be forward, backward, bidirectional, or random.

Forward selection methods begin with an empty set and add relevant descriptors one at a time, whereas backward elimination begins with a full set of descriptors and removes redundant descriptors one at a time. [76, 95, 96] A bi-directional search combines forward and backward search techniques, and it begins at both ends. A random search chooses a starting point at random and proceeds in one direction. [95]

The best first subset selection method was used to generate the descriptor subsets that were used in this study. [96] The best first search method finds the best descriptor in the descriptor space by making local changes to the current subset rather than incrementally adding descriptors to subsets generated thus far. [100] The method combines a greedy hill climbing algorithm with backtracking to search the descriptor space for a more promising subset using an evaluation function, and it can search in any direction. [100,102] As such, if the search path for descriptors does not contain any 'good' descriptors, it returns to a previous subset with more promising descriptors and continues the exploration.

#### **2.5.5.1 Descriptor subset evaluation and stopping criterion**

Three methods for evaluating generated descriptors are the filter, wrapper, and hybrid methods [97]. Filter methods select descriptors independently of the learning algorithm, whereas wrapper descriptor selection methods select descriptors concurrently with the learning algorithm. [103–105] Hybrid methods are a combination of the two methods. An evaluation is performed for each of the generated subsets to determine the goodness of the subset in comparison to the previous one. The best subset is kept after the



stopping criterion is reached, which is set when either the number of predefined descriptors or iterations is reached. [106] In the face of a large descriptor space, the stopping criterion is required to control the descriptor subset generation process.

In this study, the correlational based subset evaluator (CFS) and wrapper methods were used to evaluate descriptor subsets. The CFS is a heuristic filter method for evaluating subsets by removing features with low correlation with the class. [105,107] The CFS subset assumes that all of the descriptors are highly correlated with the activity but not with one another. In **Equation 2.12**, the heuristic is represented as a variant of Pearson's correlation equation. [105]

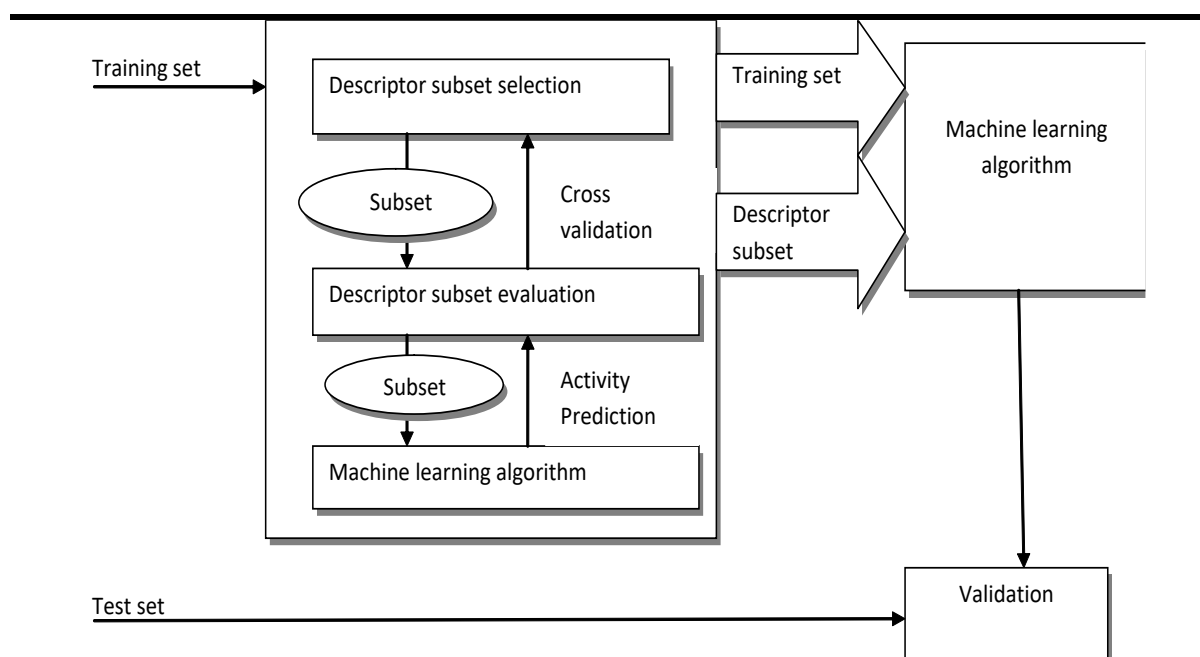
$$G_s = \frac{k\bar{r}_{cf}}{k + k(k - 1)\bar{r}_{ff}}$$

#### **Equation 2.12**

Where  $G_s$  is the evaluation of the subset of  $S$  containing  $k$  features,  $\bar{r}_{cf}$  is the average correlation value between the descriptors and classes and  $\bar{r}_{ff}$  is the average correlation between two descriptors.  $k\bar{r}_{cf}$  is a measure of how a descriptor subset can predictive the correct class and  $k + k(k - 1)\bar{r}_{ff}$  is a measure of correlation between the descriptors.  $G_s$  is then used to select only the best descriptors that are highly correlated with the class.

Wrapper methods were used to select descriptors concurrently during the learning process. [97, 99, 100] The study employed widely used machine learning algorithms such as Naïve Bayes, [108] IB1, [109] J48, [110] and SMO. [111] Machine learning algorithms build a model from training data. Depending on the algorithm, the models are presented in various ways. For example, a Naïve Bayes algorithm presents the model as a probabilistic summary, whereas a J48 tree presents the model as a decision tree. For descriptor subset selection, evaluation, model building, and validation WEKA

workbench version 3.8.2 was used. <sup>[112]</sup> Before the classification using the training set, the attribute selected evaluator is used first. Internally, the training set is divided into different subsets of descriptors for training and evaluation, with the best subset yielding the highest accuracy measure being chosen for model building. In this study, descriptor subsets were evaluated using five-fold cross validation. <sup>[100]</sup> **Figure 2.3** summarises the wrapper subset evaluation.



**Figure 2.3** The Wrapper subset evaluation illustration. <sup>[100]</sup>

## 2.5.6 Other Machine learning Algorithms

### 2.5.6.1 Naïve Bayes

The Naïve Bayes classifier is a simple probabilistic classifier that is based on the Bayes theorem. <sup>[108]</sup> The Bayes formula, shown in **Equation 2.13** is used to derive the Naïve Bayes theorem. The Bayes theorem can predict the likelihood of A, occurring if B has already occurred. When a training set is provided, B represents numerical values from the subset with good descriptors, and A is the class variable that

represents compound activity. The descriptors are converted into a large set of binary features, and a weight for each feature is calculated using the Laplacian-adjusted probability estimate. The total probability estimate can be used to forecast the activity of a compound. [113, 114]

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

**Equation 2.13**

Where P(A) and P(B) are probabilities of A and B respectively.

### 2.5.6.2 Instance based learner

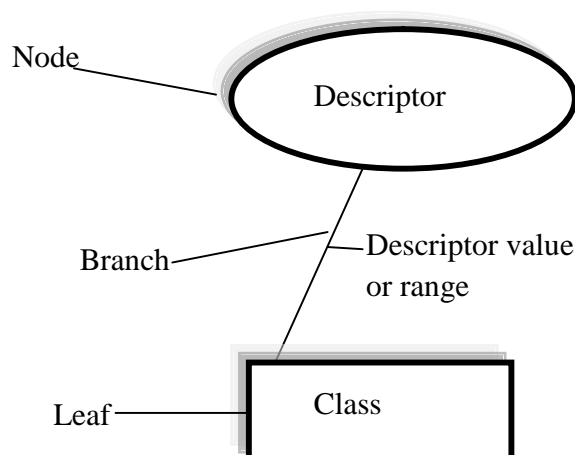
Instance-based learners (IB1) employ the nearest neighbour pattern classifier principle. When predicting the activity of a given compound, it computes similarities using a distance measure between the new compound and the stored training compounds to find a neighbour. [109] IB1 simply retrieves the stored compound closest to the compound to be classified (the nearest neighbour) and assigns the class label of the neighbour to the new compound. In IB1, 1 refers to the researcher's choice of one nearest neighbour, and the number of neighbours can be changed. The IB1 algorithm is known as a lazy algorithm because it only works during prediction.

### 2.5.6.3 J43 Trees

J48 trees are based on the C4.5 algorithm [110], which generates a decision tree as a summary of the training set. **Figure 2.4** illustrates part of a decision tree. In **Figure 2.4**, the tree is made up of nodes that represent the descriptors, branches that represent the numerical values of these descriptors, and leaves that represent the

activity. When a test set compound is introduced, its descriptors and values are examined, and a corresponding class is predicted.

---



**Figure 2.4** Illustration of part of a decision tree.

---

#### 2.5.6.4 Support vector machines

The sequential minimal optimisation (SMO) algorithm is used to train a support vector machine. To distinguish between actives and inactives, data is nonlinearly mapped through a hyperspace. Support vectors drawn from compounds in the training set are used to define the boundaries between the two classes. A hyperspace is defined as a subspace with one dimension less than the N-dimensional feature space in which it is formed. This hyperspace is the classification boundary, and its margin is the distance between two object classes in feature space separated by the hyperspace for the SVM classification. <sup>[111]</sup>

### 2.5.7 Model Validation

The robustness and predictability of a model are critical factors in model development. The training set is used to test the model's robustness, and the test set is used to test the model's predictability. Internal or external validation can be performed on the model. During model building, cross validation is used on the training set to internally validate if a model can correctly predict the class, and if there is any potential overfitting of data. N-Fold cross validation was used in this study, which entails removing a compound from the training set and using it as a test set while training the model on the remaining data. The process is repeated N times on different compounds in the training set, with the output being an average of the responses. This reduces overfitting, which occurs when a method learns the data extremely well but fails to predict biological activity for unknown compounds.

A confusion matrix, as shown in **Figure 2.5**, can also be used to evaluate the performance of the machine learning method for categorical classification problems. It is represented as a 2x2 matrix of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), with the rows representing actual class entries and the columns being predictive. <sup>[115]</sup>

---

	Inactive	Active
Inactive	TN	FP
Active	FN	TP

**Figure 2.5** Confusion Matrix.

---

This confusion matrix can be used to calculate sensitivity, specificity, accuracy, balanced classification rate (BCR), and Matthews Correlation Coefficient (MCC). True positives, and false negatives, are related to sensitivity, which describes the prediction accuracy for actives, whereas false positives, and false negatives are related to specificity, which describes the prediction accuracy for inactives. **Table 2.1** shows how to compute these validation parameters.

A Receiver Operating Characteristic (ROC) curve depicts a model's success and failure by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity). It demonstrates the model's level of precision as well as its ability to distinguish between actives and inactives. Cohen's Kappa ( $\kappa$ ) compares predicted classifications to known classifications to determine how well a classifier performs due to chance.

**Table 2.1** Validation Parameters

Parameter	Formula
<i>Sensitivity</i>	$\frac{TP}{TP + FN}$
<i>Specificity</i>	$\frac{TN}{TN + FP}$
<i>Accuracy</i>	$\frac{TP + TN}{N}$
<i>MCC</i>	$\frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}}$
<i>BCR</i>	$(Sensitivity + Specificity)/2$

### 2.5.8 Applicability domain

The domain applicability of a model is calculated to identify compounds in both the test set and the data base that can be accurately predicted using the Euclidean distance method, as shown in **Equation 2.14**. The applicability domain defined in **Equation 2.14** is a range that can be used to identify compounds within the database, or tests set that can be accurately predicted by a model built from a specific training set. Compounds in the database that fall within the threshold are assumed to be in the prediction range of the model built with the training set.

$$APD = 'd' + Z\sigma$$

#### Equation 2.14

Where ' $d$ ' is the average and  $\sigma$  is the standard deviation of Euclidean distances of the  $k$  nearest neighbours of each compound in the training set in the chemical descriptor space, and  $Z$  is the empirical parameter to control the significance level, with the default value of 0.5. If the distance from the external compound to its nearest neighbour in the training set is above the APD, the prediction is labelled unreliable.

## 2.6 Virtual Screening

Virtual screening is a computational technique for discovering drugs by searching drug databases for a specific disease. ZINC, <sup>[116]</sup> ChEMBL., <sup>[63]</sup> PubChem, <sup>[62]</sup> Maybridge and BindingDB are among the free compound databases available. These databases can be screened using two types of computational techniques: ligand-based virtual screening and structure-based virtual screening. The order in which the strategies are used in the virtual screening process varies; however, two important aspects are usually considered, database filtering and database screening.

### 2.6.1 Database filters

Given that databases used are typically large and contain a diverse range of compounds, filters must be used to filter out compounds with drug-like properties that can be used to design hits. Orally administered drugs are only effective if they reach their target in sufficient quantities and interact with the target to produce the desired biological response. <sup>[117]</sup> This drug disposition within the body can be summarized as absorption, distribution, metabolism, and excretion (ADME). Lipinski et al., <sup>[118, 119]</sup> investigated the physicochemical properties of common drugs and proposed the Rule of Five, a well-known probability index used as a solubility and permeability filter in drug design.

Solubility predicted by an octanol-water partition coefficient less than 5; molecular weight between 200 and 500 Da; hydrogen bonding, polar surface area, and charge described by less than 10 H-bond acceptors and 5 H-bond donors are among the properties of common drugs. <sup>[118]</sup> These characteristics have an impact on the in vivo and in vitro activity of orally active compounds. To filter out these undesirables, sub-structures of compounds that are known to be dyes, unstable, reactive, interfere with binding by forming aggregates, or toxic can be used. <sup>[120, 121]</sup> The Pan Assay Interference Compounds (PAINS) sub-structure filters are a common type of filter. Rhodanines, phenolic Mannich bases, hydroxyphenylhydrazones, alkylidene barbiturates, alkylidene heterocycles, 1,2,3-alkylpyrroles, activated benzofurazans, 2-amino-3-carbonylthiophenes, catechols, and quinones are among the sub-structures listed by Baell and Holloway. <sup>[122]</sup> Compounds that contain these sub-structures are filtered out and not included during database screening.



## **2.6.2 Database screening**

Following database filtering, either ligand-based or structure-based virtual screening can be performed. This thesis combined ligand-based and structure-based techniques. QSAR models were developed using quantitative structure activity relationships derived from descriptors of known compounds and were used to screen database compounds. Molecular docking, a structure-based virtual screening technique, was used to supplement the ligand-based virtual screening by structurally filtering screened compounds and predicting possible interactions from identified compounds. Docking allows for the rational validation of important residues within the binding site by supplying protein-ligand complexes. <sup>[21]</sup>

## **2.7 Pharmacokinetic analysis of orally available drugs**

The oral route is the most popular and preferred method of medication administration because it is non-invasive, convenient, and patient-compliant. However, factors such as solubility, mucosal permeability, and gastrointestinal tract stability must be considered before a drug can be taken orally. <sup>[123]</sup> As a result, it is essential to understand the physicochemical and biochemical characteristics that control oral availability. When a drug is administered, it has to cross the intestinal mucosa or the blood brain barrier (BBB) to reach the site of action. Drug molecules must partition between the apical and basolateral sides of the cell membrane to enter the cytoplasmic domain, and then into the systemic circulation of the brain. <sup>[124]</sup>

### **2.7.1 Barriers to drug delivery**

A drug's physicochemical properties are necessary for it to pass through the cell membrane and systemic circulation. The mode of absorption is influenced by the molecular weight (size), hydrogen bonding propensity (hydrogen bond donors and

acceptors), and cLogP (lipophilicity). Given that they are small and hydrophobic, most drugs prefer the transcellular route; however, hydrophilic drugs can be transported via carrier-mediated transporters either through the paracellular pathway or the transcellular pathway. [124]

The main biochemical inhibitors of drug absorption are efflux pumps, transporters, and enzymes involved in metabolism. Hydrochloric acid and proteolytic pepsins, which hydrolyze peptides and proteins at pH 2–5, are examples of metabolic enzymes. At pH 8, other proteolytic enzymes like chymotrypsin, elastase, carboxypeptidase A and B, and trypsin are active against both large and small peptides. [118]. Phase I enzymes from the CYP superfamily and phase II enzymes that are involved in metabolic activity can be found in the proximal small intestine. The CYP1, CYP2, and CYP3 sub-families make up the CYP superfamily. These CYP1A1, CYP2C, CYP2D6, and CYP3A4 enzymes are found in the intestine, and each has unique drug specificities. [124]

Small molecules cannot cross the transcellular pathway because efflux pumps like permeability glycoprotein (Pgp) prevent membrane partitioning of these molecules. The Pgp proteins recognize a wide variety of substrates with a wide range of structures. [124,125] Given these, and other barriers, drugs with the physicochemical properties described by Lipinski et al., [118] are better delivered orally. How the drug is partitioned is determined by the octanol-water partition coefficient, which describes the drug's affinity for the hydrophobic membrane region. As a result, drugs with low hydrogen-bonding potential have a low number of hydrogen bond donors or acceptors and a cLogP greater than 1 are better absorbed. Furthermore, the route of entry restricts the size of the drug, with substances weighing more than 600 g/mol failing to cross membranes. [124,125]

## 2.8 Antitumor Tests

The main aim of this research is to design compounds with antitumour properties. Compounds that have been designed must be tested if they exhibit the antitumour activity, and bioassay methods for assessing the antitumour activity are available. Bioassay methods are used preliminarily in drug discovery to screen for bioactive compounds. <sup>[126]</sup> They have been used to establish biological functions for bioactive compounds, <sup>[127]</sup> such as antibacterial, antitumour, antioxidant and phytotoxic properties. These methods offer the following advantages: <sup>[128,129]</sup>

- i. Economic,
- ii. Accurate,
- iii. Reliable, and
- iv. Convenient.

The *Agrobacterium tumefaciens* (*A. tumefaciens*) disc assay has been shown to be useful in testing novel compounds for the antitumour properties based on *A. tumefaciens* infection on potatoes, carrots, radish, and beets discs. <sup>[130–132]</sup> This method uses the *A. tumefaciens*, a gram-negative soil borne bacterium that is rod-shaped and virulent and is responsible for Crown Gall disease in plants. This disease causes a spongy or hard tumour to protrude from the stems and roots of woody and herbaceous plants, which might have deleterious effects on the plant. A tumour inducing plasmid (Ti-plasmid) which carries the oncology information in the bacteria is incorporated in the plant's chromosomal DNA. <sup>[133,134]</sup> When the plant is wounded, it releases phenols which activate the Ti-plasmid in *A. tumefaciens*. This initiates cell proliferation whilst blocking apoptosis. <sup>[131]</sup> This mechanism is similar in nucleic acid

content and histology to tumour formation mechanism in human and animal cancers.

[<sup>135</sup>, <sup>137</sup>] As such, the carrot disc assay was used as a pre-screening antitumour assay in this study.

## 2.9 References

- [1] J. Zhao, Y. Cao, L. Zhang, Exploring the computational methods for protein-ligand binding site prediction, *Comput. Struct. Biotechnol. J.* (2020) 1–10. <https://doi.org/10.1016/j.csbj.2020.02.008>.
- [2] J.P. Renaud, *Structural Biology in Drug Discovery: Methods, Techniques, and Practices*, 2020, John Wiley & Sons, 2020. <https://books.google.ps/books?id=DnTJDwAAQBAJ>.
- [3] R. Adams, C.L. Worth, S. Guenther, M. Dunkel, R. Lehmann, R. Preissner, Binding sites in membrane proteins – Diversity, druggability and prospects, *Eur. J. Cell Biol.* 91 (2012) 326–339. <https://doi.org/10.1016/j.ejcb.2011.06.003>.
- [4] G. Tiwari, D. Mohanty, An in Silico Analysis of the Binding Modes and Binding Affinities of Small Molecule Modulators of PDZ-Peptide Interactions, *PLoS One.* 8 (2013) 1–17. <https://doi.org/10.1371/journal.pone.0071340>.
- [5] M. Naderi, J.M. Lemoine, R.G. Govindaraj, O.Z. Kana, W.P. Feinstein, M. Brylinski, Binding site matching in rational drug design: Algorithms and applications, *Brief. Bioinform.* 20 (2019) 2167–2184. <https://doi.org/10.1093/bib/bby078>.
- [6] T.A. Halgren, Identifying and Characterizing Binding Sites and Assessing Druggability, *J. Chem. Inf. Model.* 49 (2009) 377–389.
- [7] L.H. Jones, Bioorganic & Medicinal Chemistry Expanding chemogenomic space using chemoproteomics, *Bioorg. Med. Chem.* 27 (2019) 3451–3453. <https://doi.org/10.1016/j.bmc.2019.06.022>.
- [8] M. Vass, A.J. Kooistra, S. Verhoeven, D. Gloriam, I.J.P. De Esch, C. De Graaf, A Structural Framework for GPCR Chemogenomics: What's In a Residue Number? in: A. Heifetz (Ed.), *Comput. Methods GPCR Drug Discov.*, Springer Science+Business Media, 2018: pp. 73–113.
- [9] D. Rognan, Review Chemogenomic approaches to rational drug design, *Br. J. Pharmacol.* 152 (2007) 38–52. <https://doi.org/10.1038/sj.bjp.0707307>.
- [10] N.K. Broomhead, M.E. Soliman, Can We Rely on Computational Predictions to Correctly Identify Ligand Binding Sites on Novel Protein Drug Targets? Assessment of Binding Site Prediction Methods and a Protocol for Validation of Predicted Binding Sites, *Cell Biochem. Biophys.* 75 (2016) 15–23. <https://doi.org/10.1007/s12013-016-0769-y>.
- [11] A.T.R. Laurie, R.M. Jackson, Methods for the Prediction of Protein-Ligand Binding Sites for Structure-Based Drug Design and Virtual Ligand Screening., *Curr. Protein Pept. Sci.* 7 (2006) 395–406. <https://doi.org/10.2174/138920306778559386>.
- [12] N.M. Hassan, A.A. Alhossary, Y. Mu, C.K. Kwoh, Protein-Ligand Blind Docking Using QuickVina-W with Inter-Process Spatio-Temporal Integration, *Sci. Rep.* 7 (2017) 1–13. <https://doi.org/10.1038/s41598-017-15571-7>.

- [13] T. Vreven, H. Hwang, B.G. Pierce, Z. Weng, Evaluating template-based and template-free protein protein complex structure prediction, *Brief. Bioinform.* 15 (2013) 169–176. <https://doi.org/10.1093/bib/bbt047>.
- [14] M. Jiang, Z. Li, Y. Bian, Z. Wei, A novel protein descriptor for the prediction of drug binding sites, *BMC Bioinformatics.* 20 (2019) 1–13.
- [15] A.M. Goswami, Computational analysis, structural modeling and ligand binding site prediction of Plasmodium falciparum 1-deoxy-d-xylulose-5-phosphate synthase, *Comput. Biol. Chem.* 66 (2016) 1–10. <https://doi.org/10.1016/j.compbiolchem.2016.10.010>.
- [16] G. Lanka, R. Bathula, M. Dasari, S. Nakkala, M. Bhargavi, G. Somadi, S.R. Potlapally, G. Lanka, R. Bathula, M. Dasari, S. Nakkala, M. Bhargavi, G. Somadi, Structure-based identification of potential novel inhibitors targeting FAM3B (PANDER) causing type 2 diabetes mellitus through virtual screening, *J. Recept. Signal Transduct.* (2019) 1–11. <https://doi.org/10.1080/10799893.2019.1660897>.
- [17] G. Macari, D. Toti, F. Polticelli, Computational methods and tools for binding site recognition between proteins and small molecules: from classical geometrical approaches to modern machine learning strategies, *J. Comput. Aided. Mol. Des.* 33 (2019) 887–903. <https://doi.org/10.1007/s10822-019-00235-7>.
- [18] Y. Loewenstein, D. Raimondo, O.C. Redfern, J. Watson, D. Frishman, M. Linial, C. Orengo, J. Thornton, A. Tramontano, Protein function annotation by homology-based inference., *Genome Biol.* 10 (2009) 207. <https://doi.org/10.1186/gb-2009-10-2-207>.
- [19] I. Halperin, D.S. Glazer, S. Wu, R.B.B. Altman, The FEATURE framework for protein function annotation: Modeling new functions, improving performance, and extending to novel applications, *BMC Genomics.* 9 (2008) 1–14. <https://doi.org/10.1186/1471-2164-9-S2-S2>.
- [20] B. Zhao, S. Hu, X. Li, F. Zhang, Q. Tian, W. Ni, An efficient method for protein function annotation based on multilayer protein networks, *Hum. Genomics.* 10 (2016) 1–15. <https://doi.org/10.1186/s40246-016-0087-x>.
- [21] D. Ghersi, R. Sanchez, Beyond structural genomics: computational approaches for the identification of ligand binding sites in protein structures, *J Struct Funct Genomics.* 12 (2011) 109–117. <https://doi.org/10.1038/jid.2014.371>.
- [22] T. Simões, D. Lopes, S. Dias, F. Fernandes, J. Pereira, J. Jorge, C. Bajaj, A. Gomes, Geometric Detection Algorithms for Cavities on Protein Surfaces in Molecular Graphics: A Survey, *Comput. Graph. Forum.* 36 (2017) 643–683. <https://doi.org/10.1111/cgf.13158>.
- [23] A.T.R. Laurie, R.M. Jackson, Structural bioinformatics Q-SiteFinder: An energy-based method for the prediction of protein – ligand binding sites, *Bioinformatics.* 21 (2005) 1908–1916. <https://doi.org/10.1093/bioinformatics/bti315>.
- [24] X.-Y. Meng, H.-X. Zhang, M. Mezei, M. Cui, Molecular Docking: A powerful approach for structure-based drug discovery, *Curr. Comput. Aided Drug Des.* 7 (2011) 146–157. <https://doi.org/10.1038/jid.2014.371>.

- [25] H.S. Lee, Y. Zhang, BSP-SLIM: A blind low-resolution ligand-protein docking approach using predicted protein structures, (2012) 93–110. <https://doi.org/10.1002/prot.23165>.
- [26] R. Dias, W. Filgueira, D. Azevedo, Molecular Docking Algorithms, *Curr. Drug Targets*. 9 (2009) 1040–1047. <https://doi.org/10.2174/138945008786949432>.
- [27] G.M. Morris, R. Huey, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell, A.J. Olson, AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility, *J. Comput. Chem.* 30 (2010) 2785–2791. <https://doi.org/10.1002/jcc.21256.AutoDock4>.
- [28] R.G. Coleman, M. Carchia, T. Sterling, J.J. Irwin, B.K. Shoichet, Ligand Pose and Orientational Sampling in Molecular Docking, *PLoS One*. 8 (2013) 1–19. <https://doi.org/10.1371/journal.pone.0075992>.
- [29] X. Du, Y. Li, Y. Xia, S. Ai, J. Liang, P. Sang, X. Ji, Insights into Protein – Ligand Interactions: Mechanisms, Models, and Methods, *Int. J. Mol. Sci.* 17 (2016) 1–34. <https://doi.org/10.3390/ijms17020144>.
- [30] A. Aleksandrov, H. Myllykallio, Advances and challenges in drug design against tuberculosis: application of in silico approaches, *Expert Opin. Drug Discov.* 14 (2019) 1–12. <https://doi.org/10.1080/17460441.2019.1550482>.
- [31] H.K. Tai, S.A. Jusoh, S.W.I. Siu, Chaos-embedded particle swarm optimization approach for protein-ligand docking and virtual screening, *J. Cheminform.* 10 (2018) 1–13. <https://doi.org/10.1186/s13321-018-0320-9>.
- [32] P.H.M. Torres, A.C.R. Sodero, P. Jofily, Key Topics in Molecular Docking for Drug Design, *Int. J. Mol. Sci.* 20 (2019) 1–29.
- [33] J.E. Jones, On the Determination of Molecular Fields II. From the Equation of State of a Gas., *Proc R Soc L. A.* 106 (1924) 463–477.
- [34] J. An, M. Totrov, R. Abagyan, Pocketome via Comprehensive Identification and Classification of Ligand Binding Envelopes, *Mol. Cell. Proteomics*. 4 (2005) 752–761. <https://doi.org/10.1074/mcp.M400159-MCP200>.
- [35] Y. Su, A. Zhou, X. Xia, W. Li, Z. Sun, Quantitative prediction of protein – protein binding affinity with a potential of mean force considering volume correction, *Protein Sci.* 18 (2009) 2550–2558. <https://doi.org/10.1002/pro.257>.
- [36] O. Trott, A.J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading., *J. Comput. Chem.* 31 (2010) 455–61. <https://doi.org/10.1002/jcc.21334>.
- [37] R. Abagyan, M. Totrov, D. Kuznetsov, A New Method for Protein Modeling and Design: Applications to Docking and Structure Prediction from the Distorted Native Conformation, *J. Comput. Chem.* 15 (1994) 488–506.
- [38] N.S. Pagadala, K. Syed, J. Tuszynski, Software for molecular docking: a review, *Biophys. Rev.* 9 (2017) 91–102. <https://doi.org/10.1007/s12551-016-0247-1>.
- [39] G. Bottegoni, I. Kufareva, M. Totrov, R. Abagyan, Four-dimensional Docking: a Fast and Accurate Account of Discrete Receptor Flexibility in Ligand Docking, *J Med Chem.* 52 (2009) 397–406. <https://doi.org/10.1038/jid.2014.371>.

- [40] G. Buchbauer, A. Klinsky, P. Weiß-greier, P. Wolschann, Ab initio Molecular Electrostatic Potential Grid Maps for Quantitative Similarity Calculations of Organic Compounds, *J. Mol. Model.* 6 (2000) 425–432.
- [41] A.L. Hopkins, C.R. Groom, The druggable genome, *Nat. Rev. Drug Discov.* 1 (2002) 727–730.
- [42] K.A. Loving, A. Lin, A.C. Cheng, Structure-Based Druggability Assessment of the Mammalian Structural Proteome with Inclusion of Light Protein Flexibility, *PLOS Comput. Biol.* 10 (2014) 1–13. <https://doi.org/10.1371/journal.pcbi.1003741>.
- [43] A.C. Cheng, R.G. Coleman, K.T. Smyth, Q. Cao, P. Souldard, D.R. Caffrey, A.C. Salzberg, E.S. Huang, Structure-based maximal affinity model predicts small-molecule druggability, *Nat. Biotechnol.* 25 (2007) 71–75. <https://doi.org/10.1038/nbt1273>.
- [44] R.P. Sheridan, V.N. Maiorov, M.K. Holloway, W.D. Cornell, Drug-like Density: A Method of Quantifying the “Bindability” of a Protein Target Based on a Very Large Set of Pockets and Drug-like Ligands from the Protein Data Bank, *J. Chem. Inf. Model.* 50 (2010) 2029–2040.
- [45] D. D. Kshatresh, K. T. Rakesh, P.O. Rajendra, Recent Advances in Protein – Ligand Interactions: Molecular Dynamics Simulations and Binding Free Energy, *Curr. Comput. Aided Drug Des.* 9 (2013) 518–531.
- [46] D.V.A.N.D.E.R. Spoel, E. Lindahl, B. Hess, G. Groenhof, GROMACS: Fast, Flexible, and Free, *J Comput Chem.* 26 (2005) 1701–1718. <https://doi.org/10.1002/jcc.20291>.
- [47] B. Frantzdale, S. J. Plimpton, M.S. Shephard, Software components for parallel multiscale simulation: an example with LAMMPS, *Eng. Comput.* 26 (2010) 205–211. <https://doi.org/10.1007/s00366-009-0156-z>.
- [48] S. Kumar, C. Huang, G. Zheng, E. Bohm, A. Bhatele, J.C. Phillips, H. Yu, L. V Kale, A. Authors, T. Views, Scalable molecular dynamics with NAMD on the IBM Blue Gene / L system, *IBM J. Res. Dev.* 52 (2008) 177–188. <https://doi.org/10.1147/rd.521.0177>.
- [49] A. Hynninen, M.F. Crowley, New Faster CHARMM Molecular Dynamics Engine, *J Comput Chem.* 35 (2014) 406–413. <https://doi.org/10.1002/jcc.23501>.
- [50] J.M. Bower, D. Beeman, M. Hucka, The GENESIS Simulation System, in: *Handb. Brain Theory Neural Networks*, 2000: pp. 1–21.
- [51] R. Salomon-ferrer, D.A. Case, R.C. Walker, An overview of the Amber biomolecular simulation package, *WIREs Comput Mol Sci.* 00 (2012) 1–13. <https://doi.org/10.1002/wcms.1121>.
- [52] J. Lee, X. Cheng, S. Jo, A.D. Mackerell, J.B. Klauda, I. Wonpil, CHARMM-GUI Input Generator for NAMD, Gromacs, Amber, Openmm, and CHARMM/OpenMM Simulations using the CHARMM36 Additive Force Field, *Biophysj.* 110 (2016) 641a. <https://doi.org/10.1016/j.bpj.2015.11.3431>.
- [53] A.W. Sousa Da Silva, W.F. Vranken, ACPYPE - AnteChamber PYthon Parser interfacE, *BMC Res. Notes.* 5 (2012) 1–8. <https://doi.org/10.1186/1756-0500-5->



[367](#).

- [54] J. Wang, W. Wang, P.A. Kollman, D.A. Case, Automatic atom type and bond type perception in molecular mechanical calculations, *J. Mol. Graph. Model.* 25 (2006) 247–260. <https://doi.org/10.1016/j.jmgm.2005.12.005>.
- [55] J. Gasteiger, M. Marsili, A new model for calculating atomic charges in molecules, *Tetrahedron*. 34 (1978) 3181–3184.
- [56] J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman, D.A. Case, Development and Testing of a General Amber Force Field, *J Comput Chem.* 25 (2004) 1157–1174.
- [57] H. Matter, Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors, *J. Med. Chem.* 40 (1997) 1219–1229. <https://doi.org/10.1021/jm960352+>.
- [58] J. Skolnick, H. Zhou, M. Gao, Are predicted protein structures of any value for binding site prediction and virtual ligand screening? *Curr. Opin. Struct. Biol.* 23 (2013) 1–7. <https://doi.org/10.1016/j.sbi.2013.01.009>.
- [59] M. Duran-Frigola, L. Siragusa, E. Ruppin, X. Barril, G. Cruciani, P. Aloy, Detecting similar binding pockets to enable systems polypharmacology, *PLoS Comput. Biol.* 13 (2017) 1–18. <https://doi.org/10.1371/journal.pcbi.1005522>.
- [60] R. Dole, R. Cimler, Variable Elimination Approaches for Data- Noise Reduction in 3D QSAR Calculations, (2015). <https://doi.org/10.1007/978-3-319-23485-4>.
- [61] S. Kausar, A.O. Falcao, An automated framework for QSAR model building, *J. Cheminform.* 10 (2018) 1–23. <https://doi.org/10.1186/s13321-017-0256-5>.
- [62] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B.A. Shoemaker, P.A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E.E. Bolton, PubChem 2019 update: improved access to chemical data, 47 (2019) 1102–1109. <https://doi.org/10.1093/nar/gky1033>.
- [63] A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J.P. Overington, ChEMBL: A large-scale bioactivity database for drug discovery, *Nucleic Acids Res.* 40 (2012) 1100–1107. <https://doi.org/10.1093/nar/gkr777>.
- [64] A. Mauri, V. Consonni, R. Todeschini, Molecular descriptors, in: *Handb. Comput. Chem.*, 2017: pp. 2065–2093. <https://doi.org/10.1007/978-3-319-27282-5>.
- [65] A. Cherkasov, E.N. Muratov, D. Fourches, A. Varnek, I. Igar, M. Cronin, J. Dearden, P. Gramatica, Y.C. Martin, V. Consonni, V.E. Kuz, R. Cramer, QSAR Modeling: Where have you been? Where are you going to? *J Med Chem.* 57 (2015) 4977–5010. <https://doi.org/10.1021/jm4004285.QSAR>.
- [66] D. Fourches, E. Muratov, A. Tropsha, Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research, *J Chem Inf Model.* 50 (2011) 1189–1204. <https://doi.org/10.1021/ci100176x.Trust>.
- [67] B. Viira, T. Gendron, D.A. Lanfranchi, S. Cojean, D. Horvath, G. Marcou, A. Varnek, L. Maes, U. Maran, P.M. Loiseau, E. Davioud-charvet, In Silico Mining for Antimalarial Structure-Activity Antimalarial Curcuminoids, *Molecules.* 21

- (2016) 1–18. <https://doi.org/10.3390/molecules21070853>.
- [68] A. Tropsha, Best Practices for QSAR Model Development, Validation, and Exploitation, *Mol. Inform.* 29 (2010) 476–488. <https://doi.org/10.1002/minf.201000061>.
- [69] A. Golbraikh, E. Muratov, D. Fourches, A. Tropsha, Dataset Modelability by QSAR, *J Chem Inf Model.* 54 (2014) 1–8. <https://doi.org/10.1007/s10955-011-0269-9>. Quantifying.
- [70] W. Shoombuatong, V. Prachayasittikul, N. Anuwongcharoen, N. Songtawee, T. Monnor, S. Prachayasittikul, V. Prachayasittikul, C. Nantasenamat, Navigating the chemical space of dipeptidyl peptidase-4 inhibitors, *Drug Des. Devel. Ther.* 9 (2015) 4515–4549.
- [71] T. Hoffmann, M. Gastreich, The next level in chemical space navigation: going far beyond enumerable compound libraries, *Drug Discov. Today.* 24 (2019) 1148–1156. <https://doi.org/10.1016/j.drudis.2019.02.013>.
- [72] A. Lin, D. Horvath, V. Afonina, G. Marcou, J.L. Reymond, A. Varnek, Mapping of the Available Chemical Space versus the Chemical Universe of Lead-Like Compounds, *ChemMedChem.* 13 (2018) 540–554. <https://doi.org/10.1002/cmdc.201700561>.
- [73] M. Vogt, How do we optimize chemical space navigation? *Expert Opin. Drug Discov.* 15 (2020) 523–525. <https://doi.org/10.1080/17460441.2020.1730324>.
- [74] N. Frimayanti, M.L. Yam, H.B. Lee, R. Othman, Validation of Quantitative Structure-Activity Relationship (QSAR) Model for Photosensitizer Activity Prediction, *Int. J. Mol. Sci.* 2 (2011) 8626–8644. <https://doi.org/10.3390/ijms12128626>.
- [75] T. Puzyn, A. Mostrag-Szlichtyng, A. Gajewicz, M. Skrzyński, A.P. Worth, Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models, *Struct. Chem.* 22 (2011) 795–804. <https://doi.org/10.1007/s11224-011-9757-4>.
- [76] T.M. Martin, P. Harten, D.M. Young, E.N. Muratov, A. Golbraikh, H. Zhu, A. Tropsha, Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? *J. Chem. Inf. Model.* 52 (2012) 2570–2578.
- [77] A. Rácz, D. Bajusz, K. Héberger, Consistency of QSAR models: Correct split of training and test sets, ranking of models and performance parameters, SAR QSAR *Environ. Res.* 26 (2015) 683–700. <https://doi.org/10.1080/1062936X.2015.1084647>.
- [78] R. Todeschini, V. Consonni, P. Gramatica, Chemometrics in QSAR, in: S. Brown, R. Tauler, R. Walczak (Eds.), *Compr. Chemom. Chem. Biochem. Data Anal.*, Elsevier, Oxford, 2009: pp. 129–172.
- [79] P.P. Roy, J.T. Leonard, K. Roy, Exploring the impact of size of training sets for the development of predictive QSAR models, *Chemom. Intell. Lab. Syst.* 90 (2008) 31–42. <https://doi.org/10.1016/j.chemolab.2007.07.004>.
- [80] V. Consonni, R. Todeschini, Molecular Descriptors, in: T. Puzyn (Ed.), *Recent Adv. QSAR Stud.*, Springer Science+Business Media, 2010: pp. 29–102.

- [https://doi.org/10.10007/978-1-4020-9783-6\\_3](https://doi.org/10.10007/978-1-4020-9783-6_3).
- [81] C.W. Yap, Software News and Updates PaDEL-Descriptor: An Open-Source Software to Calculate Molecular Descriptors and Fingerprints, *J. Comput. Chem.* 32 (2011) 174–182. <https://doi.org/10.1002/jcc>.
- [82] S. Beisken, T. Meinl, B. Wiswedel, L.F. De Figueiredo, M. Berthold, KNIME-CDK: Workflow-driven cheminformatics, *BMC Bioinformatics*. 14 (2013) 1–4.
- [83] A. Mauri, V. Consonni, M. Pavan, R. Todeschini, DRAGON software: An easy approach to molecular descriptor calculations, *MATCH Commun. Math. Comput. Chem.* 56 (2006) 237–248.
- [84] J. Sadowski, J. Gasteiger, G. Klebe, Comparison of automatic three-dimensional model builders using 639 X-ray structures, *J. Chem. Inf. Comput. Sci.* 34 (1994) 1000–1008.
- [85] J. Dong, D.S. Cao, H. Y. Miao, S. Liu, B.C. Deng, Y. H. Yun, N.N. Wang, A.P. Lu, W. Bin Zeng, A.F. Chen, ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation, *J. Cheminform.* 7 (2015) 1–10. <https://doi.org/10.1186/s13321-015-0109-z>.
- [86] K. Liu, J. Feng, S.S. Young, PowerMV: A software environment for molecular viewing, descriptor generation, data analysis and hit evaluation, *J. Chem. Inf. Model.* 45 (2005) 515–522. <https://doi.org/10.1021/ci049847v>.
- [87] N.M. O’Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: An Open chemical toolbox, *J. Cheminform.* 3 (2011) 1–14. <https://doi.org/10.1186/1758-2946-3-33>.
- [88] I. Ponzoni, V. Sebastián-Pérez, C. Requena-Triguero, C. Roca, M.J. Martínez, F. Cravero, M.F. Díaz, J.A. Páez, R.G. Arrayás, J. Adrio, N.E. Campillo, Hybridizing Feature Selection and Feature Learning Approaches in QSAR Modeling for Drug Discovery, *Sci. Rep.* 7 (2017) 1–19. <https://doi.org/10.1038/s41598-017-02114-3>.
- [89] H. Hong, Q. Xie, W. Ge, F. Qian, H. Fang, L. Shi, Z. Su, R. Perkins, W. Tong, Mold 2, Molecular Descriptors from 2D Structures for Chemoinformatics and, *J Chem Inf Model.* 48 (2008) 1337–1344.
- [90] E. Glaab, Building a virtual ligand screening pipeline using free software: A survey, *Brief. Bioinform.* 17 (2016) 352–366. <https://doi.org/10.1093/bib/bbv037>.
- [91] P.R. Duchowicz, Linear Regression QSAR Models for Polo-Like Kinase-1 Inhibitors, *Cells.* 7 (2018) 1–11. <https://doi.org/10.3390/cells7020013>.
- [92] M.K. Gupta, R. Sagar, A.K. Shaw, Y.S. Prabhakar, CP-MLR directed QSAR studies on the antimycobacterial activity of functionalized alkenols — topological descriptors in modeling the activity, 13 (2005) 343–351. <https://doi.org/10.1016/j.bmc.2004.10.025>.
- [93] J.C. Dearden, The Use of Topological Indices in QSAR and QSPR Modeling, in: *Adv. QSAR Model. Challenges Adv. Comput. Chem. Phys.*, 2017: pp. 57–88. <https://doi.org/10.1007/978-3-319-56850-8>.
- [94] S. Kausar, A.O. Falcao, Analysis and Comparison of Vector Space and Metric

- Space Representations in QSAR Modeling, *Molecules*. 24 (2019) 1–22. <https://doi.org/10.3390/molecules24091698>.
- [95] J.B.O. Mitchell, Machine learning methods in chemoinformatics, *WIREs Comput Mol Sci*. 4 (2014) 468–481. <https://doi.org/10.1002/wcms.1183>.
- [96] P.M. Khan, K. Roy, Current approaches for choosing feature selection and learning algorithms in quantitative structure–activity relationships (QSAR), *Expert Opin. Drug Discov.* 13 (2018) 1075–1089. <https://doi.org/10.1080/17460441.2018.1542428>.
- [97] M. Dash, H. Liu, Feature selection for classification, *Intell. Data Anal.* 1 (1997) 131–156. <https://doi.org/10.3233/IDA-1997-1302>.
- [98] A. Jović, K. Brkić, N. Bogunović, A review of feature selection methods with applications, in: 2015 38th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2015 - Proc., 2015: pp. 1200–1205. <https://doi.org/10.1109/MIPRO.2015.7160458>.
- [99] B. Dwyer, Rules, in: *Syst. Anal. Synth.*, 2016: pp. 295–332. <https://doi.org/10.1016/b978-0-12-805304-1.00018-7>.
- [100] A.G. Karegowda, M.A. Jayaram, A.S. Manjunath, Feature Subset Selection Problem using Wrapper Approach in Supervised Learning, *Int. J. Comput. Appl.* 1 (2010) 13–17.
- [101] R. Kohavi, G.H. John, Wrappers for feature subset selection, *Artif. Intell.* 97 (1997) 273–324. [https://doi.org/https://doi.org/10.1016/S0004-3702\(97\)00043-X](https://doi.org/https://doi.org/10.1016/S0004-3702(97)00043-X).
- [102] M. Pirlot, General local search methods, *Eur. J. Oper. Res.* 92 (1996) 493–511. [https://doi.org/10.1016/0377-2217\(96\)00007-0](https://doi.org/10.1016/0377-2217(96)00007-0).
- [103] M. Danishuddin, A. U. Khan, Descriptors and their selection methods in QSAR analysis: Paradigm for drug design, *Drug Discov. Today*. 21 (2016) 1–12. <https://doi.org/10.1016/j.drudis.2016.06.013>.
- [104] R. Dechter, J. Pearl, Generalized Best-First Search Strategies and the Optimality of A, *J. ACM*. 32 (1985) 505–536. <https://doi.org/10.1145/3828.3830>.
- [105] N. Sanchez-Marono, A. Alonso-Betanzos, M. Tombilla-Sanroman, Filter Methods for Feature Selection – A Comparative Study, in: *Int. Conf. Intell. Data Eng. Autom. Learn.*, 2007: pp. 178–187. <https://doi.org/10.1007/978-3-540-77226-2>.
- [106] M.A. Hall, G. Holmes, Benchmarking Attribute Selection Techniques for Discrete Class Data Mining, *IEEE Trans. Knowl. Data Eng.* 15 (2003) 1437–1447.
- [107] J.A. Castillo-Garit, G.M. Casañola-Martin, F. Torrens, H. Pham-The, A. Torreblanca, S.J. Barigye, Machine learning-based models to predict modes of toxic action of phenols to *Tetrahymena pyriformis*, *SAR QSAR Environ. Res.* 28 (2017) 735–747. <https://doi.org/10.1080/1062936X.2017.1376705>.
- [108] X. Xia, E.G. Maliski, P. Gallant, D. Rogers, Classification of kinase inhibitors using a Bayesian model, *J. Med. Chem.* 47 (2004) 4463–4470. <https://doi.org/10.1021/jm0303195>.

- [109] D.W. Aha, K. Dennis, M.K. Albert, Instance-Based Learning Algorithms, *Mach. Learn.* 6 (1996) 37–66.
- [110] S.L. SALZBERG, Book Review: C4.5: by J. Ross Quinlan. Inc., 1993., in: *Mach. Learn.*, 1994: pp. 235–240. [https://doi.org/10.1016/S0019-9958\(64\)90259-1](https://doi.org/10.1016/S0019-9958(64)90259-1).
- [111] J.C. Platt, Sequential Minimal Optimization: A Fast Algorithm for Training Support, (1998) 1–21.
- [112] E. Frank, M.A. Hall, I. H. Witten, T. Weka, Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for “Data Mining: Practical Machine Learning Tools and Techniques”, *Morgan Kaufmann*, Fourth Edition, 2016., Forth, 2016.
- [113] G. Mugumbate, K.A. Abrahams, J.A.G. Cox, G. Papadatos, G. Van Westen, J. Lelièvre, S.T. Calus, N.J. Loman, L. Ballell, D. Barros, J.P. Overington, G.S. Besra, Mycobacterial dihydrofolate reductase inhibitors identified using chemogenomic methods and in vitro validation, *PLoS One.* 10 (2015) 1–11. <https://doi.org/10.1371/journal.pone.0121492>.
- [114] F. Pereira, D.A.R.S. Latino, S.P. Gaudencio, A Chemoinformatics Approach to the Discovery of Lead-Like Molecules from Marine and Microbial Sources En Route to Antitumour and Antibiotic Drugs, *Mar. Drugs.* 12 (2014) 757–778. <https://doi.org/10.3390/md12020757>.
- [115] R.N. Roy, J.; Kar, S.; Das, Feature combination networks for the interpretation of statistical machine learning models: application to Ames mutagenicity, in: *SpringerBriefs Mol. Sci.*, 2015: pp. 37–60.
- [116] J.J. Irwin, T. Sterling, M.M. Mysinger, E.S. Bolstad, R.G. Coleman, ZINC: A Free Tool to Discover Chemistry for Biology, *Chem. Inf. Model.* 52 (2012) 1757–1768. <http://zinc.docking.org/>.
- [117] A. Daina, O. Michielin, V. Zoete, SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules, *Sci. Rep.* 7 (2017) 1–13. <https://doi.org/10.1038/srep42717>.
- [118] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Adv. Drug Deliv. Rev.* 64 (2012) 4–17. <https://doi.org/10.1016/j.addr.2012.09.019>.
- [119] C.A. Lipinski, Lead profiling Lead- and drug-like compounds: the rule-of-five revolution, *Drug Discov. Today.* 1 (2004) 337–341. <https://doi.org/10.1016/j.ddtec.2004.11.007>.
- [120] S.J. Chakravorty, J. Chan, M.N. Greenwood, I. Popa-Burke, K.S. Remlinger, S.D. Pickett, D.V.S. Green, M.C. Fillmore, T.W. Dean, J.I. Luengo, R. Macarrón, Nuisance Compounds, PAINS Filters, and Dark Chemical Matter in the GSK HTS Collection, *SLAS Discov.* 23 (2018) 532–545. <https://doi.org/10.1177/2472555218768497>.
- [121] J.B. Baell, J.W.M. Nissink, Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017 - Utility and Limitations, *ACS Chem. Biol.* 13 (2018) 36–44. <https://doi.org/10.1021/acscchembio.7b00903>.



- [122] J.B. Baell, G.A. Holloway, New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays, *J. Med. Chem.* 53 (2010) 2719–2740. <https://doi.org/10.1021/jm901137j>.
- [123] M.S. Alqahtani, M. Kazi, M.A. Alsenaidy, M.Z. Ahmad, *Advances in Oral Drug Delivery*, *Front. Pharmacol.* 12 (2021). <https://doi.org/10.3389/fphar.2021.618411>.
- [124] M. Laksitorini, V. Prasasty, P.K. Kiptoo, T.J. Siahaan, Pathways and Progress in improving drug delivery through the intestinal mucosa and blood-brain barriers, *Ther. Deliv.* 7 (2016) 117–138.
- [125] P. Kiptoo, A.M. Calcagno, T.J. Siahaan, Physiological, Biochemical, and Chemical Barriers to Oral Drug Delivery, in: T.J.S. Binghe Wang, Longqin Ho (Ed.), *Drug Deliv. Princ. Appl.*, Second, John Wiley & Sons, 2016, 2016: pp. 19–34.
- [126] C.-L. Zhao, W.-I. Chik, H.-J. Zhang, Bioprospecting and bioassay-guided isolation of medicinal plants — A tool for drug discovery, in: *Evid. Based Valid. Herb. Med. Transl. Res. Bot.*, Second, 2022: pp. 511–537.
- [127] Z. Chen, J. Wang, J. Yuan, Z. Wang, Z. Tu, J. Crommen, W. Luo, J. Guo, T. Zhang, Z. Jiang, Rapid screening of neuraminidase inhibitors using an at-line nanofractionation platform involving parallel oseltamivir-sensitive/resistant neuraminidase bioassays, *J. Chromatogr. A.* 1687 (2023) 463693. <https://doi.org/10.1016/j.chroma.2022.463693>.
- [128] Monthon Lertcanawanichakul, Kittisak Chawawisit, Mode of Action and Cytotoxicity of Bioactive Compounds Produced by *Streptomyces* sp. KB1, *J. Pharm. Pharmacol.* 7 (2019) 499–503. <https://doi.org/10.17265/2328-2150/2019.09.002>.
- [129] J. Hoeksma, T. Misset, C. Wever, J. Kemmink, J. Kruijtzter, K. Versluis, R.M.J. Liskamp, G.J. Boons, A.J.R. Heck, T. Boekhout, J. den Hertog, A new perspective on fungal metabolites: identification of bioactive compounds from fungi using zebrafish embryogenesis as read-out, *Sci. Rep.* 9 (2019) 1–16. <https://doi.org/10.1038/s41598-019-54127-9>.
- [130] Z. Chen, F. Xiong, A. Yu, G. Lai, Aptamer biorecognition-triggered DNAzyme liberation and Exo III-assisted target recycling for ultrasensitive homogeneous colorimetric bioassay of kanamycin antibiotic, *Chem. Commun.* 55 (2019) 3959–3962. <https://doi.org/10.1039/c8cc10107h>.
- [131] J.R. White, M. Abodeely, S. Ahmed, G. Debaube, E. Johnson, D.M. Meyer, N.M. Mozier, M. Naumer, A. Pepe, I. Qahwash, E. Rocnik, J.G. Smith, E.S.E. Stokes, J.J. Talbot, P.Y. Wong, Best practices in bioassay development to support registration of biopharmaceuticals, *Biotechniques.* 67 (2019) 126–137. <https://doi.org/10.2144/btn-2019-0031>.
- [132] O.M. Alshehri, S. Alshamrani, M.H. Mahnashi, M.M. Alshahrani, J.A. Khan, M. Shah, M.A. Alshehri, R. Zafar, M. Zahoor, M.S. Jan, S.S.U. Hassan, A. Sadiq, Phytochemical Analysis, Total Phenolic, Flavonoid Contents, and Anticancer Evaluations of Solvent Extracts and Saponins of *H. digitata*, *Biomed Res. Int.*

- 2022 (2022). <https://doi.org/10.1155/2022/9051678>.
- [133] S. Waghulde, N. Gorde, T. Baviskar, P. Patil, S. Singh, M.K. Kale, V.R. Patil, Cumulative Cytotoxicity Assay of the Aqueous and Ethanolic Extracts of the Selected Medicinal Plants Using Crown Gall Tumour Disc Bioassay, *Chem. Proc.* 3 (2021) 1–5. <https://doi.org/10.3390/ecsoc-24-08297>.
- [134] M.H. Mahnashi, Y.S. Alqahtani, B.A. Alyami, A.O. Alqarni, F. Ullah, A. Wadood, A. Sadiq, A. Shareef, M. Ayaz, Cytotoxicity, anti-angiogenic, anti-tumour and molecular docking studies on phytochemicals isolated from *Polygonum hydropiper* L., *BMC Complement. Med. Ther.* 21 (2021) 1–14. <https://doi.org/10.1186/s12906-021-03411-1>.
- [135] O. Babich, S. Sukhikh, A. Pungin, S. Ivanova, L. Asyakina, A. Prosekov, Modern Trends in the In Vitro Production and Use of Callus, Suspension Cells and Root Cultures of Medicinal Plants, *Molecules.* 25 (2020) 1–18. <https://doi.org/10.3390/MOLECULES25245805>.
- [136] N. Rahimian, H.R. Miraei, A. Amiri, M.S. Ebrahimi, J.S. Nahand, H. Tarrahimofrad, M.R. Hamblin, H. Khan, H. Mirzaei, Plant-based vaccines and cancer therapy: Where are we now and where are we going? *Pharmacol. Res.* 169 (2021) 105655. <https://doi.org/10.1016/j.phrs.2021.105655>.
- [137] C.I. Ullrich, R. Aloni, M.E.M. Saeed, W. Ullrich, T. Efferth, Comparison between tumours in plants and human beings: Mechanisms of tumour development and therapy with secondary plant metabolites, *Phytomedicine.* 64 (2019) 153081. <https://doi.org/10.1016/j.phymed.2019.153081>.

### **3 Analysis of binding sites and ligand induced conformation of nicastrin and binding modes of its inhibitors**

#### **3.1 Introduction**

Three binding sites in nicastrin were identified using molecular docking and ICM PocketFinder in this chapter. The DLID, a druggability assessment, was used to evaluate the identified pockets. The identified sites are located around nicastrin signature regions and include the DYIGS, hinge, and tetratricopeptide repeat-like (TPR-like) domains. Molecular dynamic simulations were used to confirm the stability and conformation of the ligand in the most favourable site, the DYIGS site. To identify residues in the DYIGS site that are important in nicastrin inhibition, a per residue decomposition analysis was performed. A docking analysis revealed the binding modes of nicastrin-inhibiting gamma-secretase inhibitors.

#### **3.2 Methods**

##### **3.2.1 Protein preparation**

The three-dimensional structures of the gamma-secretase complex (PDB ID 6IDF) <sup>[1]</sup> and (PDB ID 5A63) <sup>[2]</sup> were retrieved from the Protein Data Bank and the nicastrin coordinates from Chain A were extracted from these structures. Two nicastrin structures were used to select the best conformer to use in binding mode analysis. Using AutoDock Tools, <sup>[3]</sup> the nicastrin structures were prepared for docking calculations by adding Gasteiger charges, merging non-polar hydrogens, assigning the correct AutoDock4 atom types, and adding hydrogen atoms. The prepared Nicastrin structures were saved in pdbqt format.

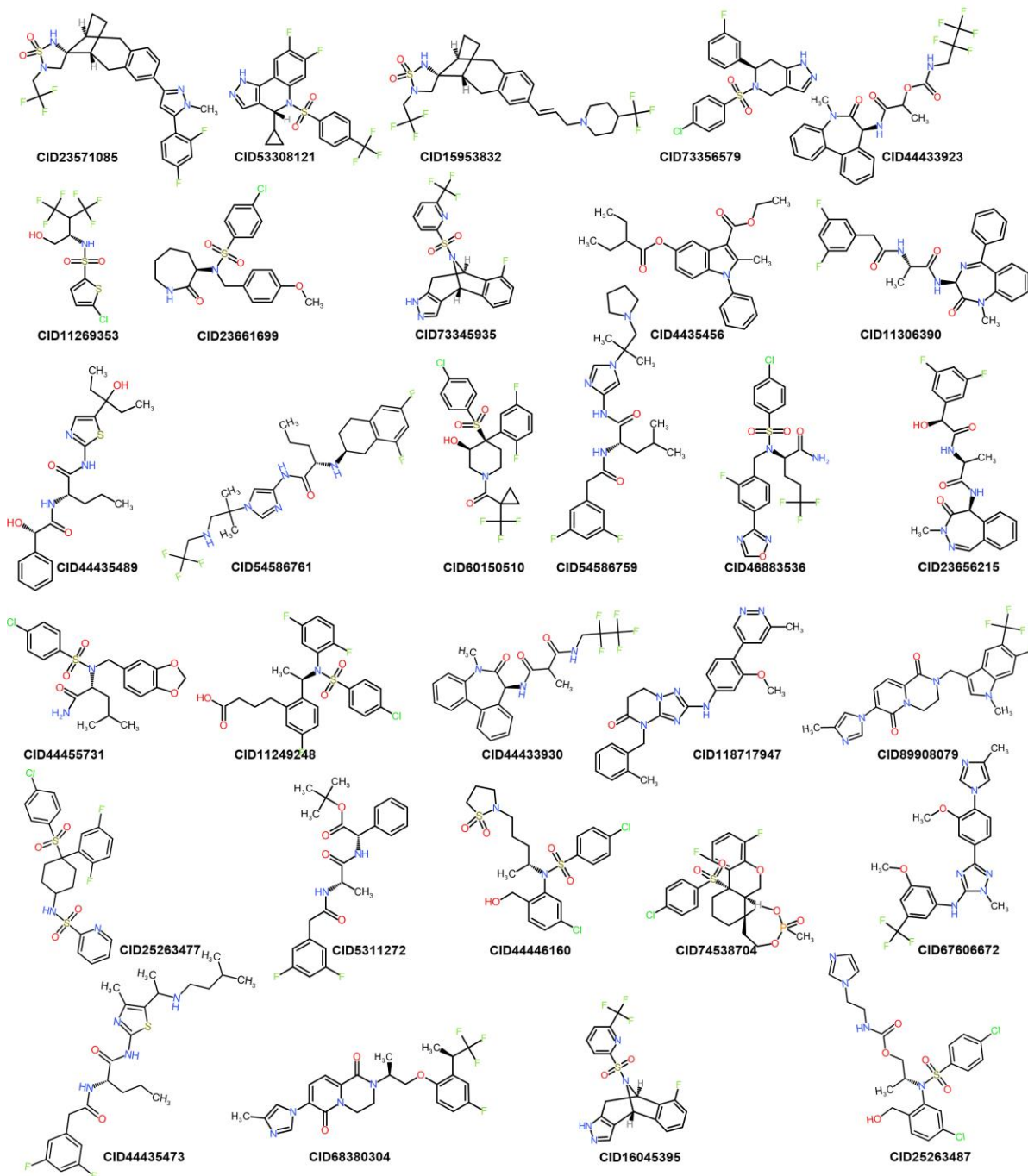


### 3.2.2 Ligand collection and curation

A dataset of human gamma secretase inhibitors (536 in total) with nicastrin bioactivity expressed as IC<sub>50</sub> values was retrieved from the PubChem database.<sup>[4]</sup> Inhibitors with a bioactivity outcome labelled *Inconclusive* and *Unspecified* impair the ability of derived models to predict bioactivity and, as a result, were removed from the dataset. Duplicates and salts were also removed as part of the curation process. All IC<sub>50</sub> values were normalised by converting them to molar units and then logarithmically transforming them to pIC<sub>50</sub> (-logIC<sub>50</sub>) values. The inhibitory potencies of the data set, expressed as pIC<sub>50</sub>, ranged from 4.3 to 11.7 and compounds with a pIC<sub>50</sub> ≥8.0 were classified as actives.

### 3.2.3 Ligand preparation for docking

The nicastrin compound dataset had a lot of common substructures and in order to select a minimal number of compounds for docking, and reduce computation time, hierarchical clustering was employed. The dataset was clustered in DataWarrior using hierarchical clustering, and 30 compounds (**Figure 3.1**) from the cluster centres were chosen for docking. The 30 ligands were prepared for docking using AutoDockTools,<sup>[3, 5]</sup> by performing energy minimisation of 200 steps using conjugate gradient and MMFF94 force field; adding Gasteiger charges, merging non-polar hydrogen atoms, assigning AutoDock4 atom types, and adding hydrogen atoms. For each ligand, the root torsion, degree of freedom, and the number of rotatable bonds were defined. The coordinates of the structures of the ligands were saved in pdbqt format.



**Figure 3.1** A set of 30 known gamma secretase inhibitors with nicastrin activity labelled using their PubChem compound identifiers.

### 3.2.4 Blind docking calculations

Blind docking was performed to identify potential binding sites in nicastrin using AutoDock Vina. For both protein conformers, 5A63 and 6IDF, the grid box was centred on the protein with a grid spacing of 0.375 Å and x, y, z dimensions of 70.66 × 122.11 × 64.21 Å and 92.34 × 67.43 × 108.94 Å, respectively. From the set of 30 ligands, ligand CID 44433923 ([1-[[[(7S)-5-methyl-6-oxo-7H-benzo[d][1]benzazepin-7-yl]amino]-1-oxopropan-2-yl]N-(2,2,3,3,3 pentafluoropropyl)carbamate) with a high nicastrin bioactivity of  $2.0 \times 10^{-6}$  μM from PubChem was selected for use in the blind docking calculations. In both cases default docking parameters were used.

### 3.2.5 Assessing druggability of the identified binding sites in nicastrin

The Internal Coordinate Mechanism (ICM) method developed by Molsoft L.L.C [6] was used to validate the predicted binding sites of the two protein structures (PDB IDs 6IDF and 5A63). The protein structures were individually prepared in ICM using receptor preparation tools by optimising hydrogen, histidine, proline, glycine, and cysteine residues. The structures were saved as ICM objects after missing hydrogens and heavy atoms were added. Potential binding pockets on the proteins were identified using ICM PocketFinder, [7] and their druggability was given by the calculated DLID score. [8] The DLID score is a metric used to evaluate binding sites where a high DLID (=1) score represents a binding site druggable by small molecules.

### **3.2.6 Molecular dynamic simulations to optimize the docked protein-ligand complexes**

Molecular dynamic simulations were performed using GROningen MAchine for Chemical Simulations (GROMACS) 2022.1 software <sup>[9]</sup> and Chemistry at Harvard Macromolecular Mechanics (CHARMM36) as an all atom forcefield to determine the stability of the docked complex as well as the intermolecular interactions over time. <sup>[10]</sup> The AnteChamber Python Parser interface (ACPYPE) portal <sup>[11]</sup> was used to generate the topology for ligand CID44433923. The complex was solvated in an octahedral TIP3P water-box with a distance of 10 Å between the box's edges and neutralized using 132K<sup>+</sup> and 120Cl<sup>-</sup> ions. This was followed by a steepest descent method used to minimize the system and set at 5000 steps. The steepest descents converged at 2368 steps when the maximum force was less than 1000 kJ/mol/nm. The system was equilibrated for 125 ps and all bonds and heavy atoms were restricted by the LINCS (Linear Constraints Solver) algorithm. The temperature and pressure were set to 310 K and 1 atm respectively and finally the system was subjected to a 50 ns production run saving the trajectories every 2 ps (.mdp file in Appendix A1).

To determine the stability of the docked complex as well as determine the intermolecular interactions over time, the root mean square deviation (RMSD), root mean square fluctuation (RMSF) and radius of gyration (RoG) were obtained from GROMACS routines for the analysis.

### **3.2.7 Calculating the free energy of binding**

The free energy of binding was calculated using molecular mechanics with generalised Born surface area solvation (MMPBSA) using gmx-mmpbsa. <sup>[12]</sup> These calculations

considered snapshots of the molecular dynamic simulations ranging from 27 to 50 ns. The binding energy calculated considered only the enthalpy of a single trajectory to minimise computing costs. A per residue energy decomposition analysis was also done to identify binding site residues that contribute to the binding energy.

### 3.2.8 Characterizing binding modes and interactions of known inhibitors in nicastrin

To characterize the binding modes and interactions of known nicastrin inhibitors, the 30 prepared ligands were docked into the DYIGS binding site of nicastrin PDBID 6IDF using the Autodock Vina tool. The coordinates of the identified DYIGS binding site in section 3.2.3 were used in setting up the grid box. The grid box for docking was centered on the protein at 171.82, 192.43, 218.77 with a 0.375 Å and x,y,z dimensions of 42.67 × 41.94 × 42.46 Å.

## 3.3 Results and Discussion

### 3.3.1 Binding sites in nicastrin

Blind docking calculations were initially performed to predict potential binding sites in nicastrin conformers. Three distinct binding sites in both conformers were identified, which are in similar locations. The identified sites (**Table 3.1**) encompass domains or signature regions within nicastrin that are specific to their function (**Figure 3.2A**).<sup>[13]</sup> These include a site with the Asp336 from the DYIGS signature (Asp336, Tyr337, Ile338, Gly339, Ser340) (DYIGS site)<sup>[14]</sup> and the TPR-like site<sup>[15]</sup> including a potential binding site positioned in a central cleft in the hinge region (Hinge region site). When compared to the other two sites, the DYIGS site had a higher affinity of -9 kcalmol<sup>-1</sup> for compound CID44433923.

**Table 3.1 Predicted binding site residues**

<b>Binding site</b>	<b>Residues</b>
DYIGS	Asn55, His58, Gln59, Ile60, Ser137, Val138, Pro141, Asn142, Asp143, Gly144, Phe145, Asn150, His158, Tyr173, Glu174, Asp175, Arg281, Arg285, Glu333, Thr334, Phe335, Asp336, Tyr337, Glu364, Leu365, Gly366, Gln367, Pro423, Pro424, Ser425, Ser426, Val440, Ala442, His444, Phe448, Tyr453, Gln454, Trp648, NAG803, NAG804, NAG805, BMA807
Hinge	Cys50, Val51, Leu53, Ile60, Gly61, Ala172, Cys248, Asp249, Arg285, Ser286, Phe287, Phe288, Trp289, Asn290, Ala292, Phe335, Tyr545, Gln552, Tyr554, Ala556, Val557, Pro560, Thr561, Asn562, Thr563, Glu650, Ser651, Arg652, Trp653
TPR-like	Thr265, Asn530, Trp533, Arg539, Gln540, Asp541, Arg543, Ser544, Leu546, Arg583, Cys586, Gln587, Pro589, Lys597, Asp598, Tyr600, Glu601, Tyr602, Ser603, Tyr604, Gln606, Leu609, Thr614, Arg616, Pro618, Arg626,

### 3.3.1.1 Tetratricopeptide repeat like site

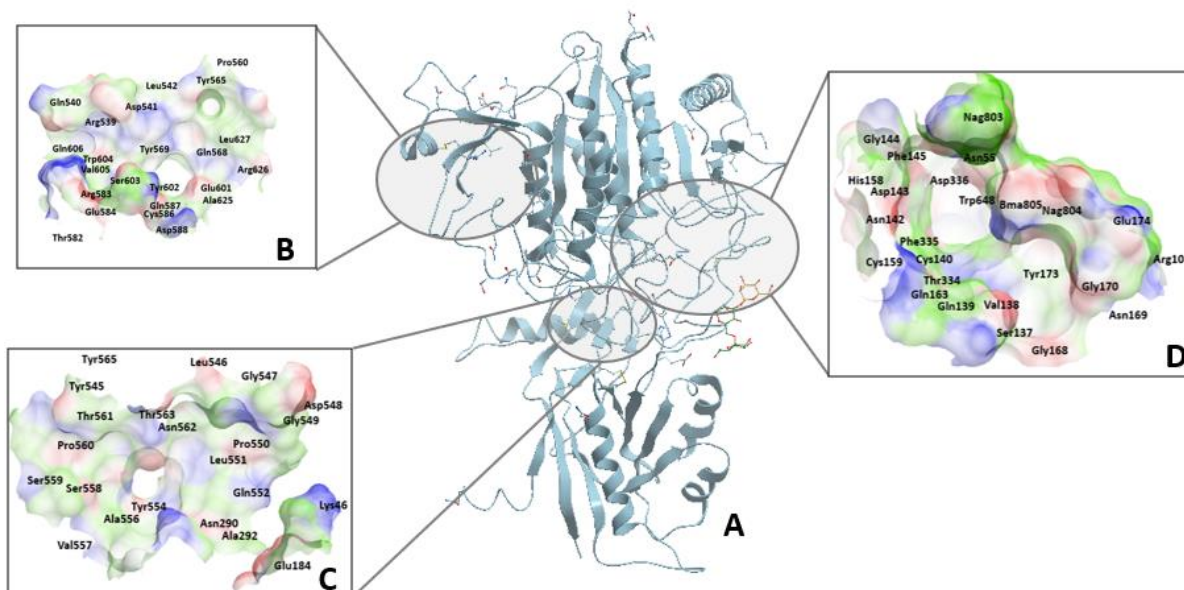
A shallow pocket located on the surface of the large lobe was predicted as a binding site and contains the TPR-like domain (**Figure 3.2B**). The TPR-like domain is homologous to TPR domain 2A and 2B helices of the HOP human protein, which binds to the C-terminal peptide of Hsp90. The TPR domains bind to substrates via side chains of  $\alpha$ -helix residues.

### 3.3.1.2 Hinge site

During substrate binding, a conserved hinge region in nicastrin homologues facilitates rotation of the large lobe relative to the small lobe. <sup>[16,17]</sup> The hinge region is made up of phenylalanine residues (Phe286 and Phe287) that interact with side chains of small lobe phenylalanine residues through van der Waals interactions. This region forms a central cleft <sup>[18]</sup> at the back of the DYIGS pocket. Blind docking studies revealed that the central cleft and surrounding residues could be a potential binding site (**Figure 3.2C**).

### 3.3.1.3 DYIGS site

The DYIGS site (**Figure 3.2D**) houses the Asp336 residue of the DYIGS motif situated in the large lobe [17,19,20]. The DYIGS residues are proposed to bind to N-termini of substrates <sup>[16]</sup> and are buried under a loop or lid (residues Ser137 to Glu168) that extends from the small lobe. <sup>[2]</sup> The orientation of the loop residues exposes just the side chain of Asp336 oriented towards the surface for interactions. During docking, ligands interact with the lid residues that line the pocket rather than interacting with the conserved residues. The presence of aromatic residues identified in the binding site such as phenylalanine (Phe145, Phe335, and Phe448), histidine (His58, His158, and His444), tryptophan (Trp648), and tyrosine (Tyr173, Tyr337, and Tyr453) is known to influence the function as well as molecular recognition in proteins. <sup>[21]</sup> Their presence in nicastrin might influence substrate recognition and recruitment. Glycans in the binding site also control the accessibility of the binding site. <sup>[22]</sup>



**Figure 3.2** Predicted binding sites in nicastrin.

The receptor pocket surface is coloured by binding property: white (neutral); green (hydrophobic surface); red (hydrogen bonding acceptor potential); blue (hydrogen bond donor potential). **A)** The binding sites are situated and hence named according to functional regions in nicastrin. These are shown as **B)** TPR-like site, **C)** Hinge site and **D)** DYIGS site all located in the large lobe of the protein.

### 3.3.2 Binding site characterization and druggability assessment

The geometry and physicochemical properties of the binding site are important in determining its quality, the nature and size of the potential drug-like ligands. Volume, surface area and buriedness of the binding site are the descriptors that characterize geometry and druggability of a binding site. These descriptors correspond to the shape and size of small molecule binders of that site. [7,23,24] Physicochemical properties of the site are important since they complement the drug-like nature of the small molecules. The geometry and physicochemical properties as well as the druggability of the binding pocket that determines the ability to bind to small drug-like molecules,



were assessed using ICM PocketFinder. [7] ICM PocketFinder generates a DLID score, which is a druggability assessment that determines the binding site's ability to interact with drug-like compounds. A positive DLID score close to 1 indicates that a site is likely to be targeted by drug-like molecules. [8]

PDB ID 5A63, the apo form of the gamma-secretase complex, and PDB ID 6IDF in complex with a notch fragment, were used in the druggability assessment. **Table 3.2** shows the three distinct sites identified through blind docking studies, along with a summary of their druggability indicated by the DLID score. The geometric differences in these sites in the conformers could be explained by ligand-induced changes in nicastrin, as observed by Bolduc et al., [17] When a ligand binds to the gamma-secretase complex's transmembrane domain, the hinge region (residue Phe287) in nicastrin rotates, affecting the hydrogen bond network and flexibility of the ectodomain, resulting in increased binding site volume as seen in the PDB ID 6IDF conformer. The increased volume in PDB ID 6IDF can positively impact the free energy of binding ligands within the binding site.

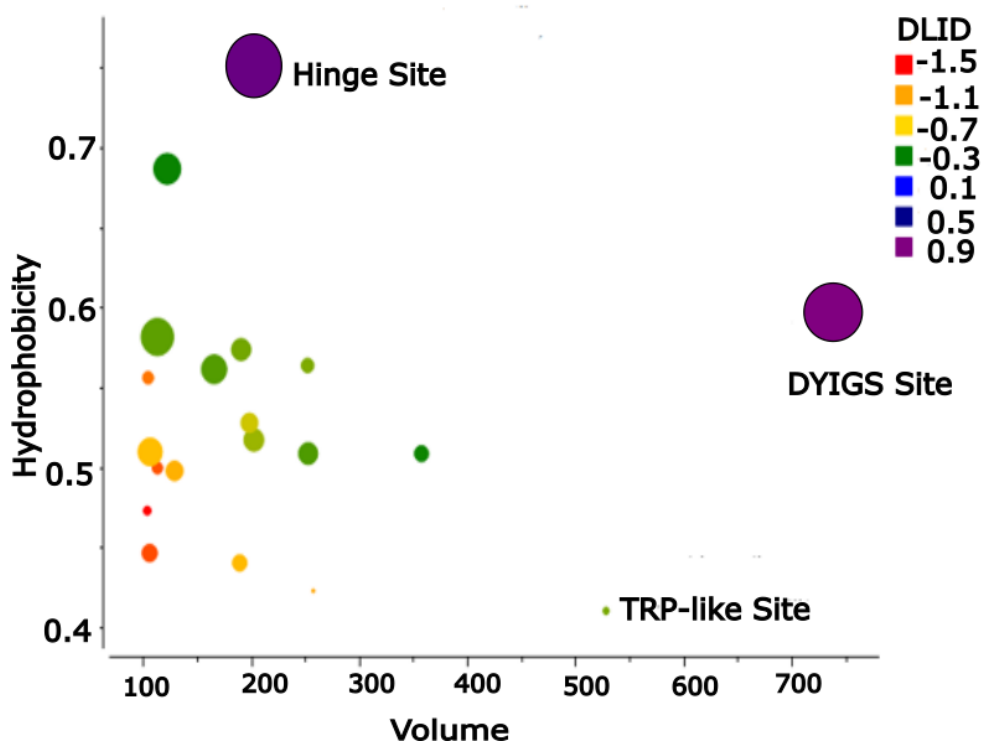
The DLID scores of two of the identified binding sites (DYIGS and Hinge) were 0.84 and 0.92, respectively, which were favourable for binding drug-like molecules. Despite having the second largest volume, the TPR-like site had a DLID score of -0.48 and a negative DLID score indicates that highly polar molecules are preferentially bound. Although the Hinge site had a higher DLID score than the DYIGS site, its volume of 200 Å<sup>3</sup> (**Table 3.2**) was less than the average volume of most ligands in the data set of 250 Å<sup>3</sup>, and thus, binding modes of ligands were not assessed in this site. Given that the volume of the ligand is known to be correlated to the binding site volume, with the ligand rarely occupying the entire binding site, so the ligands will not effectively interact with the binding site residues.

**Table 3.2** Druggability assessment of nicastrin binding sites in two different conformers

	Structure	DLID score	Volume/Å <sup>3</sup>	Buriedness	Hydrophobicity	Aromaticity
<b>DYIGS site</b>	6IDF	0.84	738.73	0.86	0.59	0.06
	5A63	0.38	535.40	0.82	0.53	0.04
<b>Hinge site</b>	6IDF	0.92	200.14	1.00	0.75	0.10
	5A63	0.54	129.00	1.00	0.75	0.07
<b>TPR-like site</b>	6IDF	-0.48	527.88	0.68	0.41	0.00
	5A63	-0.34	550.20	0.66	0.48	0.02

Given that nicastrin (PDB ID 6IDF) was the most druggable conformer due to its better geometric properties, its druggability landscape was evaluated (**Table 3.2**). The druggability landscape compared volume, hydrophobicity, and buriedness amongst other attributes that define a druggable binding site. In the druggability landscape, the volume was plotted against hydrophobicity and coloured according to buriedness in **Figure 3.3**, with larger dots indicating a druggable binding site. The druggability landscape depicts the twenty pockets identified by ICM PocketFinder, including the DYIGS, TPR-like, and hinge sites predicted by blind docking. Only two of the twenty binding sites, the DYIGS site and hinge site were predicted to be druggable by drug-like molecules.

The druggability landscape demonstrates that the druggability of binding sites in nicastrin increases with hydrophobicity. This is understandable given that hydrophobicity predominates free energy binding in protein-ligand interactions. [25] Considering the characteristics of the DYIGS site which is covered by a hydrophobic lid and the hinge site, which is surrounded by hydrophobic residues like Phe103, Leu171, Phe176, and Ile180, this greatly contributes to their hydrophobicity. [16]



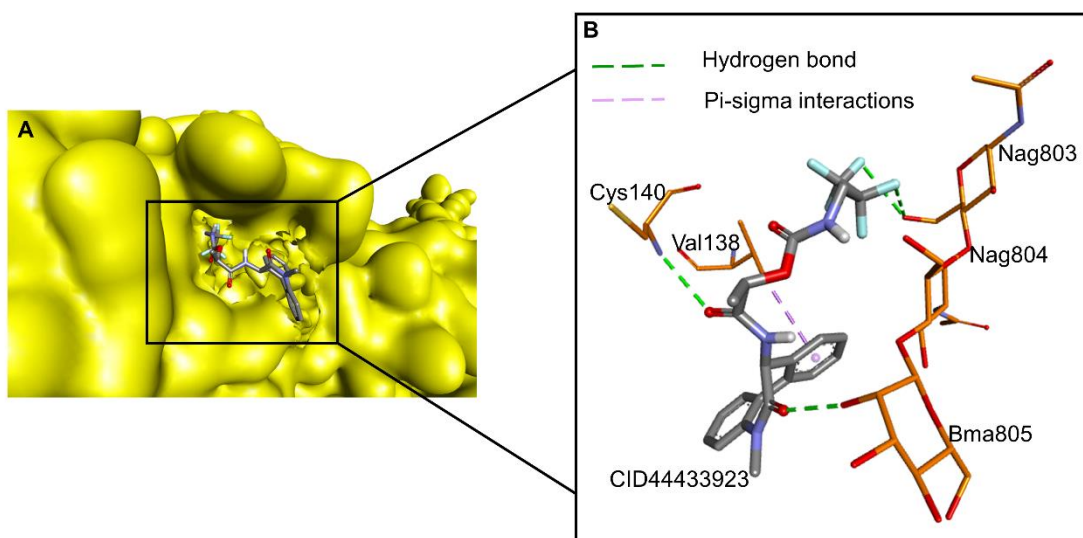
**Figure 3.3** Correlation between druggability of binding sites in nicastrin and hydrophobicity.

Volume and buriedness of pocket. Predicted binding sites are represented by dots coloured according to their DLID scores with a positive high DLID score (purple) denoting a very druggable site and a negative DLID (red) score describing a site that is very difficult to target using drug-like molecules.

### 3.3.3 Mechanism of nicastrin ligand binding

Molecular dynamic simulations were used to determine the stability of the docked complex and to better understand intermolecular interactions over time. The docked complex of compound CID44433923 in the DYIGS site (

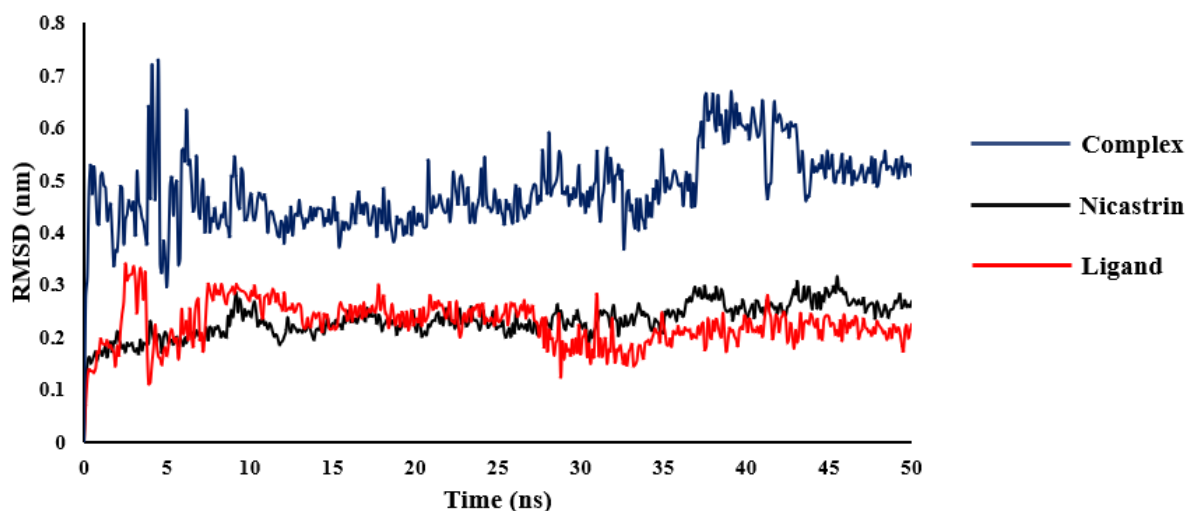
**Figure 3.4)** shows hydrogen bonds between Cys140 and oxygen on the amino group (3.2 Å) and BMA805 and the oxo group on the benzazepine (2.7 Å). Val138 established pi-sigma interactions with the pi electrons of the benzo group. To determine the stability of the docked complex, the root mean square deviation (RMSD), root mean square fluctuation (RMSF) and radius of gyration (RoG) were calculated using gromacs routines.



**Figure 3.4** Surface of nicastrin (yellow) with compound CID44433923.

**A.** Surface of nicastrin (yellow) with compound CID44433923 in stick representation and coloured by atom. **B.** The interactions between compound CID44433923 and nicastrin DYIGS binding site residues.

The RMSD characterises the overall conformational stability of the protein-ligand system by calculating changes in the protein's carbon alpha (C $\alpha$ ) and primary conformation, as well as the ligand's, over the simulation timescale. Deviations from the starting structure were noticed at the start of the simulation and after 36 ns of the simulation indicating conformational changes in nicastrin in the presence of compound CID44433923 (**Figure 3.5**). The changes could be attributed to the mobility lid domain residues in nicastrin <sup>[2]</sup> particularly Pro141, Asn142, Asp143, Gly144, Phe145 and Asn145.

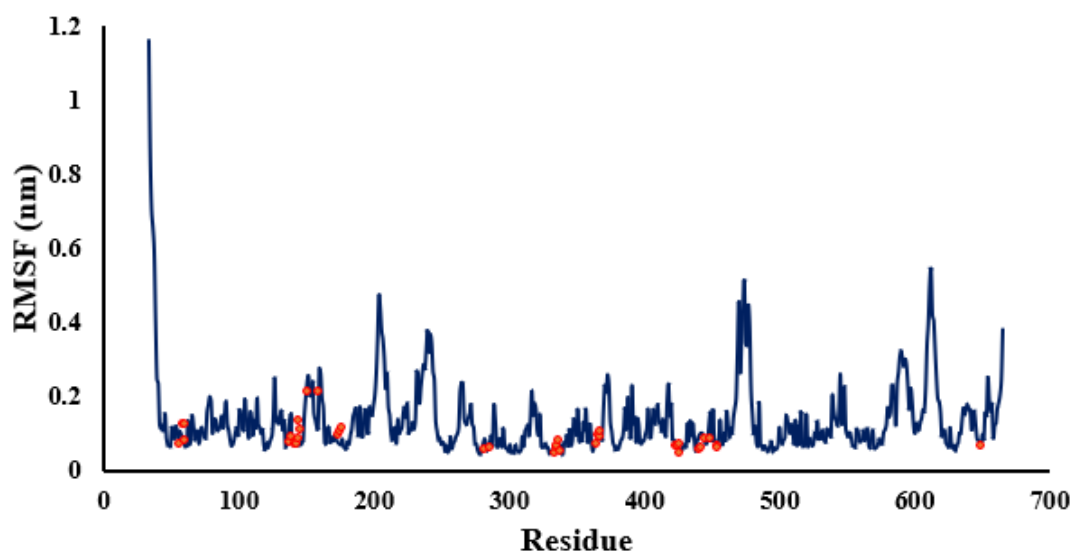


**Figure 3.5** Change in carbon alpha RMSD of nicastrin, ligand CID44433923 and complex.

Change in carbon alpha RMSD of nicastrin (black line), ligand CID44433923 (red line) and complex (blue line) plotted over 50 ns. The RMSD for nicastrin was found in the range of 0.11-0.32 nm with an average of 0.23 nm and from 0.07-0.34nm with an average of 0.22 nm for the ligand. The complex of nicastrin bound to CID44433923 had an RMSD in the range 0.27-0.72 nm with an average of 0.48 nm.

---

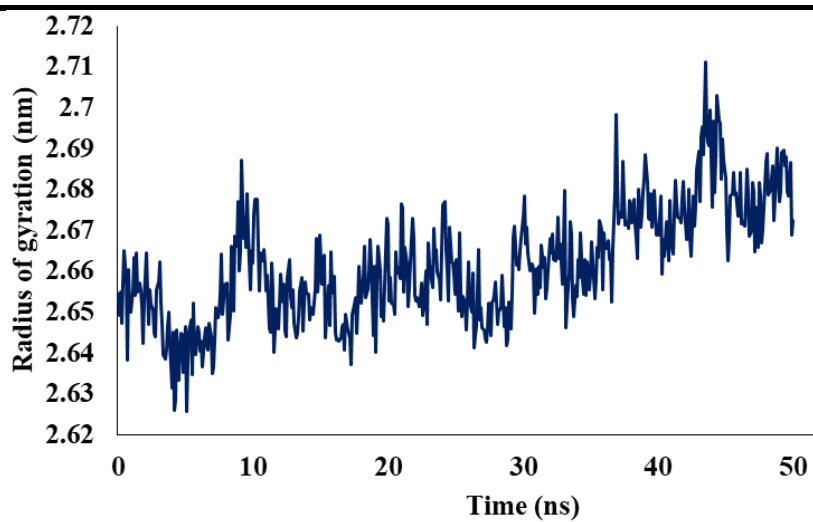
The RMSF (**Figure 3.6**) was calculated to account for the structural integrity of nicastrin when bound to the ligand. The receptor is more stable, rigid, and compact when the RMSF values per amino acid residue are low. A binding site residue RMSF value of less than 0.54 nm indicates that the ligand was stable within the binding site. The radius of gyration (**Figure 3.7**) of nicastrin also demonstrates its compactness and stability even though the lid had high fluctuations. The complex was compact, as evidenced by a range of radius of gyration of 2.65-2.67 nm with an average of 2.66 nm.



**Figure 3.6** Change in RMSF for the carbon alpha residues.

Change in RMSF for the carbon alpha residues of nicastrin bound to the ligand CID44433923 (black lines) plotted over 50 ns. The red dots indicate binding site residues identified.

---



**Figure 3.7** Radius of gyration of nicastrin carbon alpha residues of nicastrin.

---

### 3.3.4 Binding free energy calculations and per residue free energy decomposition analysis of the complex

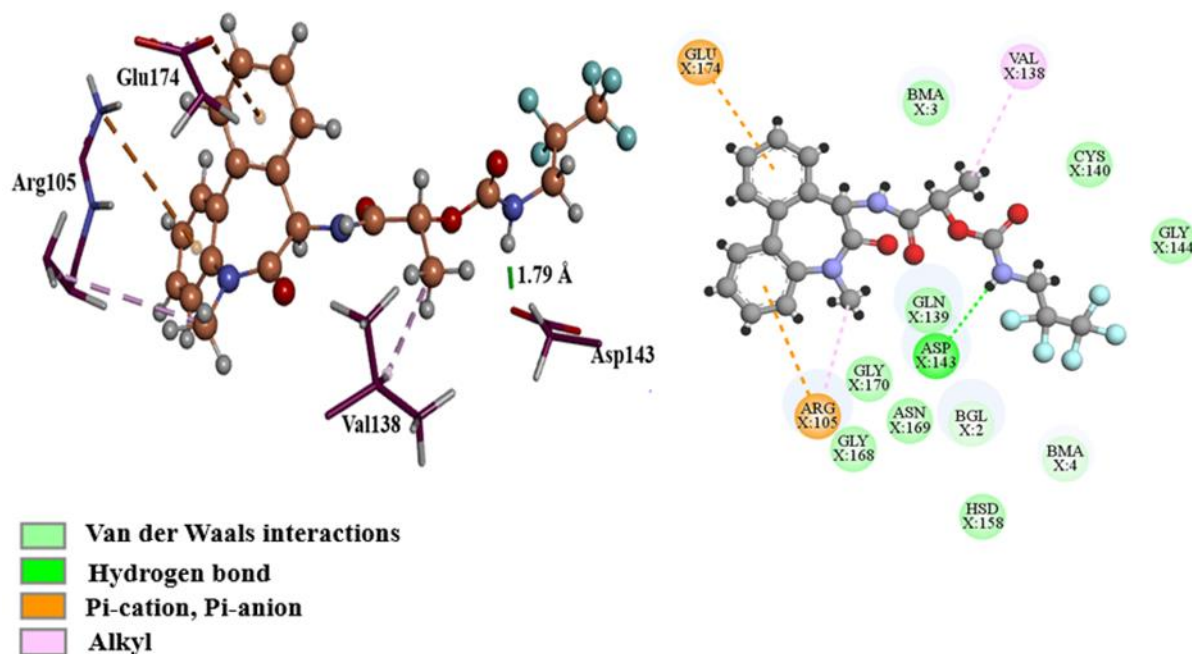
The Gibbs free energy of -11.40 kcal/mol for ligand binding to nicastrin was calculated using MM/GBSA and is shown in the **Table 3.3**. The van der Waals energy contribution was -21.24 kcal/mol and since this was the lowest energy term. It indicates that hydrophobic interactions are primarily responsible for binding free energy. The electrostatic energy, on the other hand, was -15.15 kcal/mol, indicating its importance in the binding of ligands to nicastrin in the DYIGS site.

To better understand the interactions that affect the binding's free energy, the conformation with the lowest binding energy was examined. This conformation was found at 45.6 ns with a binding energy of -19.34 kcal/mol and is presented in **Figure 3.8**. The residues Gln139, Cys140, Gly144, His158, Gly168, Asn169, Gly170, and the glycan Bma805 were involved in van der Waals interactions. Along with Val138, Arg105 had a role in hydrophobic alkyl interactions. Also, the presence of Arg105 and Glu174 encouraged electrostatic interactions, with the positively charged Arg105 creating pi-cation contacts and the negatively charged Glu174 creating pi-anion interactions with the aromatic rings. Asp143 created a strong hydrogen bond of 1.79 Å by donating a hydrogen from the amide hydrogen. Per residue free energy decomposition, using MM/GBSA, identified residues Gln139, Val138 and Arg105 as contributing to the binding energy the most, which indicates their importance in nicastrin activity. These residues were used to guide the selection of small molecules during drug design.

**Table 3.3** Energy contributions to the free energy of binding for the nicastrin-CID44433923 complex by MM-GBSA method.



Energy component	VDWaals	EEL	GSOLV	TOTAL
$\Delta$ Energy/ kcal/mol	-21.24	-15.15	24.99	-11.40

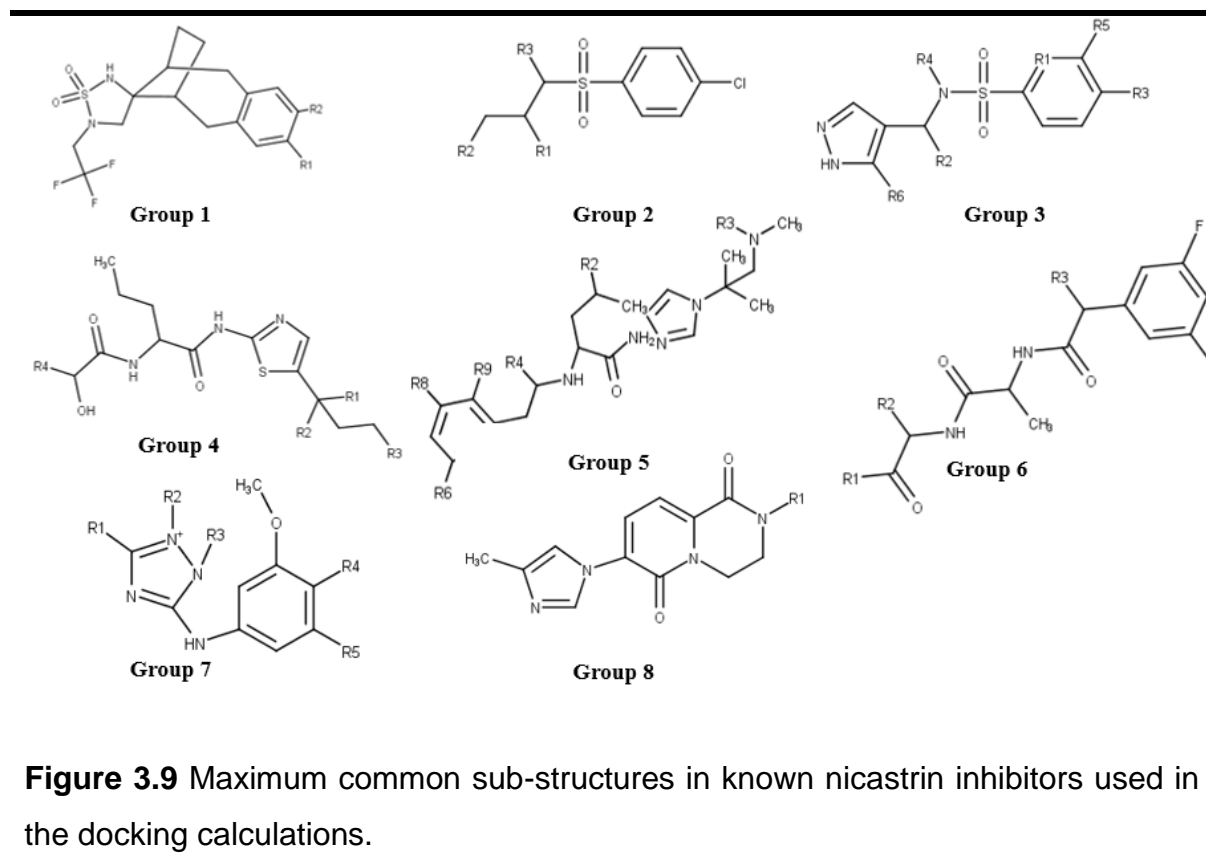


**Figure 3.8** Binding mode of CID44433923 in the DYIGS binding site with the lowest free energy of binding at 45.6 ns.

### 3.3.5 Characterization of binding modes and interactions of known inhibitors

The selected 30 ligands (**Figure 3.1**) were docked into the DYIGS site to analyse binding modes of gamma secretase inhibitors with known nicastrin activity. The compounds in the data set were first grouped based on their common substructures. **Figure 3.9** depicts the eight maximum common substructures that were obtained. The sulfonyl group is present in groups 1, 2, and 3 with the sulfonamide group linked to a phenyl ring in groups 2 and 3. Groups 4, 5, and 6 are related through the 3-aminopropanamide component within their maximum common substructures. In

general, the compounds interact with residues Val138 and Gln139 which were identified through per residue free energy decomposition analysis as contributing to the binding energy upon binding.



Compounds that bear the sulfonamide group are generally known to be effective and to have good pharmacokinetic properties. [26] In this study's diverse set of nicastrin inhibitors, the sulfonyl bearing compounds were reported to be more potent ( $IC_{50}$  values ranging from  $6.0 \times 10^{-5} \mu\text{M}$  to  $8.3 \times 10^{-3} \mu\text{M}$ ) than other compounds without the sulfonamide group. The compounds in Group 1 contains a spiro[1,2,5 thiadiazolidene-4,13'-tricyclo[8.2.1.03,8]trideca-3(8),4,6-triene]1,1-dioxide sub-structure with two compounds bearing this moiety: CID 23571085 and CID 15953832. When docked to the DYIGS site in nicastrin these compounds have similar orientation in the binding site, in which the spirocyclic sulfonamide group predicted to interact with Val138 and

Gln139 through hydrophobic interactions and hydrogen bonding between residues Tyr173 and trifluoromethyl and difluorophenyl groups in CID 15953832 and CID 23571085, respectively.

All eight compounds in Group 2 contain the 4-chlorobenzene sulfonamide moiety. As with Group 1 compounds, the sulfonyl component mostly interacts with Arg105 via hydrogen bonding; however, salt bridges between Asp 143 and Asp336 and the tertiary amine group on the sulfonamide are also observed. The orientation of Group 2 compounds differs from that of Group 1 in that the halogenated phenyl part of the substructure is mostly anchored into the binding site via hydrophobic contacts with the glycans.

Group 3 is made up of four compounds containing the methyl pyrazole conjugated to a substituted benzenesulfonamide group (CIDs 53308121, 73356579, 73345935, and 16045395). Fluoro and chloro substituents in the compounds are suggested to interact with Nag803 and Nag804 through hydrophobic interactions, similar to the orientation of Group 2 4-chlorobenzene. However, the orientation of the sulfonyl part of Group 3 interacts with Cys140. This contrasts with the interaction of sulfonyls in Groups 1 and 2 which interact with Arg105. Hydrogen bonds and hydrophobic contacts with Val138, Asp143, His158, and Bma805 were also common with the pyrazole group.

The binding mode of compounds that contain the sulfonyl moiety is elaborated by the schematic representation of the orientation of CID 15953832, a potent gamma-secretase inhibitor <sup>[1]</sup> and its interactions in the DYIGS binding site (**Figure 3.10 A and Figure 3.10 B**). Hydrophobic interactions with residues Val138, Gln139, Asn142, Asp143, Cys159, Tyr173, Asp336 and Trp648 were observed. Interactions from docking show hydrophobic contacts between the fluorine of the difluorophenyl part of

the inhibitor and the Asp336 residue, which is part of the DYIGS motif. A salt bridge between the tertiary amine group of the methylpyrazole with Asp143 was established. Hydrogen bonds are revealed between Gln163 and oxygen on the sulfonyl centre.

The 3-aminopropanamide components common in Groups 4, 5, and 6 compounds interact with residues Val138, Asp143, Nag803 and Nag804 primarily through hydrophobic interactions. The two Group 4 compounds (CIDs: 44435456 and CID44435489) have the 1,3-thiazol-2-yl]amino]-1-oxopentan-2-yl] propanamide group in their structures. The 1,3-thiazole interacts with His158 via hydrophobic contacts and the aminopropanamide with Val138. The Asp143 nitrogen forms a hydrogen bond with the oxygen on the dimethylbutanamide in CID 44435456 whilst in CID 44435489, the aspartate oxygen and Gly144 nitrogen form hydrogen bonds with phenylacetyl amino oxygens.

The common sub-structure of compounds in Group 5 is phenylacetyl amino [1-(2-methylpropan-2-yl) imidazol-4-yl] pentanamide. The phenyl group in the substructure is halogenated and electrostatic interactions between the fluorines and nitrogen of Nag804 are observed. In both compounds, the phenylacetyl amino [1-(2-methylpropan-2-yl) imidazol-4-yl] pentanamide substructure interacts with Asp143, Cys140, Bma807, His158, Bma805 and Bma806 via hydrophobic interactions. These compounds also interact with Asp336. The nitrogen on the trifluoromethylamino group forms hydrogen bonds with the oxygen in Asp336, while the fluorines form hydrogen bonds with Tyr173 and Asn142.

The 3,5-difluorophenyl-acetyl amino propanamide substructure is common to the gamma-secretase inhibitor Compound E (CID 11306390) and CID 23656215 making up Group 6. Group 6 compounds are anchored into the binding site by electrostatic

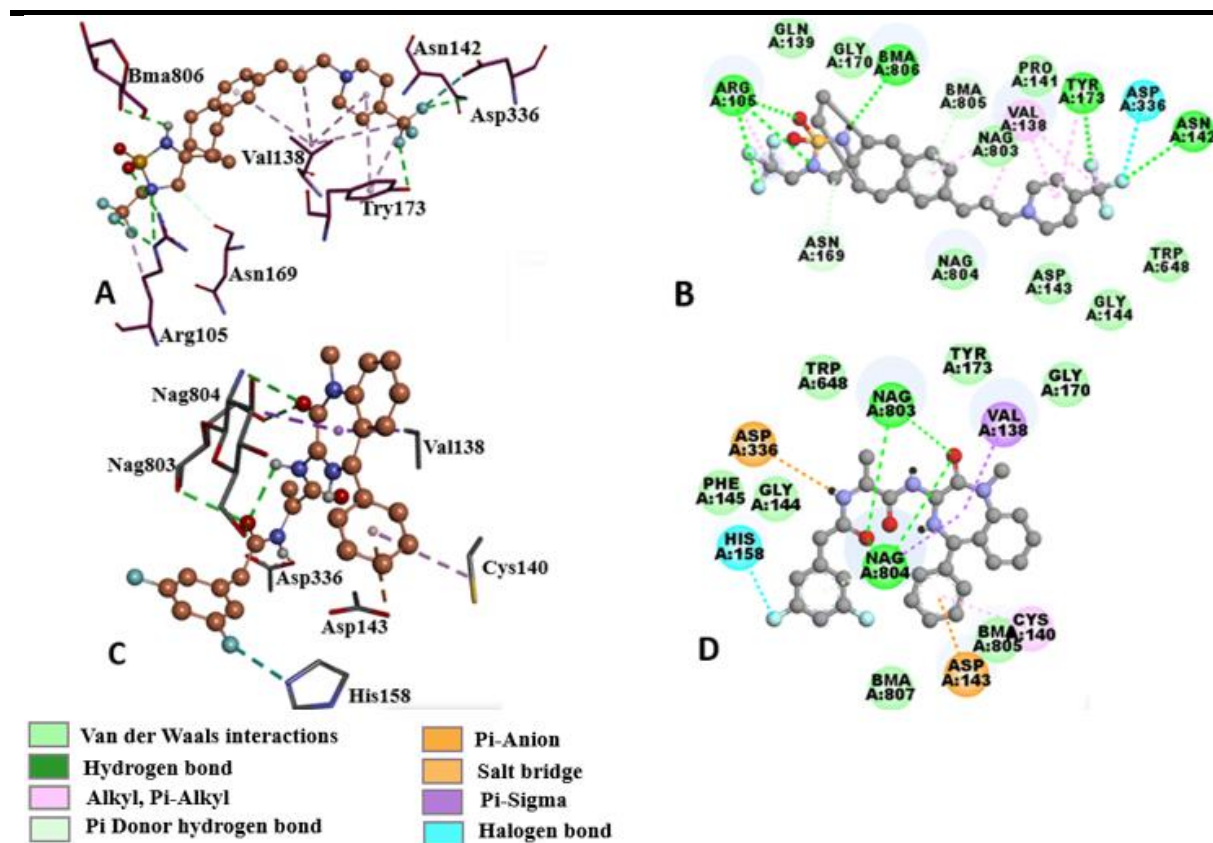
interactions between glycans Nag803 and Nag804 and propanamide oxygens. The diazepine group in both compounds interacts with Val138, whilst residues Asp143, Tyr173, Nag803 and Bma805 form hydrophobic contacts in both compounds.

The binding mode of the Groups that contain the 3-aminopropanamide moiety has been illustrated by the schematic representation of the orientation of CID 11306390, a potent L-alanine derivative <sup>[27]</sup> in the binding site (**Figure 3.10C and Figure 3.10D**). A halogen bond between His158 and the difluoro substituent was observed. Nag803 and Nag804 form hydrogen bonds with acetylamino propanamide oxygen and oxygen on the diazepine, respectively. Aps336 forms a salt bridge with the nitrogen on the acetylamino propanamide. Asp143 forms a pi-anion interaction with the phenyl group. There were also hydrophobic interactions between the phenyl-2,3-dihydro-1H-benzodiazepine and Val138, Cys140, Trp648, Tyr173, Gly170 and Gly144 and Phe145 residues.

Group 7 contains two compounds that have the 3-methoxyphenyl-2-methyl-1, 2, 4-triazol-3-amine. CID67606672 contains two methoxyphenyl groups that are positioned in such a way that one methoxyphenyl oxygen forms hydrogen bonds with Asn142 while the other methoxyphenyl forms hydrophobic contacts with His158, Bma805 and Bma807. The amino group that bridges the methoxyphenyl and triazole groups is stabilized by hydrogen bonds with Nag803 and Asp336 via hydrophobic contacts. The methoxyphenyl oxygen in CID118717947 forms a hydrogen bond with Gln163 nitrogen.

The eighth group consists of two compounds, CID89908079 and CID68380304 that have the dihydro pyrido[1,2-a]pyrazine-2,3-diol conjugated to a methyl imidazole group. Upon binding, a salt bridge is formed between tertiary amines of pyrazine-2, 3-

diol in both compounds with Asp143. CID89908079 has two of these salt bridges with Asp 143 and also with the tertiary amine of the imidazol portion with Asp336. Furthermore, a T- $\pi$  stacking exists between the phenyl group in Tyr173 and the pyrido group on the CID89908079.



**Figure 3.10 A.** Docked pose of compound CID 15953832 in the DYIGS binding site. **B.** 2D representation of binding interactions between compound CID 15953832 and DYIGS binding site residues. **C.** Docked pose of compound CID 11306390 (magenta) in the DYIGS binding site. **D.** 2D representation of binding interactions between compound CID 11306390 and DYIGS binding site residues.

### 3.4 Conclusion

This study used molecular docking and molecular dynamics to identify binding sites in nicastrin, a gamma-secretase component implicated in breast cancer and a potential drug target in cancer chemotherapy. Docking calculations identified three binding

sites; however, binding site analysis using druggability assessment identified the DYIGS site as the most favourable binding site. This site was validated by a 50 ns molecular dynamic simulation with a known inhibitor CID44433923, and free energy of binding was found to be -11.4 kcal/mol and is primarily driven by hydrophobic interactions.

A per residue decomposition analysis revealed that Gln139, Val138 and Arg105 contributed significantly to the free energy of binding. The results show that hydrophobic interactions and electrostatic forces predominate in nicastrin binding. These findings imply that these residues are important in nicastrin inhibition. Docking analysis of previously reported nicastrin inhibitors identified residues Gln139, Val138 and Asp143 as key in the interactions. This research provides an insight into the binding mechanism of small molecules and may direct drug design and development efforts towards nicastrin. In the following chapter, the DYIGS site is used in the docking studies.

### 3.5 References

- [1] G. Yang, R. Zhou, Q. Zhou, X. Guo, C. Yan, M. Ke, J. Lei, Y. Shi, Structural basis of Notch recognition by human  $\gamma$ -secretase, *Nature*. 565 (2019) 192–197. <https://doi.org/10.1038/s41586-018-0813-8>.
- [2] Bai, C X.. Yan, G. Yang, P. Lu, L. Sun, R. Zhou, S.H.W. Scheres, Y. Shi, An atomic structure of human gamma secretase, *Nature*. (2015) 1–16. <https://doi.org/10.1038/nature14892>.
- [3] G.M. Morris, R. Huey, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell, A.J. Olson, AutoDock4 and AutoDockTools4: A utomated Docking with Selective Receptor Flexibility, *J. Comput. Chem.* 30 (2010) 2785–2791. <https://doi.org/10.1002/jcc.21256.AutoDock4>.
- [4] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B.A. Shoemaker, P.A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E.E. Bolton, PubChem 2019 update: improved access to chemical data, 47 (2019) 1102–1109. <https://doi.org/10.1093/nar/gky1033>.
- [5] S.M.D. Rizvi, S. Shakil, M. Haneef, A simple click by click protocol to perform docking: Autodock 4.2 made easy for non-bioinformaticians, *EXCLI J.* 12 (2013) 830–857.
- [6] R. Abagyan, M. Totrov, D. Kuznetsov, A New Method for Protein Modeling and Design: Applications to Docking and Structure Prediction from the Distorted Native Conformation, *J. Comput. Chem.* 15 (1994) 488–506.
- [7] J. An, M. Totrov, R. Abagyan, Pocketome via Comprehensive Identification and Classification of Ligand Binding Envelopes, *Mol. Cell. Proteomics.* 4 (2005) 752–761. <https://doi.org/10.1074/mcp.M400159-MCP200>.
- [8] R.P. Sheridan, V.N. Maiorov, M.K. Holloway, W.D. Cornell, Drug-like Density: A Method of Quantifying the “ Bindability ” of a Protein Target Based on a Very Large Set of Pockets and Drug-like Ligands from the Protein Data Bank, *J. Chem. Inf. Model.* 50 (2010) 2029–2040.
- [9] M.J. Abraham, T. Murtola, R. Schulz, S. Páll, J.C. Smith, B. Hess, E. Lindah, Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers, *SoftwareX.* 1–2 (2015) 19–25. <https://doi.org/10.1016/j.softx.2015.06.001>.
- [10] B.R. Brooks, C.L. Brooks, A.D. Mackerell, L. Nilsson, R.J. Petrella, B. Roux, Y. et al, CHARMM: The biomolecular simulation program, *J. Comput. Chem.* 30 (2009) 1545–1614. <https://doi.org/10.1002/jcc.21287>.
- [11] A.W. Sousa Da Silva, W.F. Vranken, ACPYPE - AnteChamber PYthon Parser interfacE, *BMC Res. Notes.* 5 (2012) 1–8. <https://doi.org/10.1186/1756-0500-5-367>.
- [12] M.S. Valdés-Tresanco, and E.M. Mario E. Valdés-Tresanco, Pedro A. Valiente, gmx\_MMPBSA: A New Tool to Perform End-State Free Energy Calculations with GROMACS, *J. Chem. Theory Comput.* 17 (2021) 6281–6291.
- [13] W.P. Esler, W.T. Kimberly, B.L. Ostaszewski, T.S. Diehl, C.L. Moore, J. Tsai, T.



- Rahmati, W. Xia, D.J. Selkoe, M.S. Wolfe, Transition-state analogue inhibitors of  $\gamma$ -secretase bind directly to presenilin-1, *Nat. Cell Biol.* 2 (2000) 428–434.
- [14] D.M. Bolduc, M.S. Wolfe, Structure of nicastrin unveils secrets of  $\gamma$ -secretase, *Proc. Natl. Acad. Sci.* 111 (2014) 14643–14644. <https://doi.org/10.1073/pnas.1416637111>.
- [15] X. Zhang, R.J. Hoey, G. Lin, A. Koide, B. Leung, K. Ahn, G. Dolios, Identification of a tetratricopeptide repeat-like domain in the nicastrin subunit of  $\gamma$ -secretase using synthetic antibodies, *PNAS.* 109 (2012) 8534–8539. <https://doi.org/10.1073/pnas.1202691109>.
- [16] T. Xie, C. Yan, R. Zhou, Y. Zhao, L. Sun, G. Yang, P. Lu, D. Ma, Y. Shi, Crystal structure of the  $\gamma$ -secretase component nicastrin, *Proc. Natl. Acad. Sci.* 111 (2014) 13349–13354. <https://doi.org/10.1073/pnas.1414837111>.
- [17] D.M. Bolduc, D.R. Montagna, Y. Gu, D.J. Selkoe, M.S. Wolfe, Nicastrin functions to sterically hinder  $\gamma$ -secretase – substrate interactions driven by substrate transmembrane domain, *Proc. Natl. Acad. Sci. U. S. A.* 113 (2016) 509–518. <https://doi.org/10.1073/pnas.1512952113>.
- [18] J.Y. Lee, Z. Feng, X. Xie, I. Bahar, Allosteric Modulation of Intact  $\gamma$ -Secretase Structural Dynamics, *Biophysj.* 113 (2017) 2634–2649. <https://doi.org/10.1016/j.bpj.2017.10.012>.
- [19] S. Shah, S.-F. Lee, K. Tabuchi, Y.-H. Hao, C. Yu, Q. LaPlant, H. Ball, C.E. Dann, T. Südhof, G. Yu, Nicastrin Functions as a  $\gamma$ -Secretase-Substrate Receptor, *Cell.* 122 (2005) 435–447. <https://doi.org/10.1016/j.cell.2005.05.022>.
- [20] Y. Hu, Y. Ye, M.E. Fortini, Nicastrin is required for  $\gamma$ -secretase cleavage of the Drosophila Notch receptor, *Dev. Cell.* 2 (2002) 69–78. [https://doi.org/10.1016/S1534-5807\(01\)00105-8](https://doi.org/10.1016/S1534-5807(01)00105-8).
- [21] K.M. Makwana, R. Mahalakshmi, Implications of aromatic – aromatic interactions: From protein structures to peptide models, *Protein Sci.* 24 (2015) 1920–1933. <https://doi.org/10.1002/pro.2814>.
- [22] Y. Gao, X. Luan, J. Melamed, I. Brockhausen, Role of Glycans on Key Cell Surface Receptors That Regulate Cell Proliferation and Cell Death, 2021. <https://doi.org/10.3390/cells10051252>.
- [23] G. Macari, D. Toti, F. Polticelli, Computational methods and tools for binding site recognition between proteins and small molecules: from classical geometrical approaches to modern machine learning strategies, *J. Comput. Aided. Mol. Des.* 33 (2019) 887–903. <https://doi.org/10.1007/s10822-019-00235-7>.
- [24] N.K. Broomhead, M.E. Soliman, Can We Rely on Computational Predictions To Correctly Identify Ligand Binding Sites on Novel Protein Drug Targets? Assessment of Binding Site Prediction Methods and a Protocol for Validation of Predicted Binding Sites, *Cell Biochem. Biophys.* 75 (2016) 15–23. <https://doi.org/10.1007/s12013-016-0769-y>.
- [25] P.W. Snyder, J. Mecinović, D.T. Moustakas, S.W. Thomas, M. Harder, E.T. Mack, M.R. Lockett, A. Héroux, W. Sherman, G.M. Whitesides, Mechanism of the hydrophobic effect in the biomolecular recognition of arylsulfonamides by

- carbonic anhydrase, *Proc. Natl. Acad. Sci. U. S. A.* 108 (2011) 17889–17894.  
<https://doi.org/10.1073/pnas.1114107108>.
- [26] L.E. Keown, I. Collins, L.C. Cooper, T. Harrison, A. Madin, J. Mistry, M. Reilly, M. Shaimi, C.J. Welch, E.E. Clarke, H.D. Lewis, J.D.J. Wrigley, J.D. Best, F. Murray, M.S. Shearman, Novel Orally Bioavailable gamma secretase Inhibitors with Excellent in Vivo Activity, *J. Med. Chem.* 52 (2009) 3441–3444.
- [27] C.C. Shelton, L. Zhu, D. Chau, L. Yang, R. Wang, H. Djaballah, H. Zheng, Y. Li, Modulation of gamma secretase specificity using small molecule allosteric inhibitors, *PNAS.* 106 (2009) 20228–20233.

## 4 Chemical Space Analysis and Virtual screening for Nicastrin Inhibitors

### 4.1 Introduction

Despite the fact that the nicastrin has been identified as an important target for the treatment of breast cancer, clinical trials for drugs that target the gamma-secretase and nicastrin have made little progress. Therefore, in this study, nicastrin inhibitors as potential hits for breast cancer therapy, were identified using chemogenomic approaches. These include chemical space analysis of the nicastrin inhibitors that were retrieved from the PubChem database. Using the chemical space data, structure-based virtual screening was performed to identify alternative and more potent nicastrin inhibitors. The binding modes and interactions of a diverse dataset of these inhibitors into the identified DGYIS binding sites (Chapter 3) were investigated.

### 4.2 Methods

#### 4.2.1 Preparation of compound datasets

The dataset of human gamma-secretase inhibitors was collected as described in section 3.2.2 of this thesis. From this dataset 192 compounds with a  $pIC_{50} \geq 8.0$  were assigned as active nicastrin inhibitors, and the 161 compounds with a  $pIC_{50} \leq 8.0$  were considered to be inactive.

The second dataset contained 84 FDA approved breast cancer drugs retrieved from the National Cancer Institute (NCI) database. The Maybridge HitCreator™ diverse set (<http://www.maybridge.com>) was used to collect the third dataset of 14 400 drug-like compounds and finally a dataset of 3105 FDA-approved drugs was collected from

Selleckchem.com (<https://www.selleckchem.com/screening/fda-approved-drug-library>). Curation of all datasets involved removing duplicates, fragments, salts and sugars followed by standardization, and neutralization using available KNIME nodes such as Connectivity, RDKit from molecule, RDKit salt stripper, Element filter and RDKit optimize geometry. The toxicity of the compounds was predicted using toxicity risk alerts for mutagenicity, tumourigenicity, irritant, and reproductive effects in DataWarrior <sup>[1]</sup> and compounds with high toxicity risk alerts were removed from the list.

Furthermore, compounds with substructures that can interfere with binding by the formation of aggregates or compounds that are toxic by using Pan Assay Interference Compounds (PAINS) were removed using substructure filters in KNIME. 480 PAINS smart strings exported from RDKit were used in a KNIME workflow as substructure queries to search the compounds set based on the list by Baell and Holloway <sup>[2]</sup> that include rhodanines, phenolic Mannich bases, hydroxyphenylhydrazones, alkylidene barbiturates, alkylidene heterocycles, 1,2,3-aralkylpyrroles, activated benzofurazans, 2-amino-3-carbonylthiophenes, catechols and quinones amongst others. The curated datasets separately contained 353 nicastrin inhibitors, 2314 FDA approved drugs, 34 FDA breast cancer drugs, and 14000 compounds in the Maybridge HitCreator set.

## **4.2.2 Navigating the chemical space**

### **4.2.2.1 Physicochemical property space**

To determine the chemical space of the nicastrin compounds, the physicochemical property and scaffold space of the compounds were determined. The physicochemical properties include molecular weight (MW) less than 500, hydrogen bond donors (HBD)

less than 5, hydrogen bond acceptors (HBA) less than 10 but greater than 5, cLogP less than 5, number of rotatable bonds less than 10, and a topological polar surface area of less than 140 Å.<sup>2</sup> [3, 4] These characteristics define the compounds' size, hydrogen bond propensity, lipophilicity, flexibility, and polarity. Similarly, the physicochemical properties of the FDA-approved breast cancer drugs were determined, and a comparison was performed to the approved FDA drugs, and the Maybridge HitCreator set.

#### 4.2.2.2 Profiling the drug likeness of the compounds

The compounds in the nicastrin, approved FDA drugs, FDA breast cancer drugs, and Maybridge datasets were classified based on the five medicinal chemistry rules: drug-like, [3] extended drug-like, [4] lead-like, [5] fragment-like, [6] and PPI-like rules. [7] **Table 4.1** shows the physicochemical properties conditions used to formulate the rules.

**Table 4.1.** Molecular properties and conditions used to categorise compounds in datasets for drug discovery

Medicinal chemistry rules	Physicochemical property conditions
Drug-like	MW ≤ 500 Da, LogP ≤ 5 & HBD ≤ 5 & HBA ≤ 10
Extended drug-like	Drug-like rules, RotB ≤ 10, TPSA ≤ 140Å
Lead-like	MW ≤ 350 Da, LogP ≤ 3, HBD ≤ 3, HBA ≤ 3
Fragment-like	HBA ≥ 3 & MW ≤ 300 & HBD ≤ 3 & LogP ≤ 3
PPI-like	NRing ≥ 4, MW ≥ 400 Da, HBA ≥ 4, LogP ≥ 4

#### 4.2.2.3 Scaffold analysis

The diversity of the datasets can be defined by analysing the scaffolds of compounds within the datasets. Scaffolds are the core structures of organic compounds that is linked to their functional groups. [8] In this study, DataWarrior was used to generate Murcko scaffolds by removing all side chains that are attached to the central rings of compounds. The scaffold diversity of each dataset was assessed by calculating the fractions of scaffolds (Ns) relative to the total number of compounds in the dataset (M) as (Ns/M), proportion of singleton scaffolds (Nss) to the total number of compounds in the dataset as (Nss/M) and singleton fractions to the total scaffolds in the datasets as (Nss/Ns).

#### 4.2.2.4 Activity cliff analysis

An activity cliff analysis was performed to identify minimum structural changes in the compounds that have a drastic effect on biological activity to associate structural features with the biological activity of nicastrin compounds. The structure activity landscape index (SALI) was calculated based on Skeletonsphere descriptors implemented in DataWarrior to associate the varying biological activity with the compounds with similar structures (**Equation 4.1**). [1]

$$SALI = \frac{|A_i - A_j|}{1 - sim(i,j)}$$

**Equation 4.1**

where  $A_i$  is the activity of the  $i$ th molecule and  $A_j$  is the activity of the  $j$ th molecule in the nicastrin dataset, and  $sim(i,j)$  is the similarity quotient among the pair of molecules.

#### 4.2.3 Quantitative Structural Activity Relationship (QSAR) prediction using Machine Learning approaches

To generate the quantitative structure activity relationship (QSAR) models, the 353 gamma-secretase compounds with nicastrin inhibition collected from Pubchem were used in model generation. The OCHEM webserver (available at <https://ochem.eu/>) was used to calculate descriptors. In this study descriptors (1) with zero variances or more than 95% zero values, (2) that were constant for all molecules, and (3) that had a relative standard deviation of 0.001, were removed. To select descriptor subsets, the best first search method was used. The best first search method finds the best descriptor in the descriptor space by making local changes to the current subset, or else it returns to a previous subset with more promising descriptors and continues the exploration. The predictive capabilities of descriptors within the subset were assessed using either the correlation-based feature selection (CFS) or a 10-fold cross-validation loop (Wrapper) with training algorithms to identify the best method for descriptor selection. The subset with the lowest correlation among the descriptors and the highest correlation with activity was chosen. WEKA version 3.8.2 <sup>[9]</sup> was used to select descriptors, evaluate them, and build machine learning models.

The compounds were then split into training and test sets where the training set was formed by 282 compounds (80%) and the test set had 71 compounds (remaining 20% of the dataset). Machine learning based quantitative structure activity relationship (QSAR) models were generated using a number of popular techniques such as trees (J48), <sup>[10]</sup> Naïve Bayesian (NB), <sup>[11]</sup> nearest neighbour classifier, IB1 <sup>[12]</sup> and sequential minimisation optimisation, SMO <sup>[13]</sup> algorithms. To ensure that the developed models

can be used to predict activity and reliability, the predictive performance of the models was evaluated both internally and externally <sup>[14, 16]</sup>: based on Sensitivity, specificity, accuracy, balanced classification rate (BCR), and Matthews correlation coefficient (MCC). These parameters are calculated using the confusion matrix's true positive (TP), false negative (FN), false positive (FP), and true negative (TN) rates. <sup>[12]</sup>

In addition to these parameters, the Receiver Operating Characteristic curve (ROC) and Cohen's Kappa ( $\kappa$ ) are used in model validation. The ROC curve depicts a model's success and failure by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity). It demonstrates the model's level of precision as well as its ability to distinguish between actives and inactives and compares the predicted classification to known classifications to determine how well a classifier performs due to chance.

#### **4.2.4 Ligand based virtual screening using generated QSAR models**

During the production stage, all the 353 nicastrin inhibitors with their calculated descriptors were used to generate quantitative structure activity relationship (QSAR) models to screen the Maybridge HitFinder<sup>TM</sup> diverse and drug-like screening set for potential nicastrin inhibitors.

#### **4.2.5 Applicability Domain**

The applicability domain of the QSAR models was calculated in order to identify compounds in the test and Maybridge screening sets that could be predicted accurately using the chosen nicastrin inhibitors. The domain similarity node in KNIME was used to compute Euclidean distances between test set compounds and their nearest neighbours in the training set, as well as Maybridge set compounds and their nearest neighbours in the production set. A threshold was established from the



average distances. [17, 18, 19] The filtered Maybridge set, obtained from the applicability domain node, was then used in structure-based virtual screening. The Euclidean distance method is defined in **Equation 4.2**:

$$APD = d' + Z\sigma$$

**Equation 4.2**

#### **4.2.6 Structure based virtual screening of the Maybridge screening set**

Structure-based virtual screening was carried out using docking calculations in AutoDock Vina of the compounds extracted from the ligand-based virtual screening. The interactions and binding energy of the four most active gamma-secretase nicastrin inhibitors in the DYIGS binding site guided the selection of potential nicastrin inhibitors from the Maybridge screening set. Using AutoDock Tools, [20] nicastrin (chain A) from the gamma-secretase PDB ID: 6IDF was prepared by adding Gasteiger charges and merging non-polar hydrogens. AutoDock4 atom types were assigned, and hydrogen atoms were added. The grid box was centred at x, y, z coordinates of 171.82, 192.43, 218.78 with a default spacing of 0.375 Å and x, y, z grid dimensions of 42.67 × 41.94 × 42.46 Å.

All ligands were energy minimised for 200 steps using conjugate gradient and MMFF94 force field [21] and prepared for docking by adding Gasteiger charges, merging non-polar hydrogen atoms, assigning AutoDock4 atom types, and adding hydrogen atoms. The root torsion, degree of freedom, and the number of rotatable bonds were also defined, and the structures were saved in pdbqt format. Autodock4 default parameters were used during docking and both protein and ligand were considered rigid. The docked compounds were ranked according to docking scores and binding site interactions to select compounds for preliminary bioassay evaluations.

#### 4.2.7 Diversity selection

Compounds that were structurally distinct in the set identified from structure-based virtual screening were chosen for experimental validation using the Datawarrior FragFp descriptor. The FragFp is a binary fingerprint based on a substructure fragment dictionary that selects fragments with little or no overlap, implying diversity.

### 4.3 Results and discussion

#### 4.3.1 Physicochemical property space

To reduce attrition and improve market potential, drug candidates should have the appropriate physicochemical properties. The physicochemical properties of the five compound datasets described above, FDA approved drugs (FDA), FDA approved breast cancer drugs (FDA\_BCD), Maybridge HitCreator (Maybridge), and nicastrin inhibitors presented as active and inactive were analysed and compared. These include the molecular weight ( $MW \leq 500$ ); the calculated logarithm of the octanol/water partition coefficient ( $cLogP \leq 5$ ); the number of hydrogen bond donors ( $HBD \leq 5$ ) and acceptors ( $HBA \leq 10$ ); the topological polar surface area ( $TPSA \leq 140 \text{ \AA}$ ); and the number of rotatable bonds ( $RotB \leq 10$ ) [3, 4]. **Figure 4.1** depicts the range and distribution of physicochemical properties for each of the five datasets.

A multivariate approach based on principal component analysis (PCA) was used to analyse the six physicochemical properties and draw the physicochemical space by reducing the dimensions in the datasets. Eigenvectors that are descriptive of the contribution of each property to the amount of variance in the PCA were loaded and the scores of the first three PCs were plotted in a 3D model as shown in **Figure 4.2**.

According to the plot analysis, the first three principal components (PCs) account for more than 85% of the variance, as shown in **Table 4.2**, indicating that these PC dimensions can provide meaningful visualisations of the physicochemical property space defined by the compounds in the datasets.

The analysis shows that most compounds in the nicastrin datasets have properties that fall within the Lipinski rule of 5 of orally available drugs; however, when compared to the FDA BCDs, which served as a reference, some notable deviations are observed within the sets. The PCA analysis revealed that there is a physicochemical space overlap between the datasets. As the FDA approved drugs extend along the PC1 and PC2 axis, the number of rotatable bonds increased, with high hydrogen bond counts and TPSA values greater than  $200\text{\AA}^2$ . This was captured in the box, and whisker plots, that Maybridge screening set and FDA approved drugs, FDA BCDs and nicastrin inhibitors have a higher molecular weight. The average molecular weight of the Maybridge screening set, and FDA approved drugs is between 342 and 345 Da, which is significantly lower than the 496, 477 and 429 Da of FDA breast cancer drugs and nicastrin active and inactive inhibitors respectively ( $p > 0.05$ ).

The FDA approved, FDA breast cancer drugs, and nicastrin inhibitors all have between 5 and 7 rotatable bonds, whereas the Maybridge screening set has compounds with an average of 4 rotatable bonds. While nicastrin inhibitors have significantly higher cLogP values than FDA approved breast cancer drugs ( $p < 0.01$ ), the cLogP of the Maybridge dataset and FDA approved drugs is not statistically different from that of breast cancer drugs ( $p > 0.05$ ). For all of the data sets, there are significantly fewer hydrogen bond acceptors, and donors, than for FDA breast cancer drugs ( $p < 0.05$ ).

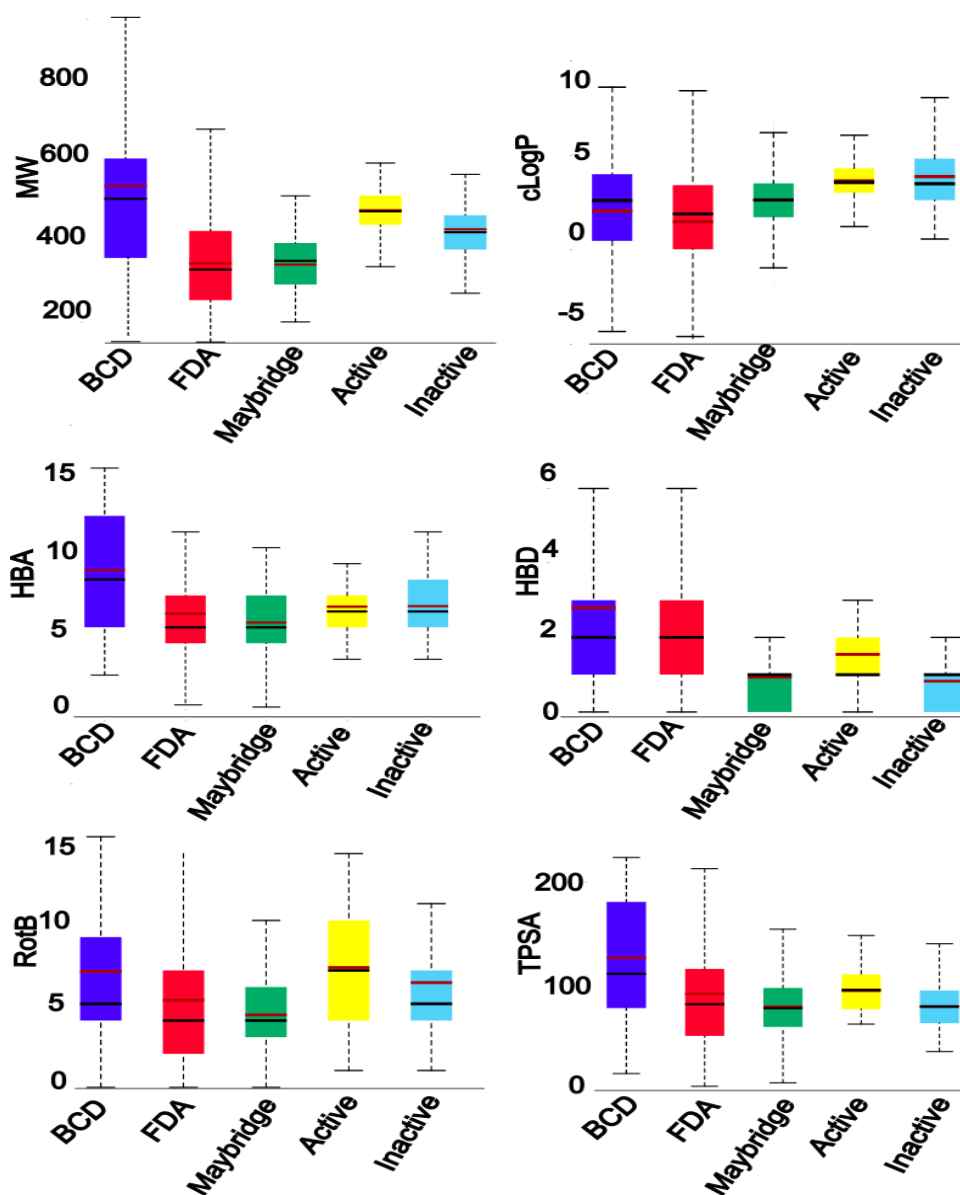
This is also explained by the highest loadings of PC1 and PC2 (**Table 4.2**): HBA (0.45), HBD (0.42), TPSA (0.55) and RotB (0.41).

When all these observations are taken into account together with the fact that FDA breast cancer drugs' TPSA averages  $126 \text{ \AA}^2$  ( $p < 0.05$ ), which is significantly higher than the average for all datasets with averages between 70 and  $94 \text{ \AA}^2$ , it explains why breast cancer drugs are mostly administered intravenously. This space can be explored for gamma-secretase inhibitors that can be administered intravenously.

However, the space occupied by the majority of the Maybridge, and FDA approved drugs overlap, and extend towards PC2 and PC3 axis providing compounds with physicochemical properties that allow for the design of orally available drugs. This property space is characterised by compounds with reduced lipophilicity indicated by cLogP values, low polar surface area, low molecular weight compounds with low hydrogen bond counts that can help to expand the property space of the GSIs to occupy a broader property space of approved drugs. PC2 and PC3 have the highest loadings from a reduction (indicated by the negative values) in cLogP (-0.72), HBD (-0.70) and RotB (-0.45) and which might imply that these properties might influence the property space (**Table 4.2**).

The Maybridge screening dataset occupies the space with lower lipophilicity, less flexibility as indicated by the number of rotatable bonds, and an acceptable total hydrogen bond count, all of which contribute to better oral bioavailability, <sup>[4]</sup> and can be used to identify orally available GSIs. This demonstrates that potential hits with different physicochemical properties can be identified from screening of the Maybridge

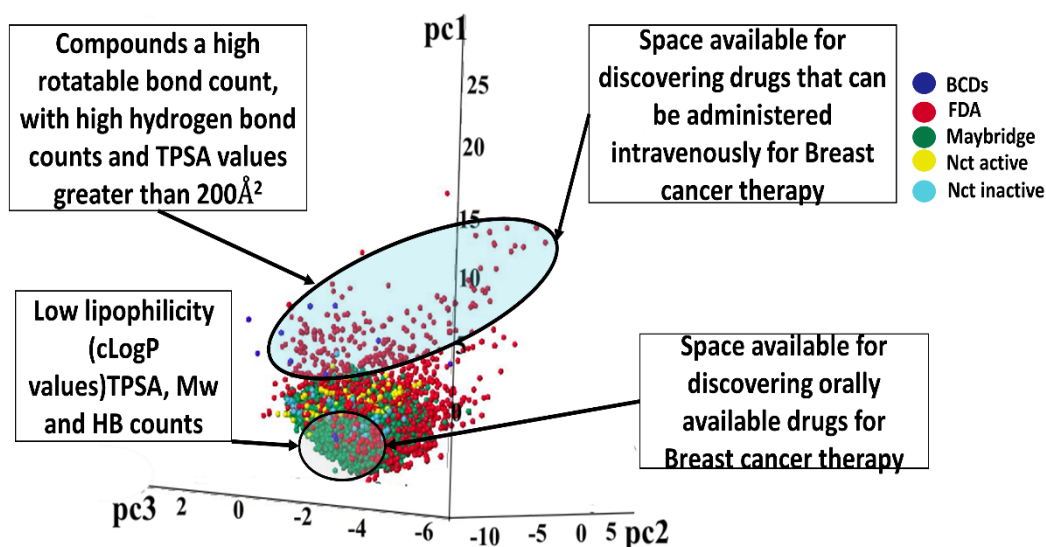
datasets for breast cancer therapy. **Figure 4.2** illustrates a physicochemical property space that is devoid of GSIs but can be used to discover novel GSIs.



**Figure 4.1** Physicochemical property distribution of FDA approved breast cancer drugs (BCD), FDA approved drugs (FDA), Maybridge HitCreator dataset (Maybridge), nicastrin actives (Active) and nicastrin inactive inhibitors (Inactive). The physicochemical properties include molecular weight (MW), octanol water partition coefficient (cLogP), hydrogen bond acceptors (HBA) and donors (HBD), rotatable bonds (rotB) and topological polar surface area (TPSA) are presented as box plots. Each box represents values between the first and third quartiles; the red line shows the mean value of the data; the black line is the median; and the whiskers indicate the top and bottom quarters of the data.

**Table 4.2** Eigen values and explained variance of Principal components

PC and Explained variance (%)	PC1	PC2	PC3
Molecular weight	0.3996	-0.2905	0.5574
cLogP	0.021	-0.7166	0.0943
HBA	0.4485	0.3761	0.3575
HBD	0.4184	0.026	-0.7017
TPSA	0.5467	0.2439	0.0146
Rotatable bonds	0.4059	-0.4478	-0.2451



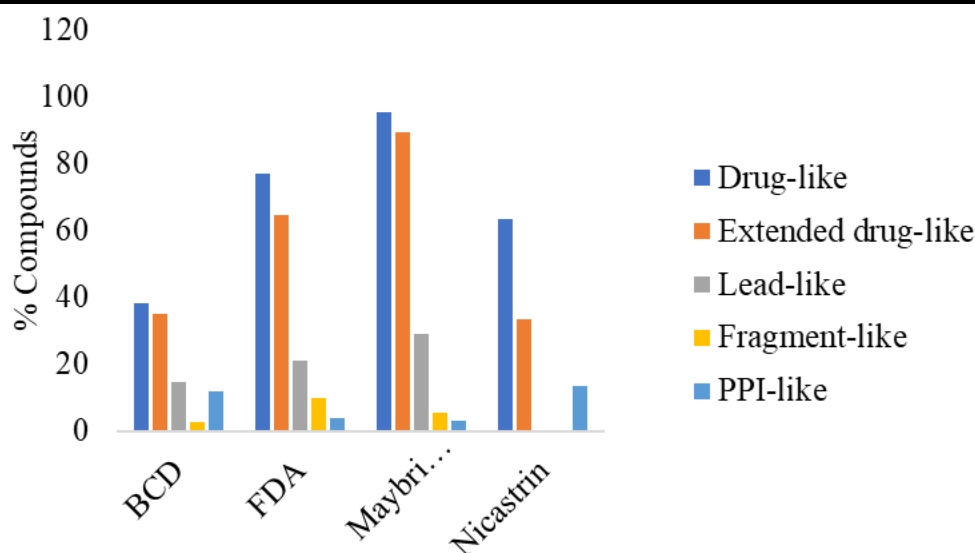
**Figure 4.2** Property-based chemical space of FDA approved breast cancer drugs (BCD), FDA approved drugs (FDA), Maybridge HitCreator dataset (Maybridge), nicastrin actives (Active) and nicastrin inactive inhibitors (Inactive) as principal components (PCs) of physicochemical properties of pharmaceutical relevance: molecular weight, hydrogen bond acceptors, hydrogen bond donors, topological polar surface area, rotatable bonds, and the octanol-water partition coefficient.

#### 4.4 Profiling of drug-likeness

Drug-likeness, extended drug-likeness, lead-likeness, fragment-likeness, and protein-protein interaction (PPI) likeness of nicastrin inhibitors and compounds in the FDA approved drugs, FDA breast cancer drugs and Maybridge datasets were also analysed. The Maybridge dataset had the highest percentage of compounds that met the drug like, extended drug like, and lead like criteria, with 95.5%, 89.3%, and 29%, respectively (**Figure 4.3**). With respect to drug-likeness and extended drug-likeness, the FDA approved drugs dataset had more than 50% of compounds adhering to both rules. Therefore, these findings suggest that compounds in the Maybridge dataset can be used in drug discovery when screening for oral drugs.

Since most breast cancer compounds are administered intravenously, as evidenced by the low percentage of drug-like breast cancer compounds, nicastrin inhibitors can be used to design orally available drugs, since 64% of compounds are drug like. All datasets had very small fractions (less than 30%) of lead-like, fragment-like and PPI like compounds. Since molecular fragments and lead like molecules are required during lead-discovery and optimisation stages of drug discovery, larger fractions of fragment-like, and lead-like, were not expected in the FDA approved drugs, and breast cancer drugs sets, as these compounds had already passed the discovery and optimisation stages.





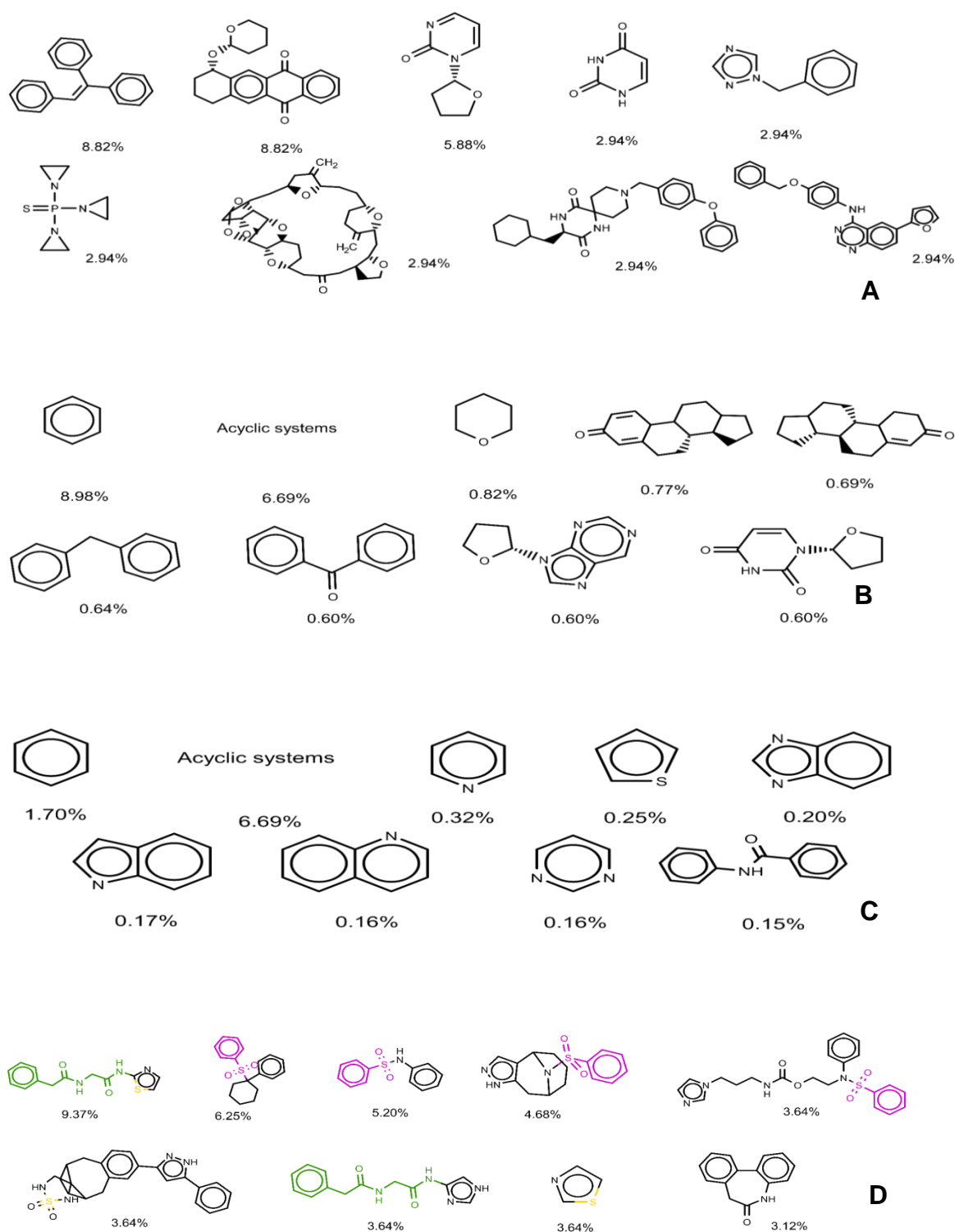
**Figure 4.3** Drug-like, extended-drug like, lead-like, fragment-like, PPI like profiles of FDA approved breast cancer drugs (BCD), FDA approved drugs (FDA), Maybridge HitCreator dataset (Maybridge), and nicastrin inhibitors (Nicastrin).

#### 4.4.1 Scaffold chemical analysis

Scaffold analysis was done to characterise nicastrin inhibitors to inform the identification of compounds in the Maybridge screening dataset, while also monitoring uniqueness of scaffolds by characterising the scaffold diversity of FDA approved drugs. To profile the scaffold diversity of nicastrin inhibitors against the Maybridge dataset, compounds were decomposed into Murcko frameworks and side chains. [22] The scaffold diversity of each dataset was analysed by computing the fraction of scaffolds relative to the total number of compounds in the dataset ( $N_s/M$ ), proportion of singleton scaffolds to the total number of compounds in the dataset ( $N_{ss}/M$ ) and singleton fractions to the total scaffolds in the datasets ( $N_{ss}/N_s$ ).

Based on Ns/M, Nss/M and Nss/Ns the scaffold diversity order decreased in the following order; Maybridge > FDA BCD > FDA approved > nicastrin inhibitors. While Ns/M generally describes the scaffold diversity distribution with a particular dataset, Nss/M and Nss/Ns reveals the proportion of unique scaffolds with respect to all compounds and scaffolds in a dataset. The proportion of singleton fractions to the total scaffolds in the datasets was high in all datasets (>0.70). This shows that a greater fraction of compounds in Maybridge dataset has unique scaffolds that can be used in the exploration of novel drug scaffolds for breast cancer. **Figure 4.4** shows the nine most common scaffolds in the six datasets used in this study, including acyclic systems.

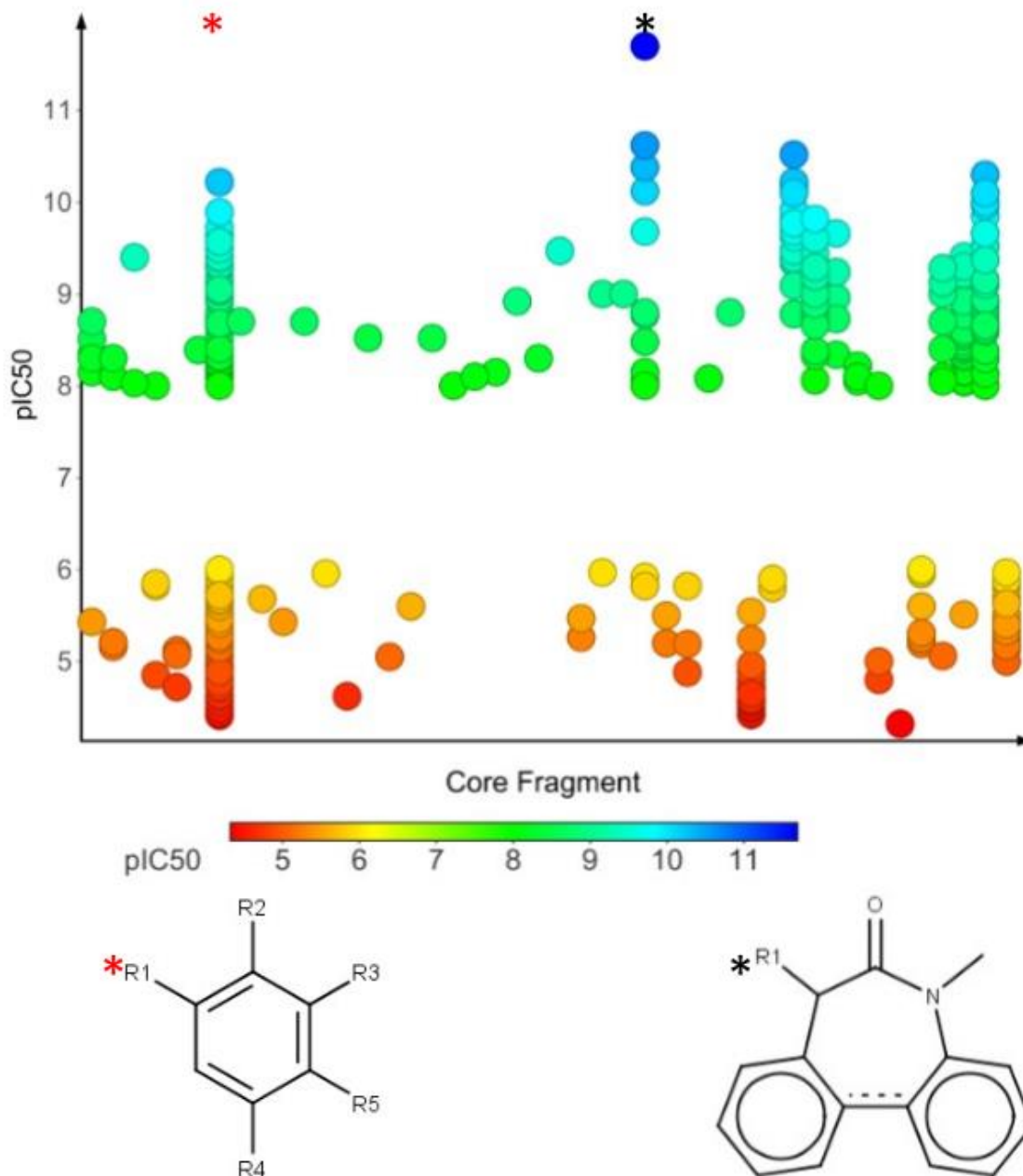
The N-(2-phenylacetyl)-N-1,3-thiazol-2-ylglycinamide scaffold (9.37%) has the highest occurrence in the nicastrin inhibitors. A sulfur bond or a sulfonyl benzene is also present in four of the most common scaffolds in nicastrin inhibitors. None of the most frequent scaffolds in nicastrin inhibitors were found in the Maybridge screening dataset, and other datasets, implying that the Maybridge screening dataset can be proposed as a novel source of chemical scaffolds for nicastrin inhibition and breast cancer therapy.



**Figure 4.4** The most common scaffolds including acyclic systems found in: A. FDA breast cancer drugs, B. FDA approved drugs, C. Maybridge HitCreator set, and D. nicastrin inhibitors.

#### 4.4.2 Structure activity relationships and Activity cliff analysis of gamma-secretase inhibitors

A structure activity relationship was used in which compounds were decomposed into their core fragments and R-groups. In this investigation the most central ring system scaffolds, which are topologically closest to the centre of the molecule were used, where all atoms and bonds from a given compound that were not a part of any rings were removed and the unsubstituted ring systems were retained. <sup>[1]</sup> For the 353 nicastrin inhibitors, 44 core fragments were produced. The structure activity relationships highlighted R-groups on the core fragments that significantly altered the activity. For instance, in nicastrin inhibitors, compounds with the R3 group on the benzene core fragment are highly active, whereas R1, R4, and R5 groups on the benzene core fragments reduced the activity. This effect of R-groups on the core fragments is shown in **Figure 4.5**.



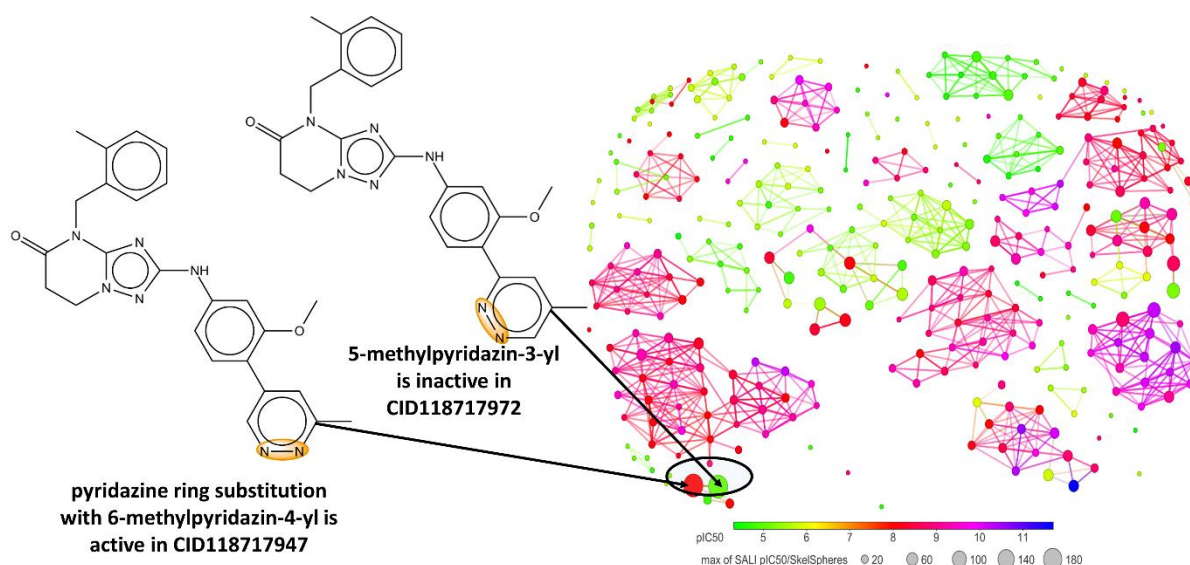
**Figure 4.5** A scatter plot of activity as pIC50 against core fragment.

The points are coloured according to their pIC50 values with red dots and blue dots corresponding to low and high values respectively.

In addition to the R-groups on core fragments, a similarity check of nicastrin inhibitors, and effects of minor structural changes in compounds that affect their bioactivity, was conducted (**Figure 4.6**). This was accomplished by analysing the bioactivities using

the structure-activity landscape index (SALI) [23]. In Datawarrior, a SALI index for a pair of compounds measures how bioactivity changes as compound structure changes. Each point in the displayed SALI plot represents a different compound, and the colour and size of each point are scaled based on the pIC<sub>50</sub> and SALI value. The distance between compounds is proportional to the compound similarity.

The SALI plot (**Figure 4.6**) shows that nicastrin inhibitors are chemically diverse since they are scattered throughout the chemical space. Pairs of compounds with similar structural characteristics but different biological activities were identified by activity cliff analysis. For example, a pair of nicastrin compounds with CIDs 118717947 and 118717972 (**Figure 4.6**) had a very high similarity value of 0.98 and bioactivity pIC<sub>50</sub> values of 8 and 5. Given that the small difference on the substitution position of nitrogen atoms on the pyridazine ring, this pair of compounds had a SALI value of 196.95, the highest amongst all pairs (**Figure 4.6**). Compound CID 118717947 with 6-methylpyridazin-4-yl is active, while compound CID 118717972 with 5-methylpyridazin-3-yl is inactive. This demonstrates how activity cliffs can be used to improve activity in nicastrin inhibitors.



16

**Figure 4.6** A structure-activity landscape index (SALI) plot of nicastrin inhibitors. Each point in the displayed SALI plot represents a different compound, and the colour and size of each point are scaled based on the pIC50 and SALI value. A comparison of structural similarity between a pair of nicastrin inhibitors CID 118717947 and 118717972 with a similarity of 0.98 is highlighted with the highest SALI value of 196.95.

#### 4.4.3 Quantitative Structural Activity Relationship (QSAR) prediction using Machine Learning

Quantitative structure-activity relationships using machine learning was done to identify features important for nicastrin inhibition. In the OCHEM web server, 306 descriptors were calculated, and after removing redundant descriptors, only 187 were subjected to descriptor reduction methods. The Best-first search method yielded descriptor subsets that could be used to build models. To assess the predictive capabilities of the subsets produced, the correlation feature-based selection method

(CFS), as well as J48 trees, Naïve Bayes, IB1 and SMO algorithms within a cross-validation loop (Wrapper), were used.

**Table 4.3** shows the descriptor sets that were evaluated by the CFS and Wrapper methods. To produce robust models of nicastrin activity, the CFS and SMO methods selected subsets with 13 descriptors each, while the IB1, J48 trees, and NB selected subsets with 10, 7, and 4 descriptors, respectively. The evaluated subsets included topological, electro topological, constitutional, electro geometric, and hybrid descriptors. Most reduction methods shared electrotopological and constitutional descriptors such as SsH and naAromAtom.

Electro topological descriptors such as SsH, SsssCH, SdssC, SsO, and SsssN consider an atom's electronegativity, and buriedness, and provide a value that depicts the atom's accessibility for interaction with other atoms in another molecule. Topological descriptors such as MDEO-11 provide distance and adjacency information that characterise oxygen connectivity, whereas MDEC-14, C3SP2, and C1SP2 descriptors detail carbon connectivity and hybridisation. To capture shape, size, symmetry, and atom distribution, hybrid descriptors are calculated from the x, y, z coordinates of a molecular structure. BCUT and WHIM descriptors are two examples of hybrid descriptors. Constitutional descriptors reflect the chemical composition of the compound, such as the number of acid (nAcid) or aromatic groups (naAromAtom), but do not provide atom connectivity.



**Table 4.3.** Subsets of descriptors selected for the development of the different Machine Learning models

Descriptor	CFS	J48	IBK	NB	SMO
Topological	SsssCH SsH SdssC SssO SsCl C1SP2 C3SP2 MDEO-11 khs.ssNH	SsH SsssN C3SP2 MDEC-14	SaasN SdssC SaasC SssssC SsCl MDEO-11 VP-0 Khs.dssC khs.ssNH	SsssCH SsH SssO	SsssCH SsH SaasN SssO SaasC SsssN SssssC MDEC-22 khs.ssNH
Constitutional	naAromAtom	nAcid naAromAtom		naAromAtom	naAromAtom
Electro-geometric	tpsaEfficiency				
Hybrid	tpsaEfficiency BCUTw-1h Wlambda2.unity	BCUTp-1h	WV.unity		BCUTp-1h Wlambda2.ur Weta3.unity

#### 4.4.3.1 Validation

This section displays the results of external validation for each machine learning method. **Table 4.4** compares the performance of the methods as weighted averages. A paired T-test revealed no significant difference in the results of the models built using the various classification methods at the 5% significance level, implying that all these methods can be used interchangeably based on the training results.

#### 4.4.3.2 J48 trees

The C4.5 algorithm was used to generate J48 decision trees for the classification of nicastrin inhibitors. The tree constructed from the CFS subset of descriptors had a

ROC area of 0.80, contained 15 internal nodes and 16 leaves from 13 descriptors and correctly classified 84.5% of the compounds in the test set. The Wrapper J48 decision tree was also evaluated, and found to outperform the CFS-based method, producing a model that correctly classified 90.1% of the test set compounds while using half the number of descriptors. The wrapper J48 method reduced the false positive rate in the CFS-based method from 15.8% to 11%, resulting in a ROC area of 0.91. Sensitivity and specificity are good indicators of a model's ability to correctly classify active compounds as active and misclassify inactive compounds as active with a low error rate (false positives). Interestingly, as shown in **Table 4.4**, the descriptor subsets chosen in both models correctly predicted 81.8% of the inactives.

#### **4.4.3.3 IB1**

To distinguish actives from inactives, an instance-based learning algorithm, based on the nearest neighbour pattern classifier, was developed. The accuracy measures were the same for the CFS-based subset as well as the subset evaluated using the Euclidean distance to find the nearest neighbour. The wrapper method used ten descriptors to build the model, whereas the CFS method used thirteen descriptors to produce models with a 90.1% accuracy and a ROC area of 0.91. With a false positive rate of 10.6% compared to 11% for the wrapper model, the CFS-based model was more selective to inactive compounds.

#### **4.4.3.4 Naïve Bayes**

The Naïve Bayes (NB) model was built using the 13 descriptors evaluated via CFS had an accuracy of 87.3% and a ROC area of 0.90. The use of the NB as an attribute selector improved the model's accuracy to 91.6% while lowering the false positive rate

to 0.09, the lowest of all machine learning models. This demonstrates that the model was most selective for inactive compounds.

#### **4.4.3.5 SMO**

The radial basis function kernel was used to convert the non-linear data to linear form and to distinguish between active and inactive compounds. The hyper-plane distinguished between active and inactive compounds with 85.9% accuracy and a ROC area of 0.85. Although using the SMO rather than the CFS to evaluate the descriptor subset increased the model's sensitivity the ROC area by 1.4% and 2%, respectively, the model's specificity decreased to 75.8%, reducing the model's ability to segregate inactive compounds.

#### **4.4.3.6 Model performance**

Cohen's Kappa is a measure of agreement between a model's ability to classify a compound and its correct class. A model's random predictive ability is 0.5, so a perfect model has a value of 1. <sup>[24]</sup> In general, the models (**Table 4.4**) had good predictive ability, with the wrapper NB having exceptional predictive power with a kappa value of 0.83.

Sensitivity and specificity are critical metrics for selecting a model that will correctly classify compounds in a database, thereby maximising resources during drug development. The wrapper J48 model was the most sensitive to actives, whilst the wrapper NB was the most specific to the classification of inactive compounds and hence used in the screening for nicastrin actives.

**Table 4.4.** Summary of Model performance

<b>Classifier</b>	<b>Sensitivity</b>	<b>Specificity</b>	<b>Accuracy</b>	<b>ROC</b>	<b>MCC</b>	<b>BCR</b>	<b>Cohen's Kappa</b>
J48+CFS	86.8	81.8	84.5	0.80	0.69	84.3	0.69
IBK+CFS	94.7	84.8	90.1	0.91	0.80	89.8	0.80
NB+CFS	92.1	81.8	87.3	0.90	0.75	87.0	0.74
SMO+CFS	92.1	78.8	85.9	0.85	0.72	85.5	0.71
Wrapper J48	97.4	81.8	90.1	0.91	0.81	89.6	0.80
Wrapper IBK	97.4	81.8	90.1	0.91	0.81	89.6	0.80
Wrapper NB	94.7	87.9	91.6	0.87	0.83	91.3	0.83
Wrapper SMO	97.4	75.8	87.3	0.87	0.76	86.6	0.74

#### 4.4.3.7 Applicability domain

According to the applicability domain calculation, the test set's percentage of reliable predictions was 100%. Based on training set data, all the 71 test set compounds were within the APD threshold of 1.954 and were reliably predicted; thus, the validation performed on the test set was accepted as reliable. [25]

#### 4.4.4 Interpretation of J48 and NB models

The J48 trees and NB algorithms were used to assess the predictive capabilities of the subsets produced within a cross-validation loop that described nicastrin activity using topological, electro topological, constitutional, and hybrid descriptors. Because the J48 trees were sensitive to actives and the NB models were specific to inactive compounds, they were used consecutively.

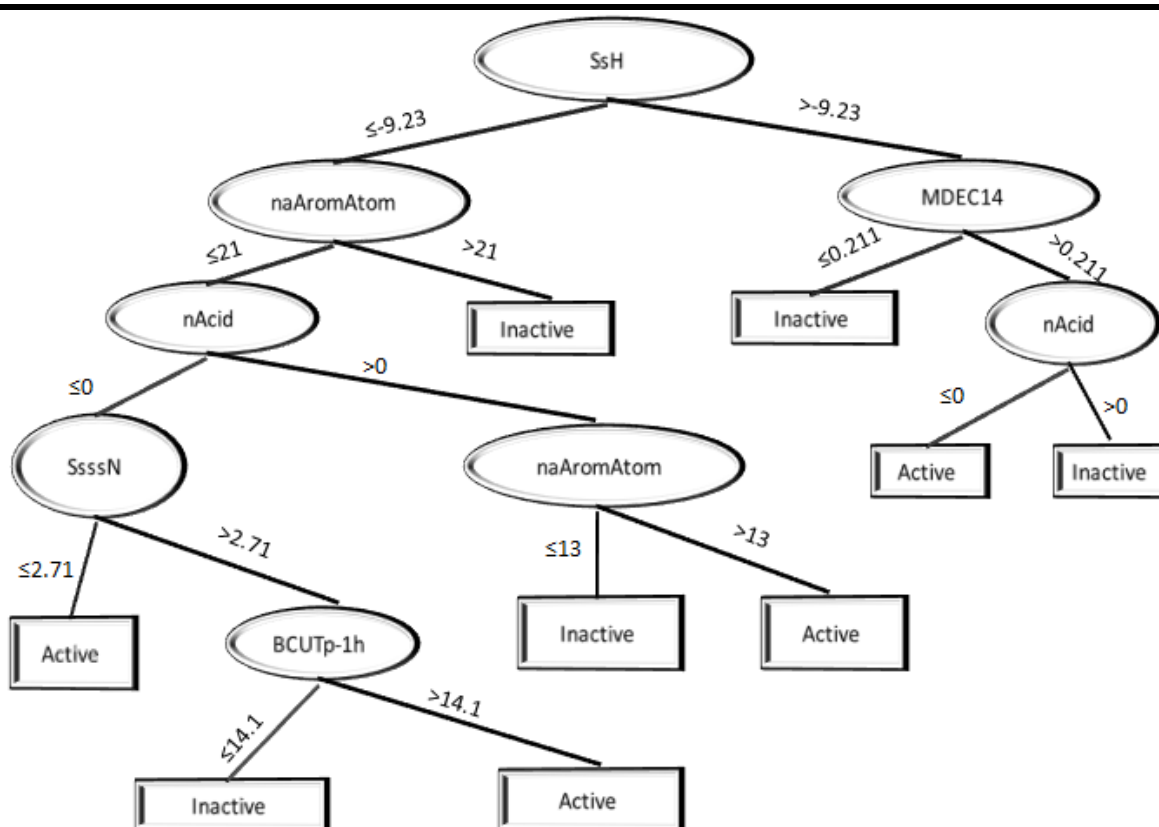
The J48 trees model (**Figure 4.7**) demonstrates that the presence of heteroatoms such as halogen, nitrogen, and sulfur atoms, the degree of branching, chain length, and the presence of cyclic structures all contribute to the activity of nicastrin inhibitors. The SsH feature, referring to the sum of polar hydrogens, and ultimately, the polarity of the molecule, is the most important descriptor identified by the J48 trees model. Nicastrin compounds with halogen, and sulfur, atoms increase the valence state electronegativity of hydrogen atoms. Furthermore, the SsssN feature, which is the sum of singly-bound amines, was used to describe the potency of nicastrin inhibitors. When amines are bound to phenylsulfonyl or morpholine groups, the compounds become inactive.

The MDEC-14 descriptor, which accounts for the geometric mean of the topological path lengths between all primary and quaternary carbons describes the separation of a branched molecule's side chains from its main body. From the J48 trees model, it was observed that compounds can only be active if their BCUTp-1h eigenvalues are high. The BCUTp-1h descriptor combines connectivity and atomic polarisability information. The phenylsulfonamide group is an example; when bound to an alkyl chain, the compound is active. However, if it is bound to an aromatic group, it is inactive due to electron delocalisation, resulting in high polarisability and low eigenvalues. Furthermore, compounds with no more than three 5-6 membered aromatic rings, as defined by the descriptor naAromAtom, are active. Except for compounds containing both the phenyl sulfonamide and the 4-chlorobenzyl alcohol groups, the absence of acidic groups in the inhibitors generally favored activity. **Figure 4.7** depicts the J48 wrapper model.

The NB model showed high sensitivity to inactive compounds, and to separate inactives from the dataset, the NB wrapper model used a subset with three topological

descriptors, SsssCH, SsH, SssO, and one constitutional descriptor (naAromAtom). Both the J48 wrapper and the NB wrapper models share the SsH, and naAromAtom descriptors. The SsssCH descriptor for saturated carbon hydride is an electrotopological descriptor that provides a count of aromatic carbon groups. Its application during screening allows for the consideration of electronegativity that is buried topologically within the structure and influences activity. <sup>[26]</sup> In the case of nicastrin inhibitors, the presence of heteroatoms such as oxygen, as described by the SssO descriptor, improves the segregation of actives and inactives.

The Maybridge dataset with 14000 compounds was initially screened using the applicability domain (APD) filter to identify compounds that could be reliably predicted by the models. The APD filter identified 10012 compounds from the dataset. The 10012 compounds were then screen using the J48 and NB wrapper models and 1315 compounds were identified as active from the Maybridge dataset.



**Figure 4.7** Representation of the J48 Wrapper model.

#### 4.4.5 Structure based virtual screening of the Maybridge dataset

To identify nicastrin inhibitors, the 1315 compounds identified through QSAR screening using the J48, and NB wrapper models, were docked against the DYIGS site of nicastrin (PDB ID: 6IDF). The criteria for selection of hits were that the docked compounds interact with DYIGS residues Val138, Asp143, Arg105 and Glu174 previously identified as important for good binding affinity as shown in **Table 4.5**.

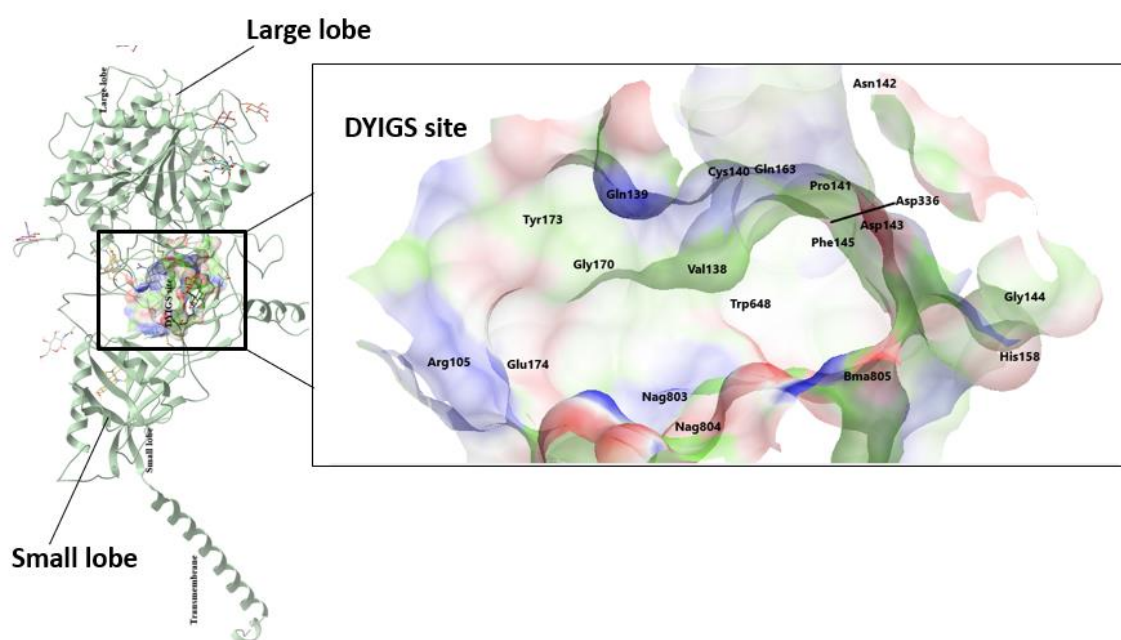
**Table 4.5.** Important binding site residues in nicastrin DYIGS binding site

Residue	Interaction
Val138	Hydrophobic alkyl interactions
Gln139	Van der Waals interactions

Asp143	Hydrogen bond
Arg105	pi-cation contacts
Glu174	pi-anion contacts

#### 4.4.6 Interactions of identified hits in the binding site

As shown in **Figure 4.8**, the DYIGS site is located between the large and small lobes. It has a large volume of 738 Å<sup>3</sup> with residues Val138, Gln139, Cys140, Asn142, Asp143, Cys159, Gln163, Tyr173, and Trp648 surrounding the conserved Asp336 of the DYIGS motif.



**Figure 4.8** The DYIGS site in nicastrin.

The binding site residues are coloured by their characteristics; lipophilic aromatic (white), non-aromatic lipophilic (green), hydrogen bond donor (blue) and hydrogen bond acceptor potential (red).

The four gamma-secretase inhibitors with nicastrin activity from PubChem dataset CIDs 44433923, 11305056, 23571070, and 11396006 were also included in the

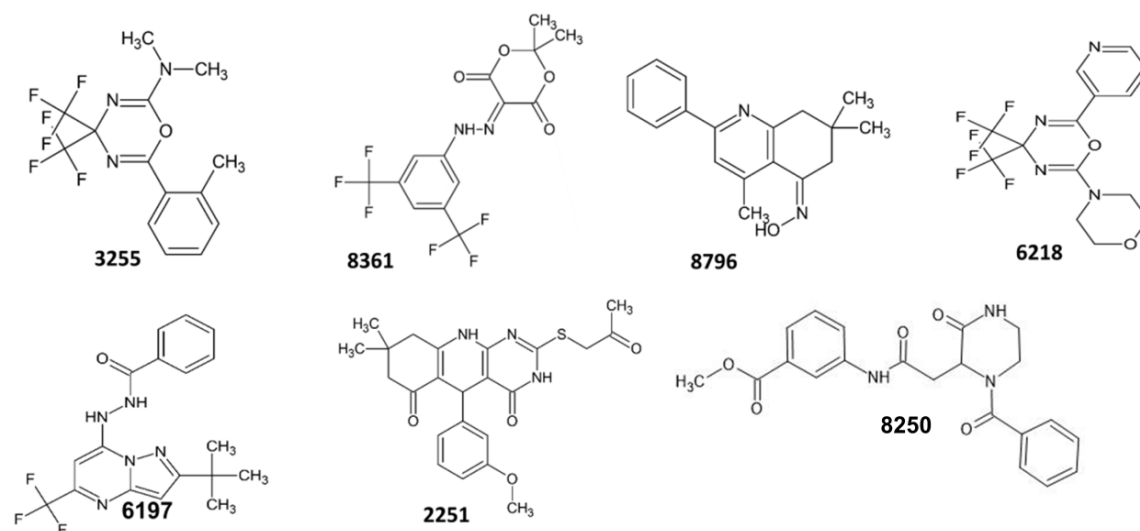


structure-based virtual screening to establish a binding affinity cut off for further selection of potential hits. From the docking results, the cut-off value was determined by considering the binding affinity CID 11396006. CID 11396006 had a binding affinity of  $-9.0 \text{ kcal mol}^{-1}$ .

In general, the polar charged aspartates Asp143 and Asp336 form salt bridges with the hit compounds' tertiary amine groups which include pyrazole, nitropiridine, indole, tetrazole, morpholine, and piperidine. These tertiary amine groups were identified as important for nicastrin inhibition by the J48 trees model. The aromatic nature of the compounds allowed them to be accommodated in the binding site via pi-sigma and pi-alkyl interactions with residues such as Val138, Cys159, and Cys140, as well as pi-anion interactions with Asp336 and pi-pi stacking with Tyr173. Complementary van der Waals interactions with hydrophobic residues such as Try173, Asn142, Pro141, Asn55, His158, Asn169, Gly170, Gln139, Phe145, Phe335, Trp648, Gly144, and glycan residues are observed around these aromatic groups as well as other alkyl groups that are mostly methyl.

Fluorine atoms in the hit compounds formed halogen bonds with the binding site residues Gln139, Gln339, Asp336, Tyr173, Glu174, Thr334, Cys140, Asp143, and Asn169. The presence of oxygen in the binding site allowed hydrogen bonds to form with polar binding site residues like Gly144, Asp143, Asp336, Cys140, Gln163, His158, and glycans. In the case of carboxamides and oxazoles, hydrogen atoms attached to the nitrogen atom allowed for hydrogen bond formation.

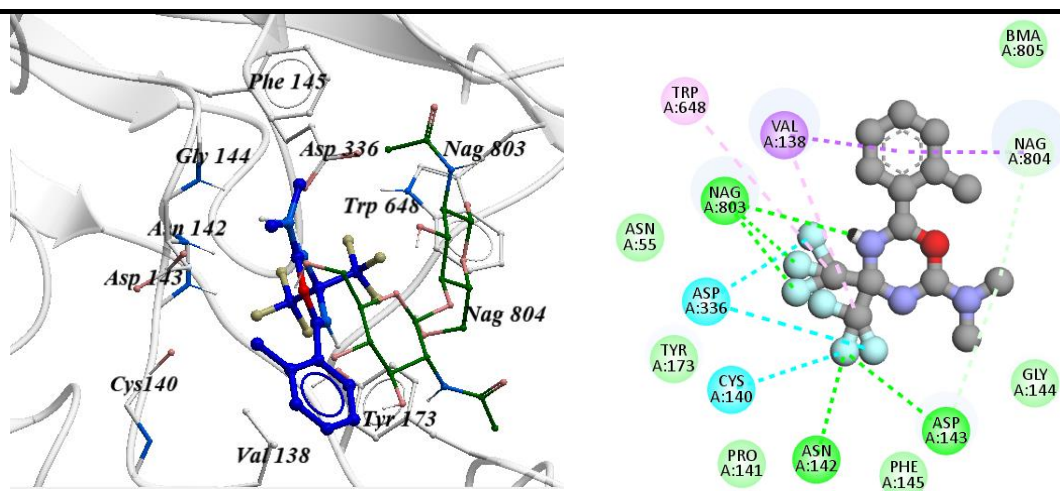
Seven compounds (**Figure 4.9**) were chosen for experimental validation after considering the structural diversity of the docked compounds from known nicastrin inhibitors and binding affinity of  $-9.0 \text{ kcal/mol}$  or less.



**Figure 4.9** Nicastrin inhibitors selected for experimental validation.

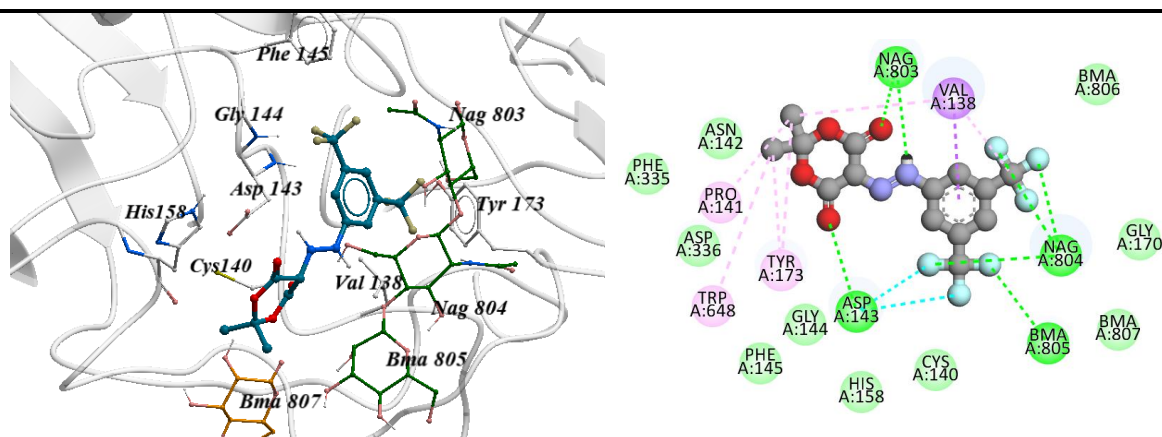
The compounds were docked and analysed, and significant interactions were observed between the ligands and the binding site residues. Compound 3255 (N, N-dimethyl-6-(2-methylphenyl)-4,4-bis(trifluoromethyl)-1,3,5-oxadiazin-2-amine) binds to the target site (**Figure 4.10**) via three hydrogen bonds with Asn142, Asp143, and Nag803 as well as halogen bonds with Asp336 and Cys140 via the bis(trifluoromethyl) group (trifluoromethyl).

Carbon hydrogen bonds are also formed between the dimethyl group attached to the amine and Asp143 and Nag804. Pi-interactions, which are important for drug stabilisation in the binding site via charge transfer, were also realised. The methylphenyl group interacts with Val138 and Nag804 via pi-sigma interactions, whereas the methyl groups interact via pi-alkali with Trp648 and alkyl interactions with Val138.



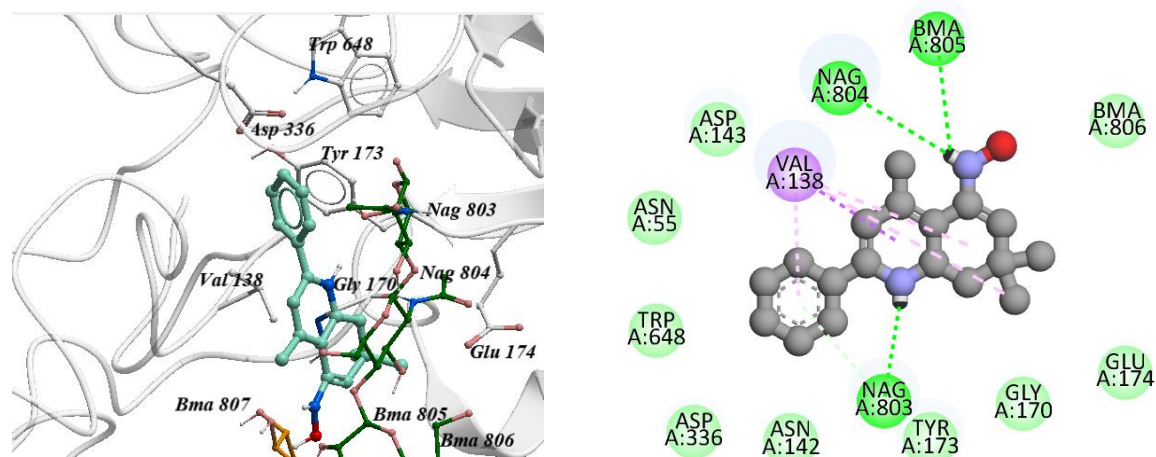
**Figure 4.10** Molecular docking 2D mode of interaction of compound 3255 (N, N-dimethyl-6-(2-methylphenyl)-4,4-bis(trifluoromethyl)-1,3,5-oxadiazin-2-amine) with DYIGS binding site in nicastrin analysed by Discovery Studio. Conventional hydrogen bond, van der Waals interactions, Carbon hydrogen bond, halogen, pi-sigma, alkyl, and pi-alkyl are shown in green, light green, and pink respectively.

The methylphenyl group of compound 8361 (5-[[3,5-bis(trifluoromethyl)phenyl]hydrazinylidene]-2,2-dimethyl-1,3-dioxane-4,6-dione) interacts with Val138 and Nag804. The bis(trifluoromethyl) group interacts with fluorine atoms in a variety of ways, including hydrogen bonds with Asn142, Asp143, and Nag803, as well as halogen bonds with Asp336 and Cys140. With Trp648 alkyl interactions, the methyl groups form pi-alkali. A carbon-hydrogen bond is also formed between the amine's dimethyl group and Asp143 and Nag804 (**Figure 4.11**).



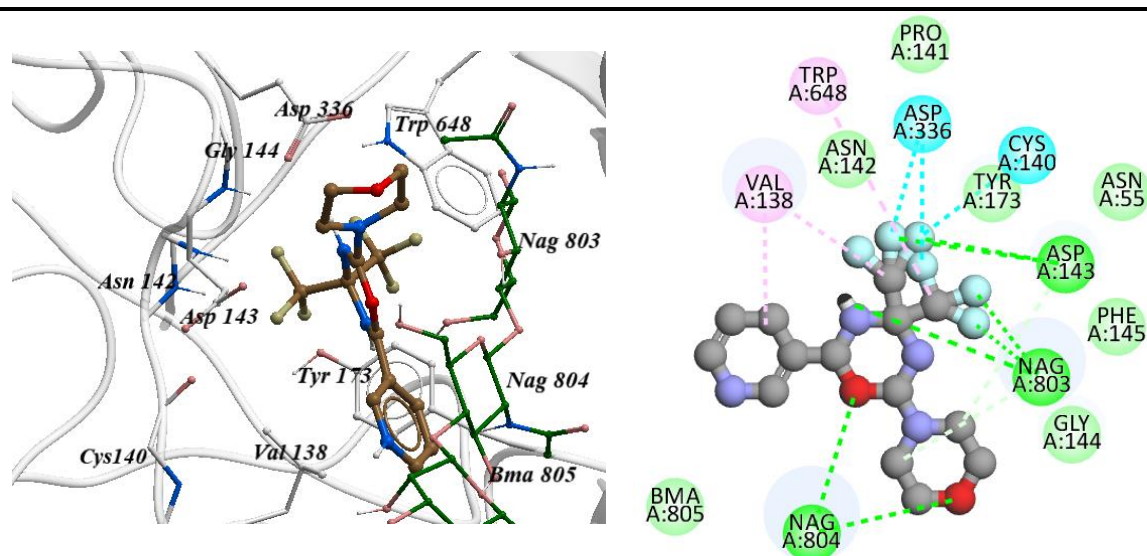
**Figure 4.11** Molecular docking 2D mode of interaction of compound 8361 (5-[[3,5-bis(trifluoromethyl)phenyl] hydrazinylidene]-2,2-dimethyl-1,3-dioxane-4,6-dione) with DYIGS binding site in nicastrin analysed by Discovery Studio. Conventional hydrogen bond, van der Waals interactions, Carbon hydrogen bond, halogen, pi-sigma, alkyl, and pi-alkyl are shown in green, light green, and pink respectively.

The interactions with compound 8796 (N-(4,7,7-trimethyl-2-phenyl-6,8-dihydroquinolin-5-ylidene) hydroxylamine) are dominated by van der Waals interactions. Van der Waals interactions with the compound are induced by the residues Asp143, Asn55, Trp648, Asp336, Asn142, Tyr173, Gly170, Glu174, and Bma806. Through pi-interactions, Val138 stabilises the compound. Pi-sigma, pi-alkyl, and alkyl interactions were observed with the 6,8-dihydroquinolin-5-ylidene group and the phenyl group, whereas conventional hydrogen bonds were observed primarily with glycan residues, Nag803, Nag804, and Bma805 with hydroxylamine hydrogens (Figure 4.12).



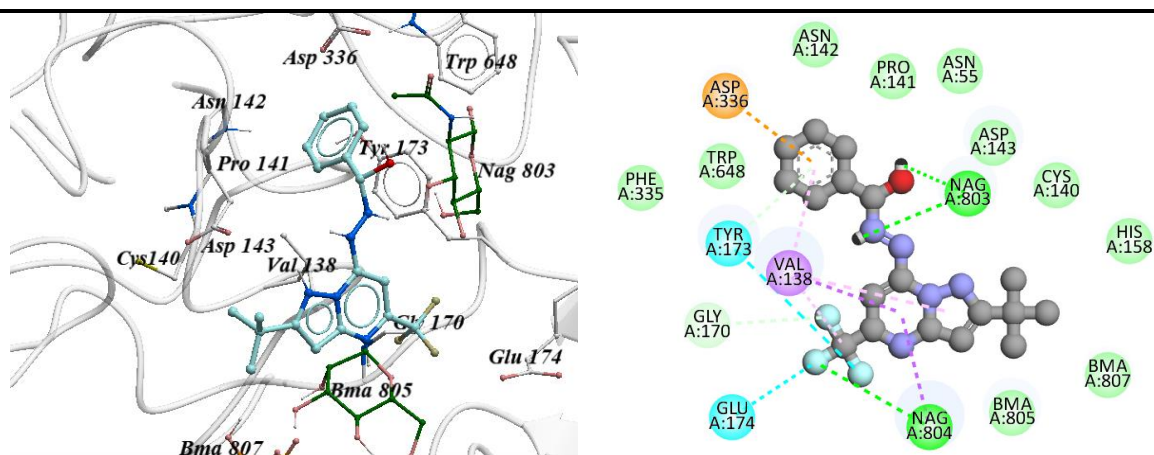
**Figure 4.12** Molecular docking 2D mode of interaction of compound 8796 N-(4,7,7-trimethyl-2-phenyl-6,8-dihydroquinolin-5-ylidene) hydroxylamine with DYIGS binding site in nicastrin analysed by Discovery Studio. Conventional hydrogen bond, van der Waals interactions, Carbon hydrogen bond, halogen, pi-sigma, alkyl, and pi-alkyl are shown in green, light green, and pink respectively.

**Figure 4.13** depicts the binding of compound 6218, 2-morpholin-4-yl-6-pyridin-3-yl-4,4-bis(trifluoromethyl)-1,3,5-oxadiazine, in nicastrin. Conventional hydrogen bonds form between Asp143, Nag803 and trifluoromethyl fluorines, Nag803 also with oxadiazine hydrogens, and Nag804 with oxadiazine and morpholine oxygens, and halogen bonds form between Cys140 and Asp336 and trifluoromethyl fluorines. Pi-alkyl interactions are observed between Val138 and the pyridine group and Trp648 and the trifluoromethyl methyl groups. Val138 also forms alkyl interactions with the trifluoromethyl group's methyl group. Van der Waals interactions are also observed with the residues Bma805, Gly144, Phe145, Asn55, Pro141, Asn142, and Tyr173.



**Figure 4.13** Molecular docking 2D mode of interaction of compound 6218 2-morpholin-4-yl-6-pyridin-3-yl-4,4-bis(trifluoromethyl)-1,3,5-oxadiazine with DYIGS binding site in nicastrin analysed by Discovery Studio. Conventional hydrogen bond, van der Waals interactions, Carbon hydrogen bond, halogen, pi-sigma, alkyl, and pi-alkyl are shown in green, light green, and pink respectively.

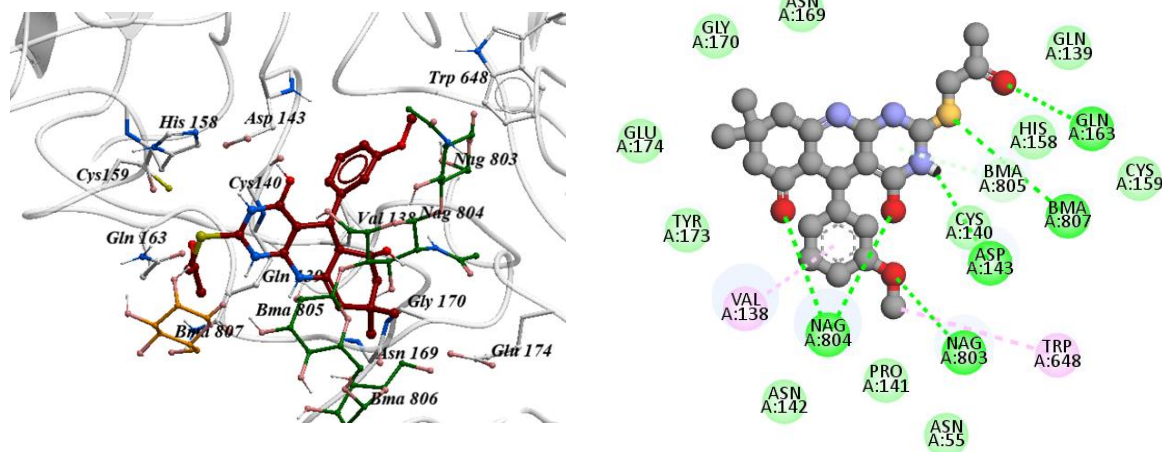
Compound 6197 (N'-[2-tert-butyl-5-(trifluoromethyl) pyrazolo[1,5-a] pyrimidin-7-yl] benzohydrazide) contains a trifluoromethyl group that forms halogen bonds with Glu174 and Tyr173, as well as conventional hydrogen bonds with Nag804 and alkyl interactions with Val138. As shown in **Figure 4.14**, the residue Val138 has numerous pi-interactions with the pyrazolopyrimidine and phenyl groups. Compound 6197 is stabilised by van der Waals interactions with the residues Phe335, Trp648, Asn142, Pro141, Asn55, Asp143, Cys140, His158, and the glycans Bma805 and 807.



**Figure 4.14** Molecular docking 2D mode of interaction of compound 6197 N'-[2-tert-butyl-5-(trifluoromethyl) pyrazolo[1,5-a] pyrimidin-7-yl] benzohydrazide with DYIGS binding site in nicastrin analysed by Discovery Studio. Conventional hydrogen bond, van der Waals interactions, Carbon hydrogen bond, halogen, pi-sigma, alkyl, and pi-alkyl are shown in green, light green, and pink respectively.

The methoxyphenyl groups in 5-(3-methoxyphenyl)-8,8-dimethyl-2-(2-oxopropylsulfanyl)-5,7,9,10-tetrahydro-3H-pyrimido[4,5-b]quinoline-4,6-dione (compound 2251) as well as the 5,7,9,10-tetrahydro-3H-pyrimido[4,5-b]quinoline form van der Waals interactions with residues such as Glu174, Gly170, Asn169, Gln139, His158, Cys159, Cys140, Asn142, Pro141 and Asn55. On the other hand, Nag804, Gln163, and Nag803 form hydrogen bonds with the diones, oxo, and methoxy substituents, respectively. Pi-alkyl interactions were observed between Val138 and Trp648 with the phenyl ring and the methyl group, respectively (**Figure 4.15**).

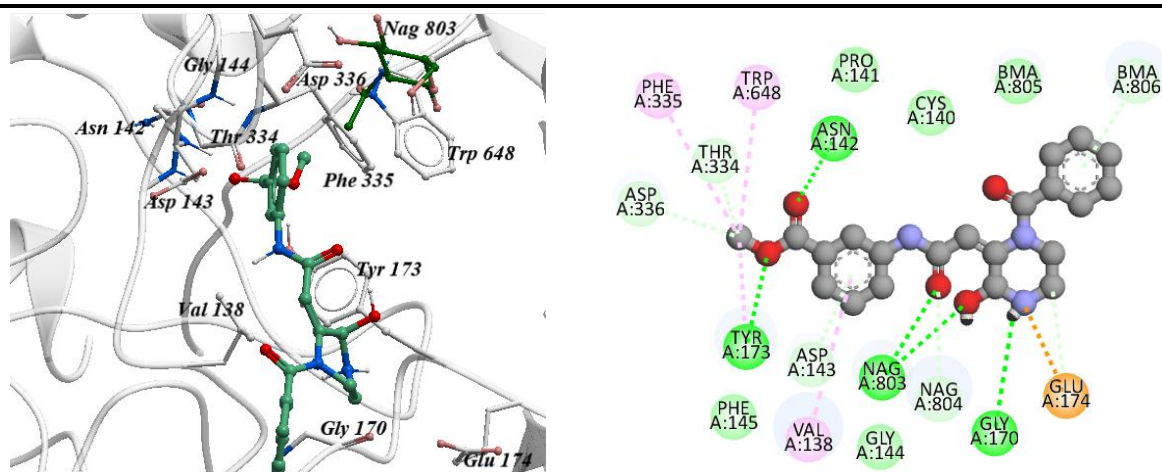




**Figure 4.15** Molecular docking 2D mode of interaction of compound 2251 5-(3-methoxyphenyl)-8,8-dimethyl-2-(2-oxopropylsulfanyl)-5,7,9,10-tetrahydro-3H pyrimido[4,5-b] quinoline-4,6-dione with DYIGS binding site in nicastrin analyzed by Discovery Studio. Conventional hydrogen bond, van der Waals interactions, Pi-Donor hydrogen bond, and pi-alkyl are shown in green, light green, and pink respectively.

Compound 8250's (methyl 3-[[2-(1-benzoyl-3-oxopiperazin-2-yl)acetyl]amino]benzoate) benzoate oxygens interact with Asn142 and Tyr173 via hydrogen bonding, methylbenzoate forms pi-alkyl interactions with Val138 through the ring and with the methyl with Phe335, Tyr173, and Trp648. In addition, the hydrogen atom bonded to the nitrogen of the oxopiperazine forms a hydrogen bond with Gly170. Nag803 forms two hydrogen bonds with oxygens on the oxopiperazine and acetyl groups. Glu174 and the nitrogen of the piperazine interact via a pi-anion contacts (**Figure 4.16**).





**Figure 4.16** Molecular docking 2D mode of interaction of compound 8250 methyl 3-[[2-(1-benzoyl-3-oxopiperazin-2-yl)acetyl]amino]benzoate with DYIGS binding site in nicastrin analysed by Discovery Studio. Conventional hydrogen bond, van der Waals interactions, Carbon hydrogen bond, halogen, pi-sigma, alkyl, and pi-alkyl are shown in green, light green, and pink respectively.

#### 4.5 Conclusion

The physicochemical properties and scaffold space of nicastrin inhibitors were investigated to inform the quantitative structure activity relationships that were used to identify potential inhibitors for breast cancer therapy from the Maybridge HitCreator dataset. The screening dataset was diverse in terms of physicochemical properties and scaffolds, implying that the Maybridge HitCreator set used in this study could be a novel source for gamma-secretase compounds. Given that nicastrin inhibitors are so diverse, they can be structurally modified to create novel breast cancer compounds. Scaffold analysis identified specific connectivity containing a sulfon, sulfonamide, or sulfonamide connected to a non-aromatic ring and a halide or a halide connected to a benzene ring as associated with high activity for nicastrin inhibition.

The J48 trees model demonstrates that the presence of heteroatoms such as halogen, nitrogen, and sulfur atoms, the degree of branching, chain length, and the presence of cyclic structures all affect the activity of nicastrin inhibitors. From this information, seven nicastrin inhibitors were identified. This study discovered scaffolds and compounds that could aid in the discovery of effective and marketable gamma-secretase inhibitors for the development of breast cancer drugs. The physicochemical property and pharmacokinetic analysis as well as antitumour tests are discussed in the following chapter.

## 4.6 References

- [1] T. Sander, J. Freyss, M. Von Korff, C. Rufener, DataWarrior: An open-source program for chemistry aware data visualization and analysis, *J. Chem. Inf. Model.* 55 (2015) 460–473. <https://doi.org/10.1021/ci500588j>.
- [2] J.B. Baell, G.A. Holloway, New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays, *J. Med. Chem.* 53 (2010) 2719–2740. <https://doi.org/10.1021/jm901137j>.
- [3] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Adv. Drug Deliv. Rev.* 23 (1997) 3–25.
- [4] D.F. Veber, S.R. Johnson, H.Y. Cheng, B.R. Smith, K.W. Ward, K.D. Kopple, Molecular properties that influence the oral bioavailability of drug candidates, *J. Med. Chem.* 45 (2002) 2615–2623. <https://doi.org/10.1021/jm020017n>.
- [5] T.I. Oprea, A.M. Davis, S.J. Teague, P.D. Leeson, Is There a Difference between Leads and Drugs? A Historical Perspective, *J. Chem. Inf. Comput. Sci.* 41 (2001) 1308–1315. <https://doi.org/10.1021/ci010366a>.
- [6] M. Congreve, R. Carr, C. Murray, H. Jhoti, A “Rule of Three” for fragment-based Lead discovery. *Drug Discov. Today.* 8 (2003) 876. [https://doi.org/10.1016/S1359-6446\(03\)02765-X](https://doi.org/10.1016/S1359-6446(03)02765-X).
- [7] S. Milhas, B. Raux, S. Betzi, C. Derviaux, P. Roche, A. Restouin, M.J. Basse, E. Rebuffet, A. Lugari, M. Badol, R. Kashyap, J.C. Lissitzky, C. Eydoux, V. Hamon, M.E. Gourdel, S. Combes, P. Zimmermann, M. Aurrand-Lions, T. Roux, C. Rogers, S. Müller, S. Knapp, E. Trinquet, Y. Collette, J.C. Guillemot, X. Morelli, Protein-Protein Interaction Inhibition (2P2I)-Oriented Chemical Library Accelerates Hit Discovery, *ACS Chem. Biol.* 11 (2016) 2140–2148. <https://doi.org/10.1021/acscchembio.6b00286>.
- [8] G.W. Bemis, M.A. Murcko, The properties of known drugs. 1. Molecular frameworks, *J. Med. Chem.* 39 (1996) 2887–2893. <https://doi.org/10.1021/jm9602928>.
- [9] E. Frank, M.A. Hall, I.H. Witten, T. Weka, Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). *The WEKA Workbench. Online Appendix for ‘Data Mining: Practical Machine Learning Tools and Techniques’*, Morgan Kaufmann, Fourth Edition, 2016., Forth, 2016.
- [10] S.L. Salzberg, Book Review: C4.5: by J. Ross Quinlan. Inc., 1993., in: Mach. Learn., 1994: pp. 235–240. [https://doi.org/10.1016/S0019-9958\(64\)90259-1](https://doi.org/10.1016/S0019-9958(64)90259-1).
- [11] X. Xia, E.G. Maliski, P. Gallant, D. Rogers, Classification of kinase inhibitors using a Bayesian model, *J. Med. Chem.* 47 (2004) 4463–4470. <https://doi.org/10.1021/jm0303195>.
- [12] R.N. Roy, J.; Kar, S.; Das, Feature combination networks for the interpretation of statistical machine learning models: Application to Ames mutagenicity, in:

- Springer Briefs *Mol. Sci.*, 2015: pp. 37–60.
- [13] J.C. Platt, Sequential Minimal Optimization: A Fast Algorithm for Training Support, (1998) 1–21.
- [14] S. Ghanbarzadeh, S. Ghasemi, A. Shayanfar, H. Ebrahimi-najafabadi 2D-QSAR study of some 2, 5-diamino benzophenone farn esyltrans ferase inhibitors by, (2015) 484–495.
- [15] M.C. Sharma, 2D QSAR studies of the inhibitory activity of a series of substituted purine derivatives against c-Src tyrosine kinase, *Integr. Med. Res.* 10 (2016) 563–570. <https://doi.org/10.1016/j.jtusci.2015.11.002>.
- [16] F. Pereira, D.A.R.S. Latino, S.P. Gaudêncio, QSAR-Assisted Virtual Screening of Lead-Like Molecules from Marine and Microbial Natural Sources for Antitumour and Antibiotic Drug Discovery, (2015) 4848–4873. <https://doi.org/10.3390/molecules20034848>.
- [17] A. Golbraikh, E. Muratov, D. Fourches, A. Tropsha, Dataset Modelability by QSAR, *J Chem Inf Model.* 54 (2014) 1–8. <https://doi.org/10.1007/s10955-011-0269-9.Quantifying>.
- [18] T.M. Martin, P. Harten, D.M. Young, E.N. Muratov, A. Golbraikh, H. Zhu, A. Tropsha, Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling, *J. Chem. Inf. Model.* 52 (2012) 2570–2578.
- [19] P.M. Khan, K. Roy, Current approaches for choosing feature selection and learning algorithms in quantitative structure–activity relationships (QSAR), *Expert Opin. Drug Discov.* 13 (2018) 1075–1089. <https://doi.org/10.1080/17460441.2018.1542428>.
- [20] G.M. Morris, R. Huey, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell, A.J. Olson, AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility, *J. Comput. Chem.* 30 (2010) 2785–2791. <https://doi.org/10.1002/jcc.21256.AutoDock4>.
- [21] T.A. Halgren, Performance of MMFF94\*, Scope, Parameterization, *J. Comput. Chem.* 17 (1996) 490–519. <http://journals.wiley.com/jcc>.
- [22] Y. Hu, D. Stumpfe, J. Bajorath, Computational Exploration of Molecular Scaffolds in Medicinal Chemistry, *J. Med. Chem.* 59 (2016) 4062–4076. <https://doi.org/10.1021/acs.jmedchem.5b01746>.
- [23] R. Guha, J.H. Van Drie, Structure - Activity landscape index: Identifying and quantifying activity cliffs, *J. Chem. Inf. Model.* 48 (2008) 646–658. <https://doi.org/10.1021/ci7004093>.
- [24] S. Egieyeh, J. Syce, S.F. Malan, A.C. Id, Predictive classifier models built from natural products with antimalarial bioactivity using machine learning approach, (2018) 1–15.
- [25] B. Viira, T. Gendron, D.A. Lanfranchi, S. Cojean, D. Horvath, G. Marcou, A. Varnek, L. Maes, U. Maran, P.M. Loiseau, E. Davioud-charvet, In Silico Mining for Antimalarial Structure-Activity Antimalarial Curcuminoids, *Molecules.* 21 (2016) 1–18. <https://doi.org/10.3390/molecules21070853>.

- [26] W. Luo, S. Medrek, J. Misra, G.J. Nohynek, Predicting human skin absorption of chemicals: Development of a novel quantitative structure activity relationship, *Toxicol. Ind. Health*. 23 (2007) 39–45.  
<https://doi.org/10.1177/0748233707077430>.

## **5 Biological evaluation, and physicochemical and pharmacokinetic property profiling of hit compounds**

### **5.1 Introduction**

In the previous chapter, seven compounds were identified as potential anticancer hits through virtual screening. In this chapter, the seven compounds were analysed in silico for their physicochemical and pharmacokinetic properties, as well as biologically for antibacterial susceptibility and antitumour activity using carrot disc assays. Physicochemical properties such as molecular weight, hydrogen bond donor, hydrogen bond acceptor, lipophilicity, aqueous solubility, topological polar surface area, number of rotatable bonds, and molar reactivity were used to evaluate drug-likeness.

Pharmacokinetic properties were used to assess the compounds' absorption, distribution, metabolism, excretion, and toxicity. Human intestinal absorption was predicted using the Caco-2 permeability and the Madin-Darby Canine Kidney cell permeability coefficient. The BOILED-EGG model was used to predict passive human intestinal absorption, blood brain barrier penetration, and permeability glycoprotein substrates. The drug distribution was calculated using plasma protein binding and equilibrium bound, and unbound serum proteins. The metabolism of drugs was evaluated by determining whether the compounds were inhibitors, non-inhibitors, substrates, or non-substrates of the CYP450 system.

The half-life and clearance of the compound were used to calculate its excretion. Compound toxicity was determined by calculating the likelihood of hERG blockage and human hepatotoxicity. To assess the antibacterial susceptibility of hits, a biological

antibacterial evaluation was performed using a slightly modified Bauer-Kirby method. The carrot disc assay was then used as a preliminary antitumour assay.

## **5.2 Methods**

### **5.2.1 Culturing of the *Agrobacterium tumefaciens***

The previously isolated and characterised *A. tumefaciens* was cultured on Mackonkey agar and then subcultured on nutrient agar media to produce pure culture of the *agrobacterium* in the laboratory at the University of Malawi, Chancellor College, Microbiology Laboratory. Using a loop, a single colony was picked from the sub-culture and inoculated into nutrient agar media using the streak plate method. <sup>[1]</sup> The process was repeated and from each plate, forming one replica. Following this, the petri dishes were kept in the incubator for three days at 28 °C. After incubation, the pure culture appeared on nutrient agar media, which was later used for the various tests for *Agrobacterium tumefaciens* confirmation, and slants were sub-cultured on nutrient agar media for storage.

### **5.2.2 Test for antibiotic resistance of the hits**

The antibiotic sensitivity of the target compounds was determined using a slightly modified Bauer-Kirby method. <sup>[2]</sup> Tetracycline (30 µg/L) and gentamicin (10 µg/L) were used as antibiotics. Whatman No. 1 filter paper discs of 6 mm in diameter were impregnated with 10 µL of the antibiotic solution and the target compounds at concentrations of 100 µM and 10 µM, respectively, before being air dried. The disc was then placed on seeded Luria-Bertani (LB) agar plates. 20 µL standard bacteria cultures (10<sup>8</sup> cfu/ml) were used for preparing seeded agar plates. The petri dishes were incubated at 30 °C for 24 hrs. The antibiotic susceptibility was determined by

measuring the size of the inhibition zone. The inhibition zones were measured using a digital Vernier Caliper and interpreted as Susceptible (S), Intermediate (I) and Resistance (R) based on the Bauer-Kirby protocol.

### 5.2.3 Carrot disc assay

Antitumour activity was evaluated by the procedure followed by Hussain et al [3] and Lellau and Liebezeit<sup>[4]</sup> with slight modifications. A fresh culture of *A. tumefaciens* was prepared by inoculating 100 mL (1.3%) autoclaved nutrient broth pH 7.4 with 10 µL of stock culture. This media was incubated at 28 °C for 48 hours to yield 5 x10<sup>9</sup> cells per mL. Carrots were surface sterilised with 20% HgCl<sub>2</sub> solution 20 min before being cut (5 x 5) mm by a sterilised cork borer.

Seven carrot disks treated with 10 µM or 100 µM compound along with positive and negative controls in the center were placed in autoclaved petri dish containing 1.5% agar medium. Samples 300 µL of each was mixed with 50 µL of cultured *A. tumefaciens* and poured 50 µL of each on the carrot disks. The entire experimental work was carried out in laminar air flow. Petri plates were incubated at 28 °C for 21 days and sprayed with Lugol's solution (potassium iodide 10% and Iodine 5% in distilled water). Tumours were counted under a microscope. Each sample was replicated five times. The percentage inhibition was calculated using **Equation 5.1**.

$$\% \text{ Inhibition} = \left( 1 - \frac{\text{no. of tumors in sample}}{\text{no. of tumors in negative control}} \right) \times 100$$

**Equation 5.1**



#### **5.2.4 Assessment of oral availability and pharmacokinetic property evaluation of hits**

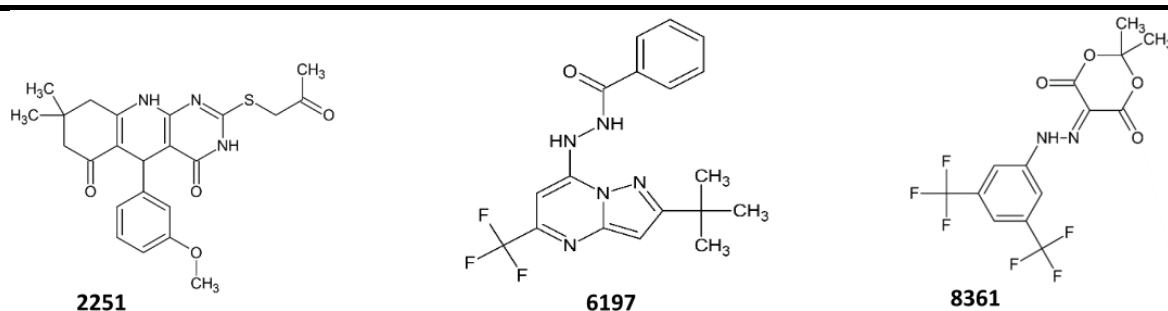
The aim is to identify hits that are less toxic, have good oral bioavailability and have optimum physicochemical properties. The SWISS ADME <sup>[5]</sup> and ADMETlab 2.0 <sup>[6]</sup> servers were used to predict the ADME properties. Lipinski rule of 5, <sup>[7]</sup> bioavailability score, <sup>[8]</sup> Ghose's <sup>[9]</sup> and Veber's <sup>[10]</sup> rules were used to guide the calculation of physicochemical properties. Molecular parameters such as molecular weight, hydrogen bond donor, hydrogen bond acceptor, lipophilicity, aqueous solubility, and topological polar surface area, number of rotatable bonds and molar reactivity were used to assess drug-likeness.

Volume of distribution ( $V_{DSS}$ ) and plasma proteins ( $F_u$ ) were calculated to determine the distribution of derivatives. Data for major human cytochrome P450 (CYP) isoforms involved in drug metabolism, including CYP2C9, CYP2D6 and CYP3A4, were also generated. Excretion routes of the compounds were determined by predicting the total clearance as well as the renal OCT2 substrate. Toxicity was determined by calculating LD50, hepatotoxicity, skin sensitisation, cellular toxicity and hERG liability for each compound. In addition, the BOILED-EGG model <sup>[11]</sup> of the molecules was predicted to reveal the capacity of gastrointestinal absorption and the permeability of the blood brain penetration barrier, both of which are key parameters in design of anticancer compounds.

## 5.3 Results and discussion

### 5.3.1 Biological evaluation of antitumor properties

The antibiotic resistance test was initially done to check for antibacterial activity of the identified hits, that could otherwise be misinterpreted as antitumour activity by inhibiting the growth of *A. tumefaciens* and cell viability. **Table 5.1** shows the hits to which the *A. tumefaciens* was susceptible and resistant. The three susceptible compounds identified (**Figure 5.1**) were **2251**; 5-(3-methoxyphenyl)-8,8-dimethyl-2-(2-oxopropylsulfanyl)-5,7,9,10-tetrahydro-1H-pyrimido[4,5-b] quinoline-4,6-dione (26 mm); **8361**; 5-[[3,5-Bis(trifluoromethyl)phenyl] hydrazinylidene]-2,2-dimethyl-1,3-dioxane-4,6-dione (21 mm); and **6197** N'-[2-tert-butyl-5-(trifluoromethyl) pyrazolo[1,5-a] pyrimidin-7-yl] benzohydrazide (22 mm). These compounds had zones of inhibition which were comparable to the standard antibiotics used in the viability test, tetracycline and gentamicine with zones of inhibition of 24 mm and 22 mm respectively.

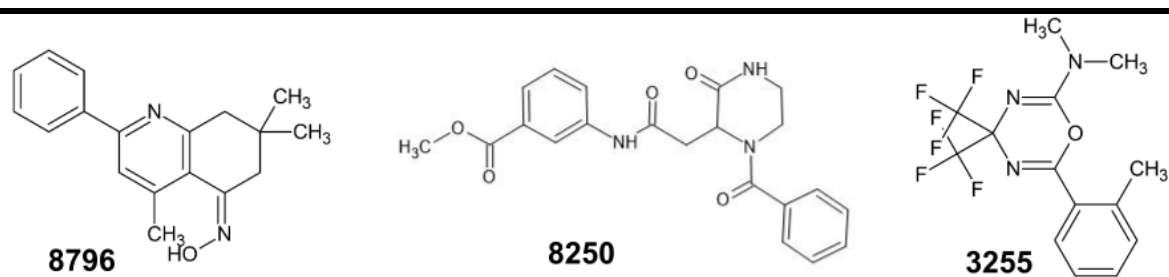


**Figure 5.1** Hit compounds with antibacterial susceptibility.

The seven hits were then tested for antitumour activity and **Table 5.1** shows the effects of the hits on the crown gall tumour inhibition on carrot discs. The percentage

inhibitions ranged from 15% to 87.5%. **6218**; 2-Morpholino-6-(3-pyridyl)-4,4-bis(trifluoromethyl)-4H-1,3,5-oxadiazine had the least percentage inhibition of 15%, whilst **8796**; N-(4,7,7-trimethyl-2-phenyl-6,8-dihydroquinolin-5-ylidene) hydroxylamine had the highest percentage inhibition of 87.5%. Compounds that inhibited the A.

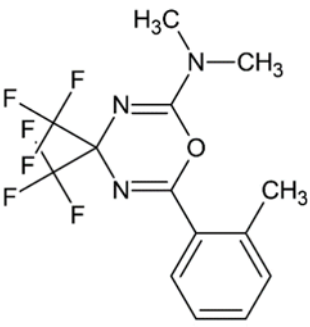
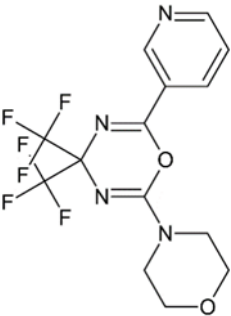
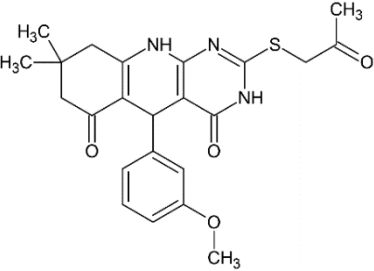
*tumefaciens* with %inhibition greater than 20% and without antibacterial susceptibility are shown in **Figure 5.2**.

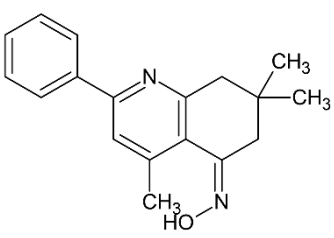
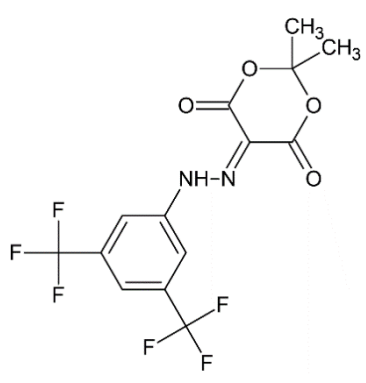
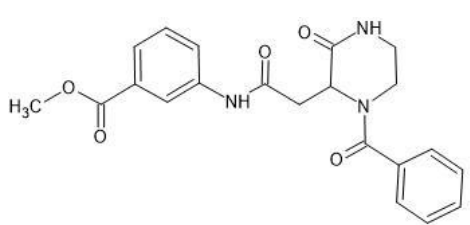
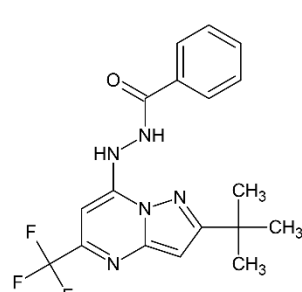


**Figure 5.2** The three active hits.

---

**Table 5.1** Resistance to antibiotics and percentage inhibition of *A. tumefaciens* by the identified hits

Compound ID	Zone of Bacterial Inhibition (mm)		Antibacterial Inhibition	% Inhibition of Tumours
	100µM	10µM		
 <p><b>3255</b></p>	6.69	6.62	R	45.00 ± 05.60
 <p><b>6218</b></p>	7.54	6.83	R	15.00 ± 10.21
 <p><b>2251</b></p>	26.40	27.99	S	57.00 ± 08.33

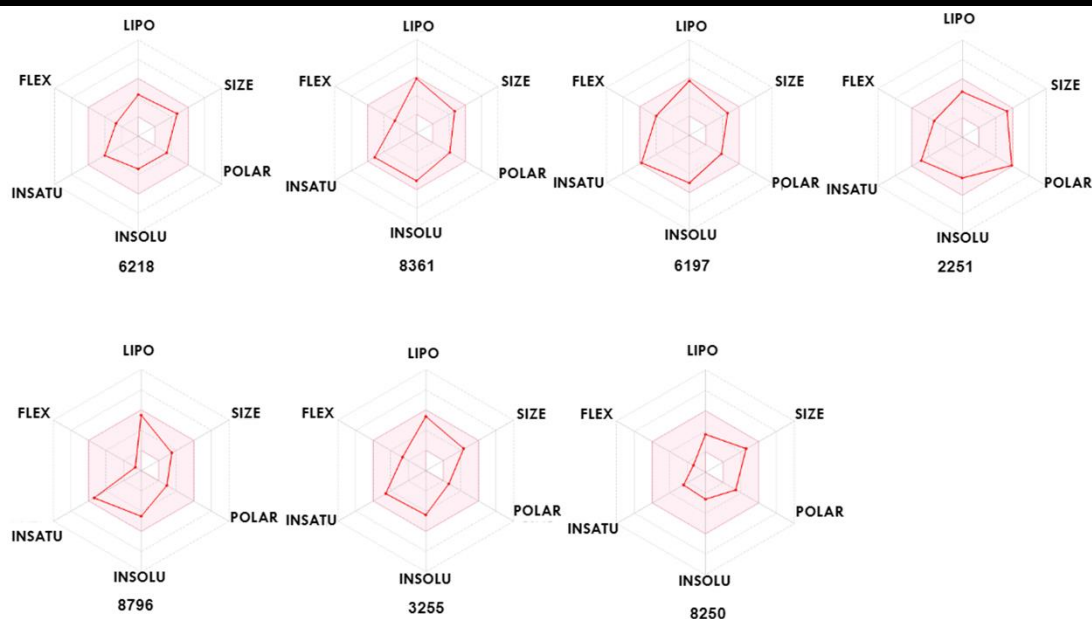
 <p style="text-align: center;"><b>8796</b></p>	8.18	9.42	R	87.50 ± 05.50
 <p style="text-align: center;"><b>8361</b></p>	21.46	23.06	S	30.02 ± 10.20
 <p style="text-align: center;"><b>8250</b></p>	7.53	8.15	R	79.00 ± 05.66
 <p style="text-align: center;"><b>6197</b></p>	21.50	22.11	S	52.50 ± 11.82
<b>Tetracycline</b>	23.92		S	
<b>Gentamicin</b>	21.60		S	

### 5.3.2 Physicochemical property prediction

To increase the chances of the identified hits passing through pre-clinical and clinical phases of drug development, the physicochemical properties for bioavailability [12,13] were evaluated together with those of 20 known orally available, and FDA approved, breast cancer drugs. The important physicochemical properties that were evaluated are molecular weight (MW), number of rotatable bonds (nRot), number of hydrogen bond acceptors (HBA) and donors (HBD), topological polar surface area (TPSA), octanol-water partition coefficient (logP), aqueous solubility (logS). [5]

The seven hit compounds, FDA approved and orally available breast cancer drugs were subjected to the SwissADME webtool, and the compressed view of these properties is presented graphically for the seven hits only using bioavailability radar plots presented in **Figure 5.3**. Different properties that include solubility, flexibility, lipophilicity, saturation, and molecular weight were predicted and did not show any outliers outside the shaded pink limits for the identified hits. However, there were deviations in the approved breast cancer drugs.

The already approved drugs are generally soluble, but nearly half of them have a MW greater than 500 g/mol; TPSA values greater than 140 Å and are flexible. These findings suggest that the seven hits have good oral bioavailability and acceptable drug-likeness properties, as they are small, relatively soluble, have acceptable hydrogen bond donor propensity, a low polar surface area in the range 37-80 Å and are relatively flexible with 1 to 7 rotatable bonds.



**Figure 5.3** Bioavailability radar plots of hit compounds evaluated using swissADME webtool. Lipophilicity (LIPO): XLOGP3 between -0.7 and +5.0, Molecular weight (MW) (SIZE): 150-500 g/mol; Polarity (POLAR): TPSA between 20 and 130; Solubility (INSOLU): log S not higher than 6; Saturation (INSATU): fraction of carbons in the sp<sup>3</sup> hybridization not less than 0.25; and Flexibility (FLEX): no more than 9 rotatable bonds.

### 5.3.3 Prediction of Absorption, Distribution, Metabolism, Excretion and Toxicity

#### 5.3.3.1 Absorption and Distribution

For an oral drug to enter systematic circulation, it must pass through intestinal cell membranes via passive diffusion, carrier-mediated uptake, or active transport. <sup>[14]</sup> The human intestinal absorption of compounds was estimated by calculating Caco-2 permeability. <sup>[14, 15]</sup> Caco-2 cells are a human colon epithelial cancer cell line that has predicted values greater than -5.15 log cm/s and used as a model of human intestinal



absorption. The Madin-Darby Canine Kidney cells (MDCK) were also used as an in vitro model for permeability screening, with a permeability coefficient,  $P_{app} > 20 \times 10^{-6}$  cm/s considered to have high passive MDCK permeability. All seven hits were predicted to permeate the intestinal cell membranes because they had Caco-2 values greater than  $-5.15 \log \text{ cm/s}$  and MDCK  $\log P_{app}$  greater than  $20 \times 10^{-6}$  cm/s, whereas 42% of the approved breast cancer drugs were predicted to have low Caco-2 permeability and 74% exhibited medium passive MDCK permeability. These results are shown in **Table A1 and A2, Appendix**.

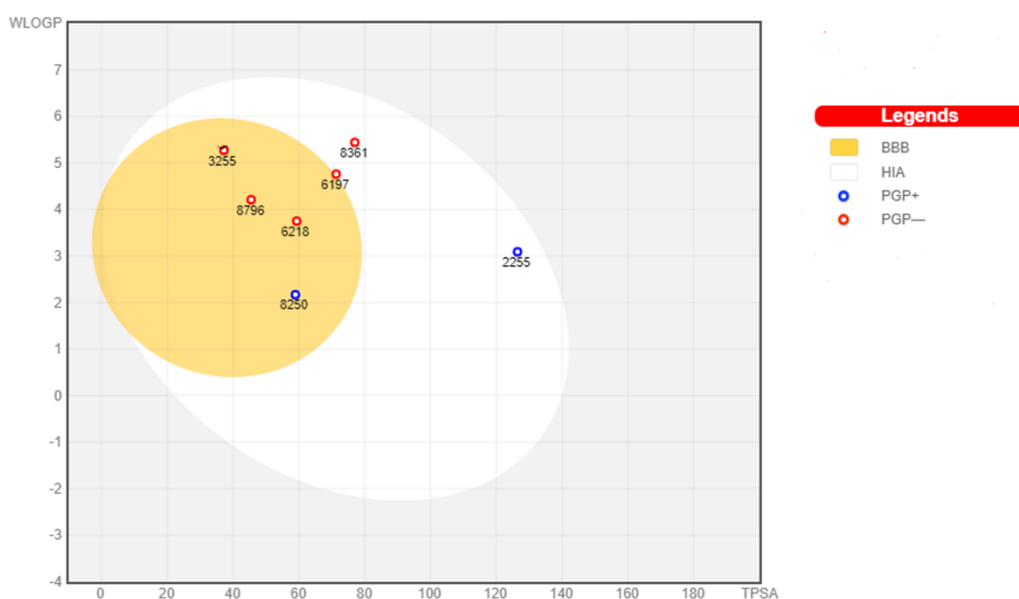
A prevalent mechanism of drug resistance in cancer is due to the over-expression of energy dependent efflux pumps, such as Pgp transporters which are modulated or inhibited by several anticancer drugs. This causes anticancer drugs to be pumped into the extracellular space, leaving a very low concentration of the drugs within the cytoplasm. As such, anticancer drugs should be non-substrates of the Pgp protein.

Compounds that are non substrates of Pgp protein are preferred because drug resistance in cancer, particularly in patients with metastatic cancers, occurs intrinsically due to the presence of energy dependent efflux pumps in cancer cell populations. <sup>[16,17]</sup> Henneman et al., <sup>[18]</sup> explained this by observing that metaplastic breast carcinoma showed intense resistance to Olaparib due to increased Pgp drug efflux transporter expression. From the *in-silico* evaluation, using the BOILED-EGG model, all of the selected hits are non-substrates of the permeability glycoprotein, Pgp, as indicated by the red spheres in **Figure 5.4**, except for compounds 8250 and 2251. Considering the approved and orally available breast cancer drugs, 48% of them are Pgp substrates, implying that low concentrations of these drugs are available for activity, while the rest are pumped out of the cell.

Further, predictions for passive human gastrointestinal absorption (HIA) and blood brain barrier (BBB) were explored using the BOILED-Egg model as shown in **Figure 5.4**. In the BOILED-Egg model, compounds are predicted to penetrate the intestinal cell membranes if they are in the white area and conversely, the BBB if they are in the yellow region. Four of the selected hits (compound 3255, 8796, 6218 and 8250) are predicted to have good blood brain barrier penetration and a high human gastrointestinal absorption, since the white and yellow regions are not mutually exclusive.

The remaining three compounds (6197, 8361 and 2251) have a high gastrointestinal absorption but do not penetrate the BBB. However, 85% of approved breast cancer drugs do not penetrate the BBB, and 28% have low gastrointestinal absorption and do not exhibit BBB permeation. The interest in hits that can cross the BBB stems from the fact that, amongst other cancers, breast cancer is the leading cause of brain metastases, which is the leading cause of death in breast cancer patients.<sup>[19–21]</sup> Given that most chemotherapy does not penetrate the BBB, there is a need to design drugs that do.

The availability of drugs depends on their distribution between their binding with plasma proteins (PPB); the equilibrium of bound and unbound serum proteins ( $F_u$ ) as well as the distribution of drugs (VD). If the PPB is predicted to be less than 90% then the therapeutic index of that drug is assumed to be high. Half of the hits (3255, 6218 and 8250) had a PPB less than 90%. All the hits had a VD within the acceptable range of 0.004-20 L/Kg. The fraction unbound was greater than 5% for all the hits except for Compound 6218.



**Figure 5.4** The BOILED-Egg model of the selected hit compounds.

### 5.3.3.2 Metabolism

Drug metabolism occurs in a variety of locations throughout the body, including the liver, intestinal walls, lungs, kidneys, and plasma. [22] However, the liver is regarded as the primary site of drug metabolism, detoxification, and xenobiotic excretion, as catalysed by the CYP450 system. [23] In the presence of another drug metabolised by the same pathway, drugs with CYP450 activity can induce drug-drug interactions, which can alter the metabolism of concurrently administered drugs resulting in drug toxicity or lowering drug concentration to the point of treatment failure. The results of the *in-silico* evaluation of the hit compounds are shown in **Table A 3** in the Appendix section, and indicates whether the hits are inhibitors, non-inhibitors and substrates or non substrates of the CYP450 family.

These results can be used as a guide when the hits are prescribed in combination with other drugs. Tamoxifen, an approved breast cancer drug, for example, when taken in combination with CYP2D6 inhibitors results in reduced tamoxifen activity due to drug-drug interactions. [24] According to the analysis, all of the compounds studied interfere with different CYP450 isoforms, but all of the hits, with the exception of compound 2251, are non substrates or non-inhibitors of at least 50% of the CYP enzymes studied, which was comparable to the set of breast cancer drugs used in this study which had 85% of the drugs as non substrates or non-inhibitors of the CYP family of enzymes.

#### **5.3.3.3 Excretion**

The excretion of a compound is an important parameter that must be studied as it describes the volume of distribution, half-life and frequency of dosing of a drug. [25] According to the results of the ADMETlab2.0 server analysis (**Table A 4, Appendix**), compounds with clearance (CL) penetration values greater than 15; between 5 and 15, and less than 5 were classified as having high; moderate and low clearance. Compounds 3255, 6197, 8250 were predicted to have a good clearance whereas 2251, 6218, 8361 and 8796 were predicted to have low clearance. In general, the hit compounds and orally available breast cancer drugs were predicted to have short half-lives of less than 3 hours, with compounds 3255 and 6218 having the shortest half-lives and 26% of approved breast cancer drugs having the longest half-lives.

#### **5.3.3.4 Toxicity**

The probability of compounds blocking hERG, a gated potassium channel responsible for the regulation of the exchange of cardiac action potential and resting potential, [26]

as well as the probability of human hepatotoxicity (H-HT) was assessed for both the hit compounds as well as the approved breast cancer drugs. The results suggest that all the hit compounds have a very low probability of blocking hERG, whilst compounds 2251, 6197 and 8361 have a high probability of inducing liver injury. When compared with the existing breast cancer drugs, 63% of the breast cancer drugs have a high probability of blocking hERG and all the breast cancer drugs have a high potential of inducing liver injury.

#### **5.4 Conclusion**

The physicochemical and pharmacokinetic antitumour properties of the seven hit compounds discovered through virtual screening were investigated. Four of the five compounds with antibacterial resistance and antitumour activity, compounds 3255, 8796, and 8250 inhibited tumour growth by more than 20%. Except for compound 8796 with a low therapeutic index and low clearance, and compound 8250 predicted to be a Pgp substrate. The four compounds were predicted to have good oral bioavailability, to be non-hERG inhibitors, and non-hepatotoxic; have short half-lives, good clearance, and high therapeutic indices. However, the three compounds 2251, 6197, and 8361 that demonstrated antibacterial susceptibility were predicted to be hepatotoxic, have long half-lives, have a high therapeutic index, and have low clearance. According to this analysis, compounds 3255, 8796, and 8250 can serve as good starting points for *in vivo* anticancer tests.

## 5.5 References

- [1] J.T. Weld, A. Gunther, A streak plate method for determining growth curves, *J. Lab. Clin. Med.* 32 (1947) 1139–1152.
- [2] A.W. Bauer, W.M.M. Kirby, J.C. Sherris, M. Turck, Antibiotic susceptibility testing by a standardized single disk method, *Am. J. Clin. Pathol.* 45 (1966) 493–496. <https://doi.org/10.1308/rcsann.2013.95.7.532>.
- [3] A. Hussain, M. Zia, B. Mirza, Cytotoxic and antitumour potential of *Fagonia cretica* L., *Turkish J. Biol.* 31 (2007) 19–24.
- [4] T.F. Lellau, G. Liebezeit, Alkaloids, Saponins and Phenolic compounds in salt marsh plants from the lower Saxonian Wadden Sea, *Senckenbergiana Maritima.* 31 (2001) 1–9. <https://doi.org/10.1007/BF03042831>.
- [5] A. Daina, O. Michielin, V. Zoete, SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules, *Sci. Rep.* 7 (2017) 1–13. <https://doi.org/10.1038/srep42717>.
- [6] G. Xiong, Z. Wu, J. Yi, L. Fu, Z. Yang, C. Hsieh, M. Yin, X. Zeng, C. Wu, A. Lu, X. Chen, T. Hou, D. Cao, ADMETlab 2.0: An integrated online platform for accurate and comprehensive predictions of ADMET properties, *Nucleic Acids Res.* 49 (2021) W5–W14. <https://doi.org/10.1093/nar/gkab255>.
- [7] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Adv. Drug Deliv. Rev.* 64 (2012) 4–17. <https://doi.org/10.1016/j.addr.2012.09.019>.
- [8] Y.C. Martin, A bioavailability score, *J. Med. Chem.* 48 (2005) 3164–3170. <https://doi.org/10.1021/jm0492002>.
- [9] A.K. Ghose, T. Herbertz, J.M. Salvino, J.P. Mallamo, Knowledge-based chemoinformatic approaches to drug discovery, *Drug Discov. Today.* 11 (2006) 1107–1114. <https://doi.org/10.1016/j.drudis.2006.10.012>.
- [10] D.F. Veber, S.R. Johnson, H.Y. Cheng, B.R. Smith, K.W. Ward, K.D. Kopple, Molecular properties that influence the oral bioavailability of drug candidates, *J. Med. Chem.* 45 (2002) 2615–2623. <https://doi.org/10.1021/jm020017n>.
- [11] A. Daina, V. Zoete, A BOILED-Egg to Predict Gastrointestinal Absorption and Brain Penetration of Small Molecules, *ChemMedChem.* 11 (2016) 1117–1121.
- [12] C. Jia, J. Li, G. Hao, G. Yang, A drug-likeness toolbox facilitates ADMET study in drug discovery, *Drug Discov. Today.* (2019). <https://doi.org/10.1016/j.drudis.2019.10.014>.
- [13] F. Zafar, A. Gupta, K. Thangavel, K. Khatana, A.A. Sani, A. Ghosal, P. Tandon, N. Nishat, Physicochemical and Pharmacokinetic Analysis of Anacardic Acid Derivatives, *ACS Omega.* 5 (2020) 6021–6030. <https://doi.org/10.1021/acsomega.9b04398>.

- [14] T. Sharma, S. Jana, Investigation of Molecular Properties that Influence the Permeability and Oral Bioavailability of Major  $\beta$ -Boswellic Acids, *Eur. J. Drug Metab. Pharmacokinet.* 45 (2020) 243–255. <https://doi.org/10.1007/s13318-019-00599-z>.
- [15] D. Vidović, N. Milošević, N. Pavlović, N. Todorović, J.Č. Panić, J. Ćurčić, N. Banjac, N. Trišović, B. Božić, M. Lalić-Popović, In silico–in vitro estimation of lipophilicity and permeability association for succinimide derivatives using chromatographic anisotropic systems and parallel artificial membrane permeability assay, *Biomed. Chromatogr.* 36 (2022). <https://doi.org/10.1002/bmc.5413>.
- [16] C. Karthika, R. Sureshkumar, M. Zehravi, R. Akter, F. Ali, S. Ramproshad, B. Mondal, M.K. Kundu, A. Dey, M.H. Rahman, A. Antonescu, S. Cavalu, Multidrug Resistance in Cancer Cells: Focus on a Possible Strategy Plan to Address Colon Carcinoma Cells, *Life.* 12 (2022) 1–15. <https://doi.org/10.3390/life12060811>.
- [17] A. Catalano, D. Iacopetta, J. Ceramella, D. Scumaci, F. Giuzio, C. Saturnino, S. Aquaro, C. Rosano, M.S. Sinicropi, Multidrug Resistance (MDR): A Widespread Phenomenon in Pharmacological Therapies, *Molecules.* 27 (2022) 1–18. <https://doi.org/10.3390/molecules27030616>.
- [18] L. Henneman, M.H. Van Miltenburg, E.M. Michalak, T.M. Braumuller, J.E. Jaspers, A.P. Drenth, R. De Korte-Grimmerink, E. Gogola, K. Szuhai, A. Schlicker, R. Bin Ali, C. Pritchard, I.J. Huijbers, A. Berns, S. Rottenberg, J. Jonkers, Selective resistance to the PARP inhibitor olaparib in a mouse model for BRCA1-deficient metaplastic breast cancer, *Proc. Natl. Acad. Sci. U. S. A.* 112 (2015) 8409–8414. <https://doi.org/10.1073/pnas.1500223112>.
- [19] Y. Wang, F. Ye, Y. Liang, Q. Yang, Breast cancer brain metastasis: insight into molecular mechanisms and therapeutic strategies, *Br. J. Cancer.* 125 (2021) 1056–1067. <https://doi.org/10.1038/s41416-021-01424-8>.
- [20] S. Sharma, S.Y. Wu, H. Jimenez, F. Xing, D. Zhu, Y. Liu, K. Wu, A. Tyagi, D. Zhao, H.W. Lo, L. Metheny-Barlow, P. Sun, J.D. Bourland, M.D. Chan, A. Thomas, A. Barbault, R.B. D'Agostino, C.T. Whitlow, V. Kirchner, C. Blackman, B. Pasche, K. Watabe, Ca<sup>2+</sup> and CACNA1H mediate targeted suppression of breast cancer brain metastasis by AM RF EMF, *EBioMedicine.* 44 (2019) 194–208. <https://doi.org/10.1016/j.ebiom.2019.05.038>.
- [21] C. Bailleux, L. Eberst, T. Bachelot, Treatment strategies for breast cancer brain metastases, *Br. J. Cancer.* 124 (2021) 142–155. <https://doi.org/10.1038/s41416-020-01175-y>.
- [22] Y. He, H.L. McLeod, Pharmacokinetics for the prescriber, *Med. (United Kingdom).* 44 (2016) 407–411. <https://doi.org/10.1016/j.mpmed.2016.04.009>.
- [23] A.M. McDonnell, PharmD, BCOP, C.H. Dang, PharmD, BCPS, Basic Review of the Cytochrome P450 System, *J. Adv. Pract. Oncol.* 4 (2013) 263–268. <https://doi.org/10.6004/jadpro.2013.4.4.7>.
- [24] F. Esteves, J. Rueff, M. Kranendonk, The central role of cytochrome p450 in xenobiotic metabolism—a brief review on a fascinating enzyme family, *J.*

- Xenobiotics*. 11 (2021) 94–114. <https://doi.org/10.3390/jox11030007>.
- [25] F. Broccatelli, C. E.C.A Hop, M. Wright, Strategies to optimize drug half-life in lead candidate identification, *Expert Opin. Drug Discov.* 14 (2019) 221–230. <https://doi.org/10.1080/17460441.2019.1569625>.
- [26] S. Su, J. Sun, Y. Wang, Y. Xu, Cardiac hERG K<sup>+</sup> Channel as Safety and Pharmacological Target, *Handb. Exp. Pharmacol.* 267 (2021) 139–166. [https://doi.org/10.1007/164\\_2021\\_455](https://doi.org/10.1007/164_2021_455).



## 6 Summary and Conclusions

### 6.1 Introduction

Although nicastrin, a gamma-secretase component, is a validated target for breast cancer therapy, binding data for small molecules known to target nicastrin show that the compounds were designed for the gamma-secretase complex in general. Binding modes and interactions for these known inhibitors, as well as their possible binding sites, are not available to inform the design of nicastrin specific inhibitors. In this study, chemogenomic means were used to identify binding sites in nicastrin and design nicastrin specific compounds.

#### 6.1.1 Nicastrin binding sites and binding modes and interactions of known inhibitors

Blind docking predicted three distinct binding sites in both conformers, which are located in similar locations. The identified sites (**Table 3.1**) encompass domains or signature regions within nicastrin that are specific to their function (**Figure 3.2A**). These include a site that contains the DYIGS signature (DYIGS site) and the TPR-like site including a potential binding site positioned in a central cleft in the hinge region (Hinge region site). The DYIGS and Hinge sites had DLID scores that were favorable for binding drug-like molecules, however, the volume of the Hinge site was small relative to that of small molecules used in the study. The TPR-like site, though having the second-largest volume had negative DLID scores in both conformers due to its low hydrophobic and aromatic character. The negative DLID score shows preferential binding to highly polar molecules that are not drug-like. The DYIGS site was used throughout the study for binding mode analysis as it was the most druggable site by drug-like compounds.

The binding site was evaluated using a 50 ns molecular dynamic simulation, free energy calculations, and residue decomposition analysis. The analysis reveals that hydrophobic interactions and electrostatic forces dominate binding in nicastrin. The residues Arg105, Gln139 and Val138 were found to contribute the most to the binding energy (**Table 3.3**; **Figure 3.4**). Known nicastrin compounds were docked in the DYIGS site and the interactions reveal that they interact with the residues Arg105, Gln139 and Val138 (**Figure 3.10**).

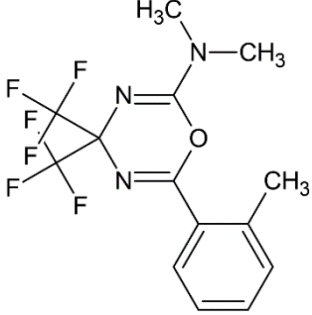
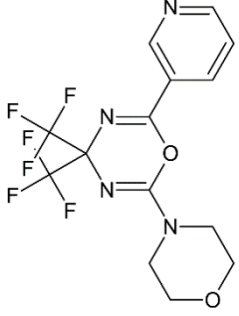
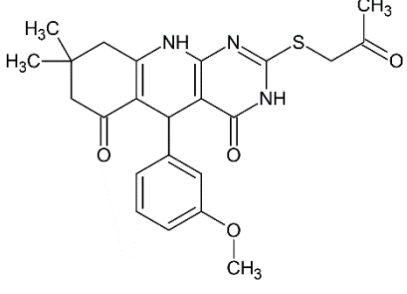
### 6.1.2 Nicastrin hits identified from virtual screening

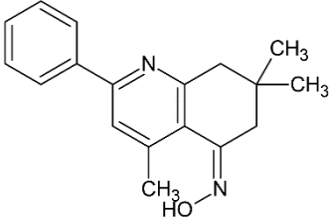
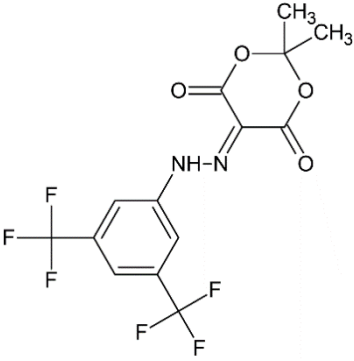
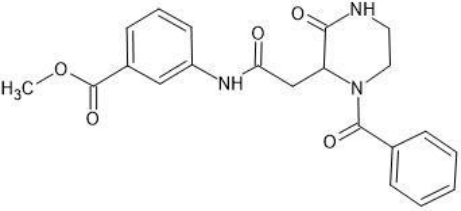
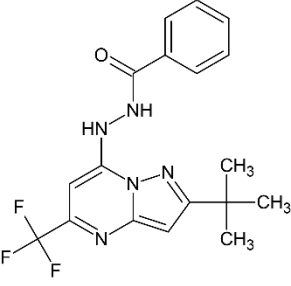
To inform the quantitative structure activity relationships that were used to identify potential inhibitors for breast cancer therapy from the Maybridge HitCreator dataset, the physicochemical properties and scaffold space of nicastrin inhibitors were investigated. High activity for nicastrin inhibition was associated with specific connectivity containing a sulfon, sulfonamide, or sulfonamide connected to a non-aromatic ring and a halide or a halide connected to a benzene ring. The degree of branching, chain length, and the presence of cyclic structures, heteroatoms like halogen, nitrogen, and sulfur atoms all had an impact on the activity of nicastrin. Seven hits were identified from the QSAR and scaffold analysis.

The seven hits were evaluated *in silico* for their physicochemical and pharmacokinetic properties as well as antitumour activity using bioassays; this evaluation provides a solid foundation for *in vivo* experiments, as summarized in **Table 6.1**. Only Compound 6218 out of the seven compounds had tumour inhibition less than 20%. This demonstrated that the models were capable of identifying nicastrin hits with anticancer activity. For the seven compounds, the percentage inhibition ranged from 30 to 87.50%. Some of the compounds did, however, exhibit an antibiotic susceptibility,

which may limit their activity. All compounds had the physicochemical characteristics required for bioavailability, according to ADMET predictions, but some of the pharmacokinetic characteristics that were flagged and shown in **Table 6.1** can be optimized for effective drug delivery.

**Table 6.1** Physicochemical, pharmacokinetic and Antitumour properties of the hits

Compound ID	Antibacterial Inhibition	% Inhibition of Tumours	Flagged ADMET properties
 <p style="text-align: center;"><b>3255</b></p>	R	45.00 ± 05.60	Long half life
 <p style="text-align: center;"><b>6218</b></p>	R	15.00 ± 10.21	Low clearance, low fraction bound, long half-life.
 <p style="text-align: center;"><b>2251</b></p>	S	57.00 ± 08.33	Pgp substrate, does not penetrate the BBB, low therapeutic index, substrate of 80% of the CYP family, low clearance and a high hepatotoxicity

 <p style="text-align: center;"><b>8796</b></p>	R	87.50 ± 05.50	Low clearance and low therapeutic index due to high plasma protein binding
 <p style="text-align: center;"><b>8361</b></p>	S	30.02 ± 10.20	Does not penetrate the blood brain barrier, low therapeutic index due to high plasma binding, low clearance and high hepatotoxicity
 <p style="text-align: center;"><b>8250</b></p>	R	79.00 ± 05.66	Pgp substrate
 <p style="text-align: center;"><b>6197</b></p>	S	52.50 ± 11.82	Does not penetrate the blood brain barrier, low therapeutic index due to high plasma binding, high hepatotoxicity.

## 6.2 Conclusion

In this work, three binding sites were identified in nicastrin. The TPR-like site, a site seen to be ideal for hydrophilic compounds, the hinge site, suitable for small hydrophobic compounds with a volume less than 250 Å and the most druggable by

drug-like compounds, the DYIGS site. Analysis of the DYIGS site revealed that hydrophobic interactions and electrostatic forces dominate binding with residues Arg105, Gln139 and Val138 contributing the most to the binding energy.

Seven compounds were identified as nicastrin inhibitors from scaffold analysis and virtual screen using machine learning models and docking and of these six having anticancer activity. Compound 6218, 2-morpholin-4-yl-6-pyridin-3-yl-4,4-bis(trifluoromethyl)-1,3,5-oxadiazine had a percentage inhibition of 15% and was off the range of anticancer compounds.

Compounds 8361(5-[[3,5-bis(trifluoromethyl)phenyl] hydrazinylidene]-2,2-dimethyl-1,3-dioxane-4,6-dione); 6197 (N'-[2-tert-butyl-5-(trifluoromethyl) pyrazolo[1,5-a]pyrimidin-7-yl] benzohydrazide) and 2251 (5-(3-methoxyphenyl)-8,8-dimethyl-2-(2-oxopropylsulfanyl)-5,7,9,10-tetrahydro-3H-pyrimido[4,5-b] quinoline-4,6-dione) showed antibacterial activity which compromised their anticancer activity.

Compounds that exhibited anticancer activity with antibacterial resistance are 3255 (N, N-dimethyl-6-(2-methylphenyl)-4,4-bis(trifluoromethyl)-1,3,5-oxadiazin-2-amine); 8796 (N-(4,7,7-trimethyl-2-phenyl-6,8-dihydroquinolin-5-ylidene) hydroxylamine) and 8250 (methyl 3-[[2-(1-benzoyl-3-oxopiperazin-2-yl)acetyl]amino]benzoate).

### 6.3 Future Work

The discovery of nicastrin inhibitors with anticancer activity has laid the foundation of development of lead compounds with proven breast cancer activity. It is important to test these hits further on breast cancer cell lines and optimize the pharmacokinetic properties. Using deep learning machine learning techniques, the compounds should be developed from hits to leads.

## APPENDIX

### A1: GROMACS mdp file for Molecular dynamic simulations of nicastrin bound to compound CID44433923

```
integrator      = md
dt             = 0.002
nsteps        = 25000000
nstlog        = 1000
nstxout       = 5000
nstvout       = 5000
nstfout       = 5000
nstcalcenergy = 100
nstenergy     = 1000;
cutoff-scheme = Verlet
nstlist       = 20
rlist         = 1.2
coulombtype   = pme
rcoulomb      = 1.2
vdwtype       = Cut-off
vdw-modifier  = Force-switch
rvdw_switch   = 1.0
rvdw          = 1.2;
tcoupl        = Nose-Hoover
tc_grps       = PROT SOL_ION
tau_t         = 1.0 1.0
ref_t         = 303.15 303.15;
pcoupl        = Parrinello-Rahman
pcoupltype    = isotropic
tau_p         = 5.0
```

compressibility = 4.5e-5  
 ref\_p = 1.0;  
 constraints = h-bonds  
 constraint\_algorithm = LINCS  
 continuation = yes;  
 nstcomm = 100  
 comm\_mode = linear  
 comm\_grps = PROT SOL\_ION;  
 refcoord\_scaling = com

**Table A 1 Absorption**

Compound	Caco2 permeability	MDCK Permeability	Human intestinal absorption	P-gp inhibitor	P-gp Substrate	F <sub>20%</sub>	F <sub>30%</sub>
2251	-4.577	2.9e-05	0.007	0.081	0.001	0.929	0.332
3255	-4.694	2.8e-05	0.005	0.815	0.003	0.004	0.008
6197	-4.961	3.1e-05	0.006	0.465	0.024	0.004	0.002
6218	-4.755	2.5e-05	0.002	0.082	0.001	0.983	0.781
8250	-4.692	0.00039	0.052	0.0	0.071	0.025	0.246
8361	-4.866	2.1e-05	0.005	0.261	0.001	0.934	0.918
8796	-4.636	3.1e-05	0.007	0.917	0.0	0.002	0.001

**Table A 2 Distribution**

Compound	PPB /%	VD L/Kg	BBB	Fu/ %
2251	94.23	0.862	0.242	2.563
3255	88.42	0.459	0.553	2.918
6197	95.7	3.304	0.888	2.738
6218	35.06	1.517	0.47	46.14
8250	87.67	1.68	0.968	6.563

8361	100.1	2.065	0.048	0.700
8796	96.76	0.69	0.95	2.44

**Table A 3 Metabolism**

Compound	CYP1A2	CYP1A2	CYP2C19	CYP2C19	CYP2C9	CYP2C9	CYP2D6	CYP2D6	CYP3A4	CYP3A4
	inhibitor	substrate	inhibitor	substrate	inhibitor	substrate	inhibitor	substrate	inhibitor	substrate
2251	0.047	0.774	0.893	0.797	0.941	0.874	0.074	0.781	0.881	0.902
3255	0.954	0.239	0.948	0.068	0.877	0.548	0.042	0.096	0.391	0.286
6197	0.947	0.925	0.946	0.127	0.934	0.518	0.389	0.197	0.2	0.278
6218	0.12	0.061	0.66	0.928	0.442	0.15	0.007	0.235	0.25	0.94
8250	0.064	0.971	0.018	0.835	0.012	0.083	0.004	0.087	0.03	0.64
8361	0.894	0.963	0.897	0.118	0.914	0.047	0.098	0.053	0.358	0.187
8796	0.766	0.162	0.915	0.086	0.901	0.871	0.043	0.645	0.144	0.323

**Table A 4 Excretion**

Compound	CL	T1/2
2251	1.325	0.35
3255	7.764	0.884
6197	7.485	0.311
6218	3.086	0.602
8250	5.454	0.119
8361	4.725	0.165
8796	1.166	0.096