



SALVADOR DA COSTA MACEDO MAGRINHO
BSc in Biochemistry

DEVELOPING A COMPUTER AIDED DRUG DESIGN APPROACH TO DISCOVERY LEAD-LIKE PMM2 ENZYME ACTIVATORS FOR CONGENITAL DISORDERS OF GLYCOSYLATION (CDG) THERAPY

MASTER IN BIOCHEMISTRY
NOVA University Lisbon
March, 2023

DEVELOPING A COMPUTER AIDED DRUG DESIGN APPROACH TO DISCOVERY LEAD-LIKE PMM2 ENZYME ACTIVATORS FOR CONGENITAL DISORDERS OF GLYCOSYLATION (CDG) THERAPY

SALVADOR DA COSTA MACEDO MAGRINHO

BSc in Biochemistry

Adviser: Florbela Pereira
Assistant Researcher, NOVA University Lisbon

Co-advisers: Paula Videira
Associate Professor, NOVA University Lisbon

Developing a computer aided drug design approach to discovery lead-like PMM2 enzyme activators for congenital disorders of glycosylation (CDG) therapy

Copyright © Salvador da Costa Macedo Magrinho, NOVA School of Science and Technology, NOVA University Lisbon.

The NOVA School of Science and Technology and the NOVA University Lisbon have the right, perpetual and without geographical boundaries, to file and publish this dissertation through printed copies reproduced on paper or on digital form, or by any other means known or that may be invented, and to disseminate through scientific repositories and admit its copying and distribution for non-commercial, educational or research purposes, as long as credit is given to the author and editor.

The outcomes of Master's thesis project have been featured in the following outlets:

- 14th ENQO & 7th ENQT - National Organic Chemistry and Medicinal Chemistry Meeting. Caparica, 20th-22nd April 2022: Poster communication entitled "Developing a computer aided drug design approach to discover lead-like pharmacological chaperones for PMM2-CDG therapy". Magrinho, S.C.M., Brasil, S., Andreotti, G., Videira, P.A. & Pereira, F.
- ProDGNE Meeting 2022. School of Science and Technology - Nova University of Lisbon, 31st May 2022: Poster communication entitled "New horizons for CDGs: Transposing current PMM2-CDG computer-aided drug design (CADD) approaches as potential routes for drug discovery in GNE myopath". Magrinho, S.C.M., Brasil, S., Andreotti, G., Videira, P.A. & Pereira, F.
- 26th Encontro Galego-Portugués de Química (Galician-Portuguese Chemistry Meeting). Facultad de Química de Santiago de Compostela, 16th-18th November 2022: Oral communication entitled "Developing a computer aided drug design approach to discover PMM2 pharmacochaperones for PMM2-CDG therapy". Magrinho, S.C.M., Brasil, S., Andreotti, G., Videira, P.A. & Pereira, F.
- 14th GLUPOR - International Meeting of the Portuguese Carbohydrate Group. School of Science and Technology - Nova University of Lisbon, 16th-18th January 2023: Poster communication entitled "Developing a computer aided drug design approach to discover lead-like PMM2 pharmacochaperones for PMM2-CDG therapy". Magrinho, S.C.M., Brasil, S., Andreotti, G., Videira, P.A. & Pereira, F.
- Review paper on drug repurposing and AI for CDG: Brasil, S., Allocca, M., Magrinho, S.C.M. *et al.* Systematic Review: Drug Repositioning for Congenital Disorders of Glycosylation (CDG). *Int. J. Mol. Sci.* (2022), 23, 8725. <https://doi.org/10.3390/ijms23158725>
- 14th ENQO & 7th ENQT abstract in conference proceedings: Pereira, F., Lourenço, A., Aires-de-Sousa, J. *et al.* 14th Edition of the Nacional Organic Chemistry Meeting and 7th Edition of the Nacional Therapeutic Chemistry Meeting. *Stand Alone Papers 2022* (2022) doi:10.3390/chemproc2022011001.

ACKNOWLEDGMENTS

À Professora Florbela Pereira que, com tanta dedicação e carinho me acolheu no seu espaço durante toda esta fase e que tão altruisticamente dedicou muito do seu tempo a ensinar-me tudo o que sei acerca de Químioinformática. Pela constante orientação, palavras de motivação, conselhos amigos e encorajamento a ir mais além, aventurando-me no mundo da investigação académica. Ficou verdadeiramente sensibilizado pela sua devoção.

À Professora Paula Videira, por tão prontamente me ter recebido no laboratório 319 e se disponibilizar para partilhar comigo o que de melhor se faz no seu grupo de investigação. Por sempre providenciar os seus conselhos e por suportar com paciência as fraquezas durante a minha aprendizagem.

A todo o grupo de Glicoinmunologia, pelos conhecimentos que me transmitiram, pelos vários ensinamentos e partilhas de experiências e pela amizade. Agradeço ainda as importantes contribuições e valioso feedback da professora Giussepina Andreotti, da doutora Mariateresa Alloca e da “tia” Diana Gallego-Martínez na elaboração do presente trabalho e no entendimento da investigação realizada.

Aos vários professores, investigadores, técnicos e colegas com que ao longo deste tempo partilhei o dia-a-dia, por proporcionarem um ambiente tão próspero para a prática da Ciência. Agradeço em especial aos meus companheiros de Licenciatura e Mestrado, com que partilhei muitas aventuras por estes anos (especialmente à Patrícia, Joana e Vanessa), a todos vós, *os quiero mucho*.

Aos paroquianos da Igreja do Monte da Caparica e em especial à Dona Carmina que ao longo dos anos, sempre que ia ao Terço da manhã antes de ir para a FCT, fazia questão de, num constante gesto de compaixão para com o próximo, incluir propositadamente nas intenções dos mistérios uma prece pelos estudantes da faculdade.

A todos os meus amigos, desde os mais antigos aos mais recentes, que, aturando-me, sempre tiveram uma palavra de apoio para me dar força e alento nesta tão importante etapa da minha formação. Que saiba sempre expressar-vos da melhor maneira a minha gratidão e o quanto gosto de todos vocês.

A toda a minha estimada família, ao Avô Manel, à Avó Zefa, ao Avô Mário e à querida Avó Picucha que de junto de Deus me acompanha. Aos meus tios e tias que tanto me acariciam e aos meus cerca de trezentos primos mais novos. A todos fico muito grato pela constante preocupação comigo e pelo amor e orgulho com que me acompanham desde sempre.

Aos meus queridos Pai e Mãe e irmãos Rosário e Manelinho, pelo incondicional apoio, amor e paciência com que tão entusiasticamente acompanham o meu percurso académico ao longo dos anos e me proporcionam uma vida confortável e feliz, encorajando-me sempre a perseguir os meus objectivos sem desistir. Que os frutos do meu trabalho e dedicação sejam também ocasião de orgulho para eles.

“E tudo quanto fizerdes, fazei-o de todo o coração...” (Col 3, 23).

ABSTRACT

The Congenital Disorders of Glycosylation (CDG) comprise a group of rare metabolic diseases caused by inherited defects in cellular glycosylation pathways, ultimately leading to a wide range of multisystem disease phenotypes. The most common sub-type, PMM2-CDG, with >1000 reported cases worldwide, is characterised by genetic defects in the coding gene for phosphomannomutase 2 (PMM2), which catalyses the interconversion of mannose 6-P to mannose 1-P, a substrate for posttranslational *N*-glycosylation. While no effective treatment is yet available, previous functional characterisation studies in patient-observed mutations have suggested the possibility of designing pharmacochaperone (PC) therapies to stabilize PMM2 structure and partially rescue its residual activity.

Using an experimentally-validated compound dataset containing over 10,000 drug-like molecules whose interaction with human PMM2 enzyme, namely thermal stability and IC₅₀ (half maximal inhibitory concentration), had been previously assayed, two machine learning strategies were developed to unveil possible PCs for hPMM2 activation: 1) A quantitative structure-activity relationship (QSAR) classification model to predict the interaction of submitted molecules with PMM2 and 2) a QSAR regression model to estimate an IC₅₀ value for such ligand-protein interactions. These QSAR models later served as computational tools to perform a virtual drug screen to search for and select hits with the desired PC profile. Compounds yielding interesting results were submitted to docking studies to further explore possible protein-ligand interactions and the complex's overall affinity.

Promising hits included quinolone derivatives, phenylpropanoic acid derivatives, and vitamins, amongst others. Some of these compounds were acquired and underwent *in vitro* experiments in patient-derived fibroblast cell lines to further analyse their effect on mutant PMM2 activity from cell extracts using different molecular biology techniques. The applications of different biomarkers for this disease model were identified and studied. This experimental work allowed to elaborate on the therapeutic potential of selected hits and provided notions on defining the models' predictive capabilities for such complex structure-biological activity relationships.

Keywords: *N*-glycosylation, PMM2-CDG, CDG therapies, Computer-Aided Drug Design, Drug repurposing, Pharmacological chaperoning, Virtual drug screen

RESUMO

Os Distúrbios Congénitos da Glicosilação (CDG) representam um grupo de doenças metabólicas raras causadas por defeitos hereditários nos mecanismos de glicosilação, dando origem a uma gama ampla de fenótipos, com um impacto multissistémico. O subtipo mais comum, PMM2-CDG, com mais de 1000 casos relatados em todo o mundo, é caracterizado por defeitos no gene codificante da fosfomanomutase 2 (PMM2), que catalisa a inter-conversão de manose 6-P em manose 1-P, um substrato para *N*-glicosilação pós-traducional. Embora não exista nenhuma terapêutica eficaz, estudos de caracterização funcional em mutações observadas em pacientes sugeriram a possibilidade de projectar terapias de farmacochaperonas (PC) para estabilizar a estrutura da PMM2 e resgatar parcialmente a sua actividade.

Usando um conjunto de dados com mais de 10000 compostos validados experimentalmente cuja interacção com a PMM2 humana, nomeadamente estabilidade térmica e IC_{50} (metade da concentração inibitória máxima), foram previamente testadas, duas estratégias de aprendizagem automática foram desenvolvidas para desvendar possíveis PCs para activação da hPMM2: 1) Um modelo classificativo de relações quantitativas de estrutura-actividade (QSAR) para prever a interacção das moléculas submetidas para com a PMM2 e 2) um modelo de regressão QSAR para estimar um valor de IC_{50} para as tais interacções. Esses modelos QSAR serviram posteriormente como ferramentas computacionais para realizar uma triagem virtual de fármacos com o intuito de pesquisar e seleccionar *hits* com o perfil desejado. Os compostos com resultados interessantes foram submetidos a estudos de *docking* molecular para explorar mais aprofundadamente as interacções proteína-ligando e a afinidade geral do complexo.

Compostos promissores incluem derivados de quinolonas, de ácido fenilpropanóico, vitaminas, entre outros. Quatro destes foram adquiridos e submetidos a experimentação *in vitro* em linhas de fibroblastos derivadas de pacientes para explorar o seu efeito na actividade da PMM2 a partir de extractos celulares usando diferentes técnicas de biologia molecular. As aplicações de diferentes biomarcadores para este modelo de doença foram também identificadas e estudadas. Este trabalho experimental permitiu aprofundar o potencial terapêutico de *hits* seleccionados e definir as capacidades de previsão dos modelos QSAR para relações tão complexas entre estrutura química e actividade biológica.

Palavras-chave: *N*-glicosilação, PMM2-CDG, Terapias para CDG, Desenho computacional de fármacos, Reaproveitamento de fármacos, Chaperones farmacológicos, Triagem virtual de fármacos

CONTENTS

1	INTRODUCTION.....	1
1.1	Glycosylation in Health and Disease, a brief overview.....	1
1.1.1	Glycosylation, an essential process for life	1
1.1.2	<i>N</i> -linked glycosylation	2
1.1.3	CDGs: When glycosylation goes wrong	4
1.2	PMM2-CDG, a phosphomannomutase deficiency.....	6
1.2.1	Background	6
1.2.2	Structural and functional characterisation of PMM2	9
1.2.3	Impact of mutations in PMM2 enzyme function.....	12
1.2.4	Therapies for PMM2-CDG.....	15
1.3	Cheminformatics: A Computer-Aided Drug Design (CADD) approach to PMM2-CDG therapies	19
1.3.1	Chemical Structure Representations & Molecular Descriptors.....	20
1.3.2	Structure-Property and Structure-Activity Relationships.....	23
1.3.3	Machine Learning in CADD approaches	25
1.3.4	Molecular Docking.....	30
1.4	Biological studies applied to PMM2-CDG drug development	33
1.4.1	Fibroblasts as disease study models for PMM2-CDG.....	33
1.4.2	Molecular biomarkers to study PMM2-CDG <i>in vitro</i> models.....	34
1.4.3	Enzymatic assays to measure PMM2 activity	36
1.5	Goals for this thesis	37

2	MATERIALS & METHODS	39
2.1	Cheminformatics component.....	39
2.1.1	Model 1: QSAR classification model to predict ligand interactions with PMM2.....	39
2.1.2	Model 2: QSAR regression model to predict theoretical IC ₅₀ value	44
2.1.3	Virtual screening of approved drugs	47
2.1.4	Molecular Docking studies.....	48
2.2	Biological studies	50
2.2.1	Fibroblast culturing	50
2.2.2	Preparation of test compounds for the biological studies.....	51
2.2.3	Resazurin assays.....	51
2.2.4	Assays to evaluate PMM2 enzyme activity in fibroblast extracts.....	53
2.2.5	Western blotting studies with fibroblasts extracts.....	54
3	RESULTS & DISCUSSION	57
3.1	Cheminformatics component.....	57
3.1.1	Building a compound dataset for QSAR model construction	57
3.1.2	QSAR classification Model 1 for interaction prediction	59
3.1.3	QSAR regression Model 2 for IC ₅₀ estimation.....	77
3.1.4	Virtual screening of approved drugs	91
3.1.5	Molecular Docking studies.....	92
3.2	Biological component.....	103
3.2.1	Cell viability assays.....	105
3.2.2	Evaluation of PMM2 enzyme activity in fibroblast extracts.....	108
3.2.3	Western blotting with fibroblast extracts	110
4	CONCLUSION & FUTURE WORKS.....	113

LIST OF FIGURES

Figure 1. General process of <i>N</i> -linked glycosylation.	4
Figure 2. Overview map of the congenital disorders of glycosylation (CDG).....	5
Figure 3. Scheme of surrounding pathways involving PMM2 in human metabolism.	7
Figure 4. Comparison by primary sequence alignment of hPMM2 with isozyme hPMM1 and related PMM counterparts from model organisms <i>M. musculus</i> , <i>S. cerevisiae</i> and <i>L. mexicana</i>	9
Figure 5. Structure representation of human phosphomannomutase 2 and primary sequence.....	10
Figure 6. Simplified scheme of the reactional mechanism for catalysis in hPMM2.	12
Figure 7. PMM2-CDG distribution of amino acid substitutions found in PMM2-CDG patient pathogenic variants..	13
Figure 8. Repurposing candidates for PMM2-CDG.....	16
Figure 9. Simplified illustration of a pharmaco-chaperoning approach for PMM2-CDG therapy.	18
Figure 10. Schematic illustration of a Random Forest algorithm setup	27
Figure 11. Illustrative example of a Support Vector Machine..	29
Figure 12. Neuron Network schematic in a supervised learning setting.	30
Figure 13. Typical docking workflow.	31
Figure 14. Schematic depiction of different ICAM-1 isoforms.	34
Figure 15. The typical western blotting workflow.	36
Figure 16. PMM2 phosphomannomutase kinetic assay pathway.....	37
Figure 17. General workflow of the CADD approach	38
Figure 18. Subway map representing the general workflow employed to build the predictive QSAR models.	40
Figure 19. Simplified scheme of the workflow employed for the virtual drug screening approach.	48
Figure 20. Representation of hit compounds described in the work of Yuste-Checa and collaborators.	58
Figure 21. Benzisothiazolinone 2,3-dihydro-1,2-benzothiazol-3-one scaffold.....	59
Figure 22. Figure X. Variation of model performance evaluation metrics with different class weights.	64
Figure 23. Bar chart for Matthews's Correlation Coefficient values for trial models using different FP/descriptor types.	66

Figure 24. Bar chart for Matthews's Correlation Coefficient values for the try-out models..	70
Figure 25. Bar charts with comparisons between Sensitivity (SE), Specificity (SP) and Matthews's Correlation Coefficient values.....	74
Figure 26. Template compounds used for calculation of the model's applicability domain.....	76
Figure 27. Bar chart representing the Correlation Coefficient for trial models built with the addition of different 3D descriptors.....	81
Figure 28. Bar chart for the variation in the Correlation Coefficient values for trial models built with the different defined FP/descriptor bundles..	84
Figure 29. Scatter plot of the final RF model trained with the EState FP + all 3D set bundle.....	87
Figure 30. Scatter plot of the final SVM model trained with the 1D&2D descriptor + CPSA bundle.	88
Figure 31. Scatter plot of the final NN model trained with the 1D&2D descriptor + CPSA bundle. ..	89
Figure 32. Testing set instances.	89
Figure 33. Ipsapirone (1,1-dioxo-2-[4-(4-pyrimidin-2-yl)piperazin-1-yl]butyl]-1,2-benzothiazol-3-one) and Elseben (2-phenyl-1,2-benzoselenazol-3-one).	92
Figure 34. A) Quinolone, B) phenylpropanoic acid and C) camptothecin scaffolds.	96
Figure 35. Structural representation of docked ligand 1-(3-chlorophenyl)-3,3-bis(pyridine-2-yl)urea in hPMM2.....	100
Figure 36. Schematic diagram representing the protein-ligand interaction established between hPMM2 and docked ligand 1-(3-chlorophenyl)-3,3-bis(pyridine-2-yl)urea.	101
Figure 37. Relative occurrence of PMM2 residues in the protein-ligand interactions described for all docked ligands.....	102
Figure 38. Compounds acquired for experimental validation with biological studies.	104
Figure 39. The underlying basis of methodology used in resazurin assays.....	105
Figure 40. Scatter plot of resazurin metabolism for control and PMM2-CDG fibroblast lines.	106
Figure 41. Bar chart of cell viability for control and PMM2-CDG fibroblast lines exposed to different concentrations of DMSO.....	107
Figure 42. Bar charts of cell viability for control and PMM2-CDG fibroblast lines exposed to different concentrations of the four lead drugs.	108
Figure 43. Enzyme activity curves as a function of the reactional time.	109
Figure 44. hPMM2 and ICAM-1 western blotting analysis..	110

LIST OF TABLES

Table 1. Brief descriptive summary of some of the molecular Fingerprints and descriptors available for calculation in PaDEL-Descriptor that were selected to be utilised in this study.....	60
Table 2. Preliminary analysis of trial RF QSAR models to evaluate the predictive performance of each FP/descriptor group.	66
Table 3. Top ten MACCS FP attributes measured by Mean Decrease Accuracy.	67
Table 4. Top ten 1D&2D attributes measured by Mean Decrease Accuracy.....	68
Table 5. Analysis of try-out RF QSAR models to evaluate the predictive performance of different sets of descriptor attributes.....	69
Table 6. Analysis of the final models trained for different ML algorithms.	72
Table 7. Evaluation with test set of the final models trained with different ML algorithms.....	73
Table 8. Tanimoto Coefficient threshold values calculated for the QSAR model's dataset compounds defined for each one of the positive and negative template molecules.	77
Table 9. Brief descriptive summary of some of the molecular three-dimensional descriptors available for calculation in PaDEL-Descriptor that were selected to be utilised in this study.	78
Table 10. Analysis of preliminary RF QSAR models to evaluate the predictive performance of each FP and 3D descriptor type.	80
Table 11. Performance evaluation for RF QSAR models built with different FP/descriptor bundles.	82
Table 12. Analysis of trial models to assess how performance is affected by attribute no.	84
Table 13. Training set analysis of the Random Forest models.	86
Table 14. Training set analysis of the Support Vector Machine models.....	87
Table 15. Training set analysis of the Neural Network models.	88
Table 16. Final test set evaluation of the QSAR regression models trained with different ML algorithms.	90
Table 17. Docking scores obtained for the template compound set.....	95
Table 18. Scaffold grouping of potential hits compounds identified with molecular docking and respective docking scores.....	97
Table 19. Chemaxom Standardizer SMILES standardization list with defined action commands displayed in hierarchical order.	129
Table 20. Recipes for various solutions used throughout the biological experiments.	130

LIST OF ABBREVIATIONS AND SYMBOLS

AI	Artificial Intelligence
CADD	Computer-Aided Drug Design
CDG	Congenital Disorder of Glycosylation
DMEM	Dulbecco's Modified Eagle Medium
DR	Drug reprofiling (or repositioning or repurposing)
drug R&D	Drug Research and Development
FN	False Negative
FP	False Positive
FPS	Molecular Fingerprints
HAD family	Haloalkanoic Acid Dehydrogenase-like superfamily
IC ₅₀	Half maximal in
ICAM-1	Inter Cellular Adhesion Molecule 1, CD54
MAE	Mean Absolute Error
MCC	Matthews's Correlation Coefficient
ML	Machine Learning
NN	Neural Network
PC	Pharmacological chaperone
PMM1 or hPMM1	Phosphomannomutase isozyme 1 (the human form unless stated otherwise)
PMM2 or hPMM2	Phosphomannomutase isozyme 2 (the human form unless stated otherwise)
PMM2-CDG	Phosphomannomutase 2 congenital disorder of glycosylation (CDG type Ia)
Q	Overall Predictability
QSAR	Quantitative Structure-Activity Relationships
R	Pearson's Correlation Coefficient
RF	Random Forest
RMSE	Root Mean Square Error
SE	Sensitivity metric
SL	Supervised Learning
SP	Specificity metric
SVM	Support Vector Machine
TC	Tanimoto Coefficient

T_m	Melting temperature
TN	True Negative
TP	True Positive
WB	Western Blotting

INTRODUCTION

This introduction summarises a comprehensive literature review on the various essential concepts approached during this study in order to better understand the biological background of glycosylation and glycobiology in health and disease; the underlying aspects of PMM2-CDG, its pathogenicity and therapeutic prospects; the computational chemistry techniques that were exploited, as well as the biological experimentation that followed.

1.1 Glycosylation in Health and Disease, a brief overview

1.1.1 Glycosylation, an essential process for life

Sugars (or glycans) are widely distributed in nature and comprise an essential component of living systems. Besides having an important biological function on their own, saccharides can be combined to create a range of different glycan structures that can be attached to a non-sugar moiety forming a glycoconjugate^{1,2}. Glycosylation is defined as the enzyme-mediated process through which a carbohydrate is covalently attached to a target macromolecule, such as a protein or lipid, in the endoplasmic reticulum (ER) and Golgi apparatus^{3,4}.

Concerning proteins, glycosylation is regarded as the most frequent and diverse form of post-translational modification (PTMs), being reportedly estimated that around 50% of human genes code for proteins with the potential to be glycosylated^{5,6}. Furthermore, it is described that at least 2% of the human genome codes for proteins involved in glycosylation⁷, which explains how this dynamic process is so tightly controlled, requiring several intervening enzymes that synthesise these glycan structures depending on several factors like glycosyl precursors, gene expression, substrate accessibility, cell type and cellular environment, amongst others⁸. Glycoconjugates can be generally found throughout the cell surface, inside the cell or secreted to the extracellular matrix⁹. Therefore, glycosylation not only influences protein maturation, structural stability and turnover, as it also modulates a variety of physiological and pathological processes, playing a vital role in, for example, cell signalling and adhesion, or receptor binding and activation^{2,10,11}. Proteins can undergo two main types of glycosylation: *N*-linked glycosylation, which produces a glycoconjugate with a sugar chain bound to the amine group of an asparagine

amino acid at the consensus sequence -Asn-X-Ser/Thr-, and *O*-linked glycosylation, where the sugar is linked to the hydroxyl oxygen of Ser or Thr side chains². In this thesis only *N*-glycosylation will be approached.

1.1.2 *N*-linked glycosylation

N-glycosylation in humans is a highly conserved co-translational process that involves a complex enzymatic pathway, as shown in **Figure 1**: In a first phase, a dolichol phosphate carrier embedded in the ER membrane facing cytosol binds to a uridine diphosphate *N*-acetylglucosamine (UDP-GlcNAc) forming a pyrophosphate linkage. This is followed by the attachment of yet another UDP-GlcNAc and subsequently five guanosine diphosphate mannose (GDP-Man) residues bind to the disaccharide. Then, the dol-PP-GlcNAc₂-Man₅ precursor is translocated across the membrane, now facing the ER lumen, where additional glycosyltransferases will further add 4 mannose and 3 glucose (Glc) units, using other dol-PP-Man and dol-PP-Glc carriers as glycan donors. The final dol-PP-GlcNAc₂Man₉Glc₃ precursor transfers its oligosaccharide chain *en bloc* to the Asn residue in the protein, forming an *N*-glycosidic bond in a reaction catalysed by an oligosaccharyltransferase complex which recognises the -N-X-S/T- consensus sequence at the nascent polypeptide chain.

Afterwards, the *N*-glycan is submitted to a series of processing steps, starting still in the ER, where an initial trimming stage is believed to be a crucial part in regulating glycoprotein folding: α -glucosidases I and II remove the terminal α 1-2Glc and the two inner α 1-3Glc units, respectively. If by any reason the protein did not fold correctly after translation, the *N*-glycan's third α 1-3Glc residue will not be removed, preventing the protein from exiting the organelle. In this situation, the ER's quality control machinery comes into action and lectin-like molecular chaperones calnexin and calreticulin are recruited to assist in correcting the protein's folding. Once proper folding is achieved, the Glc unit is cleaved followed by removal of the terminal α 1-2Man from the glycan's central arm catalysed by ER mannosidase I, ultimately signalling the glycoprotein to be exported to the Golgi apparatus. This high-mannose glycan structure is one of the main types of *N*-glycans.

At the *cis*-Golgi, further processing takes place and several Man residues are sequentially trimmed by α 1-2 mannosidases I, before proceeding to the *medial*-Golgi. In the *medial*-Golgi, high-mannose glycans can be converted into the other two main forms of *N*-glycans: hybrid and complex *N*-glycans. The biosynthetic pathways for hybrid and complex *N*-glycans start with the addition of a GlcNAc residue by *N*-acetylglucosaminyltransferase GlcNAc-TI to the terminal α 1-3Man of the sugar tree, forming the precursor for hybrid *N*-glycan synthesis. To produce complex *N*-glycans, further trimming by α -mannosidase II removes a few more terminal Man residues and a second GlcNAc is added by GlcNAc-TII, originating the building block for bi-branched, complex *N*-glycans. More branches can be later added to obtain tri- and tetra-antennary *N*-glycans. The core of complex *N*-glycans is modified by the addition of a fucose (Fuc) unit to the asparagine-bound GlcNAc by α 1-6 fucosyltransferase (FUT8) before continuing to the *trans*-Golgi. Once in the *trans*-Golgi, *N*-glycans mature by incorporating additional sugar units, resulting in elongated branched glycans³.

As the glycans move through this secretory pathway, the sequential addition of sugars to the growing glycan tree is determined by the acting enzymes, substrate specificity, competition and availability, and even differences in gene expression and organization patterns amongst cells and tissues². All these factors combined make possible to create, from a single glycan core and a limited selection of sugars, a rich diversity of *N*-glycans varying in composition, length and branching, resulting in more than 7000 *N*-glycoproteins in humans¹¹.

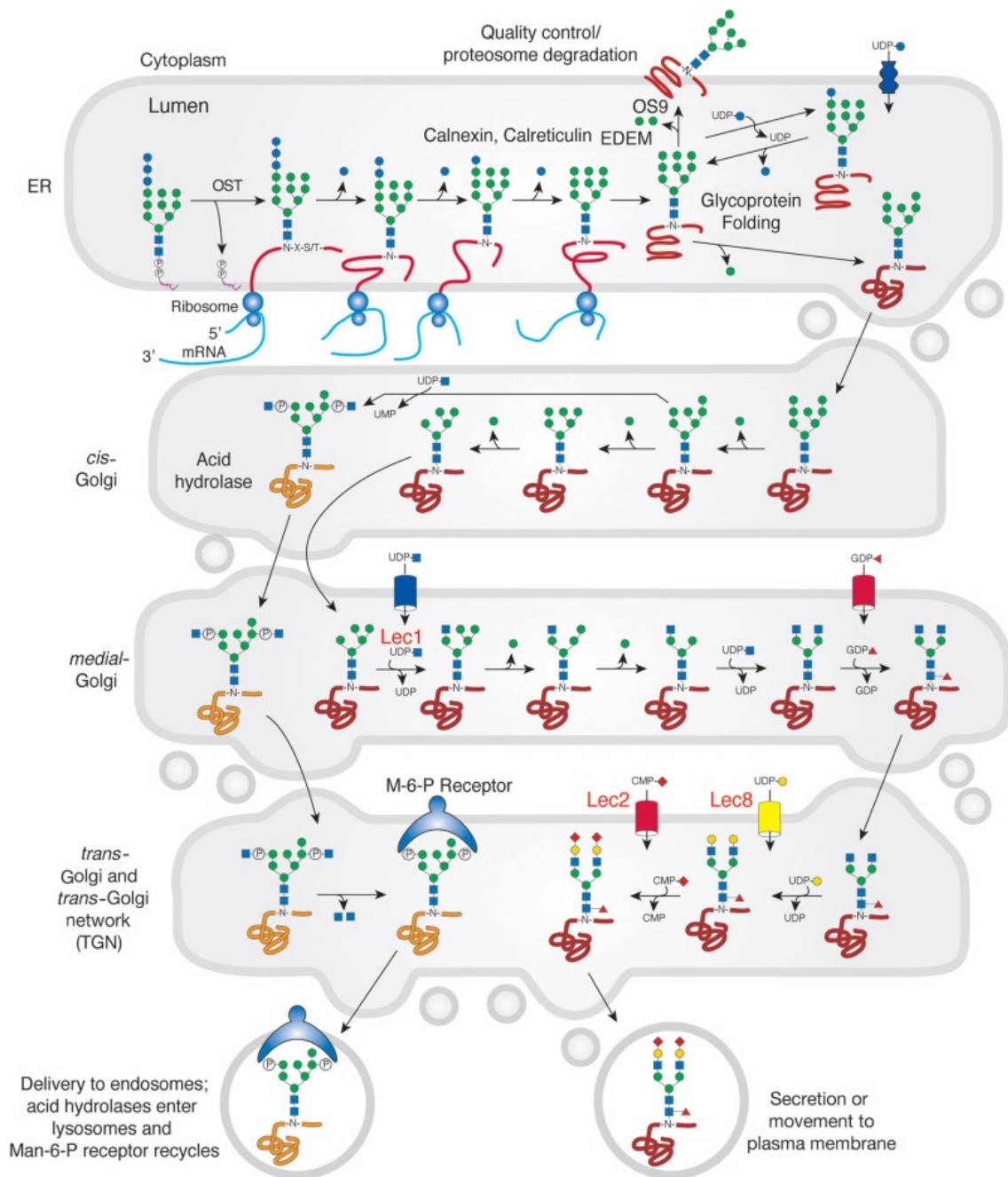


Figure 1. General process of N-linked glycosylation. The glycan precursor is assembled on a dolichol phosphate base whilst in the cytoplasm of the ER. Once in the lumen side, additional sugars are added and trimmed before being attached to the protein, with additional processing later taking place. GDP-Man and UDP-GlcNAc as glycan donors. Adapted from³.

1.1.3 CDGs: When glycosylation goes wrong

The importance of functional glycosylation becomes evident when an individual carries mutations in a gene coding for one of the many enzyme effectors of these pathways, resulting in defective biosynthesis

of glycan precursors and abnormal glycosylation. These defects originate the potentially life-threatening congenital disorders of glycosylation (CDGs)^{6,12,13}. CDGs are an ever-growing family of inherited autosomal recessive disorders (a reduced number is inherited in an autosomal dominant or X-linked fashion) and, whilst being considered rare, more than 160 types have been identified to date, although more are expected to exist given the number of glycosylation-related proteins encoded in our genome. As each affected gene causes a different disease type, CDGs can be generally grouped into four categories, according to their defective pathway, as shown below in **Figure 2**: disorders of protein *N*-glycosylation (which is the largest class); protein *O*-glycosylation; lipid glycosylation and other glycosylation pathways & multiple glycosylation pathways^{13–15}.

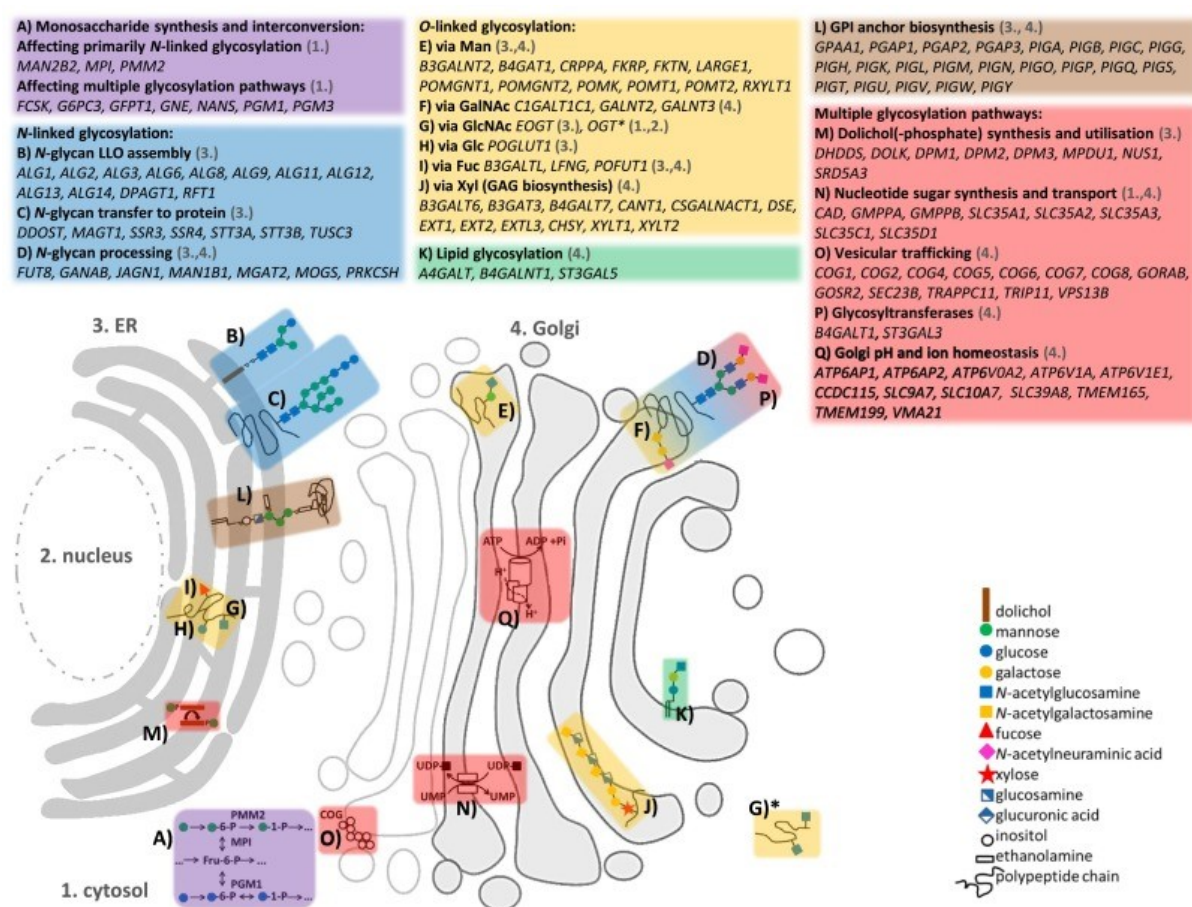


Figure 2. Overview map of the congenital disorders of glycosylation (CDG). This schematic illustration depicts the first individual 137 subtypes of CDG reported, indicating the affect gene and respective glycosylation pathway. Adapted from¹⁵.

Given that glycoconjugates have important roles in a range of physiological functions across tissues and organs, CDGs produce a great variety of clinical phenotypes with multi-system manifestations, being especially note-worthy the impacts on brain development and nervous system functions, as ~80% of

CDGs have a major neurological component. Other severely affected organs observed in patients include the hepatic, gastrointestinal, skeletal, circulatory and immune systems, amongst others^{12,15–17}.

Obtaining a diagnosis for CDG can be a challenging endeavour due to the lack of a well-defined clinical presentation, consequence of the low number of identified patients, scarcity of data and the very heterogeneous pathophysiological patterns verified between patients as a result of its multi-systemic nature¹³. A typical diagnosis strategy comprehends different molecular testing procedures, such as isoelectric focusing of transferrin, a glycoprotein abundant in the serum, to identify defects in *N*-glycosylation. This is the most broadly used screening test and transferrin is considered a primary biomarker for CDG. However, it can only be applied to CDGs involving *N*-glycans with sialic acid deficiencies and the existence of CDGs with normal transferrin patterns emphasise the necessity for more biomarkers¹⁶. Different mass spectrometry (MS) techniques have been suggested as supplementary analysis to profile the patient's *N*-glycome, being progressively applied in diagnosis, mainly when transferrin analysis is inconclusive^{18,19}. CDGs related to *O*-glycosylation abnormalities can be detected also through isoelectric focusing for serum apolipoprotein C-III. Moreover, recent works provide new insights for the on-going quest for biomarkers for CDG diagnosis and therapy²⁰. Despite these methods, it is currently recommended for the final diagnosis to contemplate whole genome or exome sequencing (WGS or WES) analysis, which unequivocally confirms the presence of genetic variants in glycosylation-related genes.

Treatment-wise, the options are limited to very few CDG types and, even then, since an effective curative therapeutic is yet to be established, the only treatments available are mainly symptomatic and supportive therapies¹⁶. Besides these, other current directed treatment efforts include nutritional supplementation; enzyme therapies; organ transplants and gene therapy, amongst others²¹. This project was designed with the scope of contributing for the recent efforts in developing innovative approaches to drug design & discovery for PMM2-CDG, the most common CDG, with nearly a thousand patients reported worldwide²² and an estimated incidence of 1/20000 people²³. This disorder, alongside its genetic causes, pathophysiology and current therapeutic strategies, as well as the entire thesis outline and concept, will be discussed in depth in the following chapters.

1.2 PMM2-CDG, a phosphomannomutase deficiency

1.2.1 Background

Human phosphomannomutase 2 (hPMM2, EC 5.4.2.8) is a 246-amino acid enzyme, with a little over 28 kDa, a pI = 6.35 and a molar extinction coefficient $\epsilon = 20775 \text{ M}^{-1} \text{ cm}^{-1}$ (computed in ProtParam²⁴). This cytosolic protein catalyses the interconversion between mannose 6-phosphate (Man 6-P) and its isomer, mannose 1-phosphate (Man 1-P) (*vide* **Figure 3**), in an essential step preceding the production of the GDP-Man precursors required for *N*-glycosylation in humans, as described in **Section 1.1.2**.

PMM2 requires a bisphosphate sugar co-factor to be activated, either mannose or glucose 1,6-bisphosphate (Man 1,6-BP or Glu 1,6-BP)²⁵ and it is also reported to have phosphoglucomutase activity, catalysing the interconversion of glucose 6-phosphate (Glu 6-P) to glucose 1-phosphate (Glu 1-P)^{26,27}. Due to its vital role, PMM2 is an essential enzyme and complete lack of activity is incompatible with life, as demonstrated by studies with knockout mice^{28,29} and hypothesised by the absence of patients bearing known null mutations in a homozygous pattern (*vide* Section 1.2.3)³⁰.

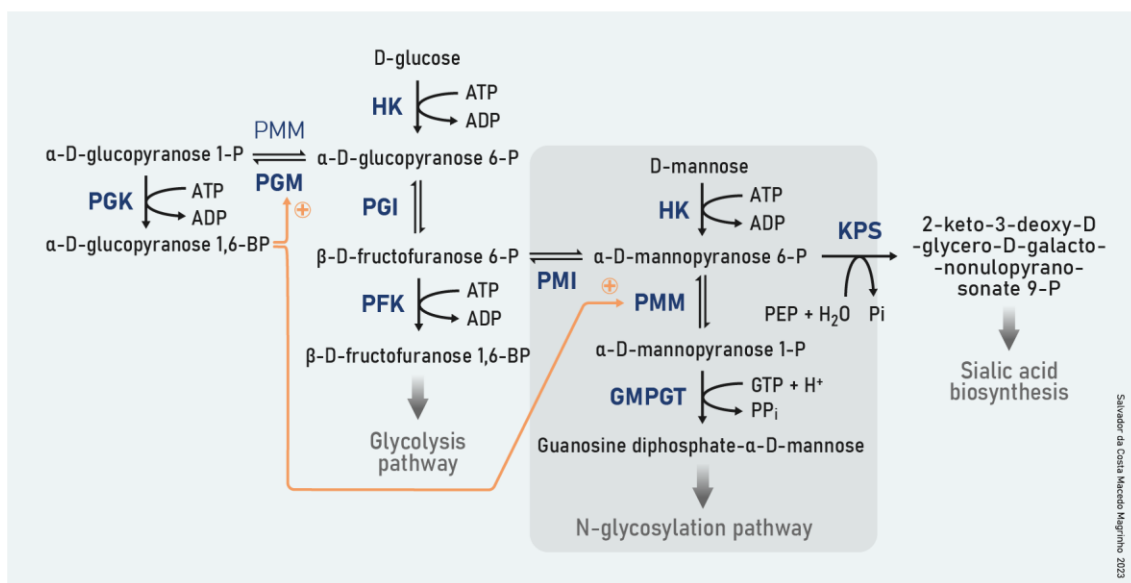


Figure 3. Scheme of surrounding pathways involving PMM2 in human metabolism.

hPMM2 is encoded in the *PMM2* gene (OMIM entry: 601785, GenBank: U85773.1) which comprises eight exons in an extension of around 20 kbp in chromosome 16p13^{31,32}. The gene is proposed to be regulated by a housekeeping promoter³³, which seems plausible given the essential role of PMM2 in cellular glycosylation. Despite its ubiquity in the human body, variability in the constitutive expression of PMM2 across tissues has been documented before²⁶, suggesting the possibility of an additional tissue-specific regulation to the gene. The highest expression is described to occur in the pancreas and liver, two organs with important production of secreted glycoproteins³⁴.

Another isoform of PMM2, phosphomannomutase isozyme 1 (hPMM1), has also been identified in humans. PMM1 contains 262 amino acids and is encoded by the paralogous *PMM1* gene (OMIM entry: 601786, GenBank: NM_002676.3), located in chromosome 22q13. It is believed that both *PMM1* and *PMM2* genes evolved from a duplication event approx. 75-110 million years ago³⁵. Whilst both enzymes require the same sugar activator and possess the same catalytic activities, PMM1 registers slightly distinct substrate specificity and enzyme kinetics, as well as an additional phosphatase activity, hydrolysing Glc-1,6-BP and Man1,6-BP in the presence of inosine monophosphate (IMP)^{29,36}. Regardless, PMM1

activity does not compensate for the loss of PMM2 in hypomorphic variants, as clearly seen in PMM2-CDG patients and other organism models³⁷. hPMM 1 and 2 share great sequence identity and similarity, 68% and 82% respectively (*vide* **Figure 4**) and, even though that most of the sites described to be crucial for protein structure and catalytic activity remain conserved in both isoforms, variations at the primary sequence level between amino-acid clusters are described and may be the fundament for the observed functional divergences³⁸. Disparities between both isozymes are also noted in tissue expression where, contrarily to the ubiquitous PMM2, PMM1 is reportedly downregulated in human adults, being only found in brain, lungs, liver, pancreas and endocrinal glands. Perhaps the most striking contrast is that, concerning pathogenetics, unlike *PMM2* whose disease-associated mutations are well known and comprise a key aspect for the study of PMM2-CDG (this topic is further developed in **Section 1.2.3**), no mutations in *PMM1* have been linked to human diseases till today²⁷. In fact, the physiological role of PMM1 is not yet defined: It is thought to be involved in neural and endocrinal development, rather than protein glycosylation like PMM2, although studies in *PMM1* knock-out mice models showed that normal development was not affected. This may be a result of enzyme redundancy (a recurrent consequence of gene duplication) with constitutive PMM2 compensating PMM1 deficiency^{37,39}. To date, the effects of PMM1 mutations on a PMM2-null background have not yet been studied, since producing a PMM2 knock-out mouse model has been proven difficult due to early embryonic mortality²⁸.

Apart from human PMM1 and PMM2, related PMM enzymes are essential and highly conserved throughout eukaryotes⁴⁰, including in well-studied model organisms. Below, **Figure 4** presents the primary sequences of a few selected PMMs, such as *L. mexicana*'s and yeast SEC53, showcasing their similarity and conserved clusters. This observation attests, once again, the underlying importance of glycosylation in living systems.

PMM2	<i>H. sapiens</i>	-----MAAPGPALCLFDVDGTLTPARQKITKEMDDFLQKLR-QKIKIGVVGGSDFEKV	52
PMM1	<i>H. sapiens</i>	MAVTAQAARRKERVLCFLFDVDGTLTPARQKIDPEVA AFLQKLR-SRVQIGVVGGSDYCKI	59
PMM2	<i>M. musculus</i>	-----MATLCLFMDGTLTPARQKITEEMDGF LQKLR-QKTKIGVVGGSDFEKL	48
SEC53	<i>S. cerevisiae</i>	MSIAEFAYKEKPEITLVLFVDVDGTLTPARLTVSEEVKRTLAKLR-NKCCIGFVGGSDLSKQ	59
PMM	<i>L. mexicana</i>	-----MGSKAILLFDVDGTLTPERNPETHDMKEALLKARAAGFKLG VVGGSDFAKQ	51
		: : * * * : * * * * : : * * * : * : * * * * *	
PMM2	<i>H. sapiens</i>	QEQLG--NDVVEKYDYVFPENGLVAYKDGKLLCRQNIQSHLGEALIQDLINYLCLSYIAKI	110
PMM1	<i>H. sapiens</i>	AEQLGDGDEVIEKFDYVFAENGTVQYKHGRLLSKQTIQNHLGEELLQDLINFLCLSYMALL	119
PMM2	<i>M. musculus</i>	QEQLG--NDVVEKYDYVFPENGLVAYKDGKLLCKQNIQGHLEGDVQLINYLCLSYIANI	106
SEC53	<i>S. cerevisiae</i>	LEQLG--PNVLDEFDYSFSENGLTAYRLGKELASQSFINWLGEKYNKLAVFILRYLSEI	117
PMM	<i>L. mexicana</i>	KEQLG--ESILEDFDYVSENGLLAYKDGKEFHRNSLLRALGNEKVVA FVKCLHLIADL	109
		* * * * : : : : * * * * * : * : : : : : * * : : *	
PMM2	<i>H. sapiens</i>	KLEKKRGTFIEFRNGMLNVSPIGRSCSQEERIEFYELDKKENIRQKFVADLRKEFAGKGL	170
PMM1	<i>H. sapiens</i>	RLEKKRGTFIEFRNGMLNISPIGRSCTLEERIEFSELDKKEKIREKFVEALKTEFAGKGL	179
PMM2	<i>M. musculus</i>	KLEKKRGTFIEFRNGMLNVSPIGRSCSQEERIEFYELDKKEHIRQKFVADLRKEFAGKGL	166
SEC53	<i>S. cerevisiae</i>	DLEKKRGTFLEFRNGMINVSPIGRNASTEERNEFERDYDKEHQIRAKFVEALKKEFPDYGL	177
PMM	<i>L. mexicana</i>	DLEVQRGTFVEFRNGMFNVSPIGRNCSSQERDEFENLDKERHIREKLIRELKEAFDPDYQL	169
		: * : * * * : * * * * : * : * * * : * : * * : * : *	
PMM2	<i>H. sapiens</i>	TFSIGGQISFDVFPDGDWKRYCLRHVENDGYKTIYFFGDKTMPGGNDHEIFTDPRTMGYS	230
PMM1	<i>H. sapiens</i>	RFSRGGMISFDVFPDGDWKRYCLDSLQDSFDTIHFFGNETSPGGNDHEIFTDPRTVGHS	239
PMM2	<i>M. musculus</i>	TFSIGGQISIDVFPDGDWKRYCLRHLEHAGYKTIYFFGDKTMPGGNDHEIFTDPRTVGYT	226
SEC53	<i>S. cerevisiae</i>	TFSIGGQISFDVFPAGWDKTYCLQHVEKDGFKELHFFGDKTMVGGNDYEIFTDSRTIGH	237
PMM	<i>L. mexicana</i>	AYSVGGQISFDVFPKGWDKTYCLQFVEN-DFETIHFFGDKTSEGGNDYEIFTDSRTIGH	228
		: * * * * : * * * * * * * * : : : : * : * * * : * * * : * * * : * : :	
PMM2	<i>H. sapiens</i>	VTAPEDTRRICELLS-----	246
PMM1	<i>H. sapiens</i>	VVSPQDTVQRCREIFFPETAHEA	262
PMM2	<i>M. musculus</i>	VTAPEDTRRICELGLP-----	242
SEC53	<i>S. cerevisiae</i>	VQSPDDTVKILTELFNL-----	254
PMM	<i>L. mexicana</i>	VKTYKDTIAILEALLEDSR----	247
		* : . * * : : :	

Figure 4. Comparison by primary sequence alignment of hPMM2 with isozyme hPMM1 and related PMM counterparts from model organisms *M. musculus*, *S. cerevisiae* and *L. mexicana*. hPMM2 primary sequence is depicted in bold. Residues marked with (*) and highlighted in green are conserved across PMMs. Residues marked with (:) are only identical between at least three PMM queries and similar to the remainder, whilst for residues marked with (.) at least two are identical and similar to the remainder. Calculated sequence identity with hPMM2 is 66.67%, 90.50%, 58.13% and 55.56% and calculated sequence similarity with hPMM2 is 82%, 96%, 71% and 72% for hPMM1, murine PMM2, SEC53 and leishmanial PMM, respectively. Alignment was performed using UniProt Consortium's Basic Local Alignment Search Tool (Align)⁴¹ available on-line at <https://www.uniprot.org/align>.

1.2.2 Structural and functional characterisation of PMM2

PMM2 belongs to the haloalkanoic acid dehydrogenase-like (HAD-like) superfamily, which shares a common topology of a conserved α/β -domain, resembling a Rossmann fold, and a smaller cap domain of higher sequence heterogeneity, correlated to the different functions and substrate specificities within the family⁴²⁻⁴⁴. HAD-like proteins do not share great sequence similarity, although three adjacent sequence motifs seem to be highly conserved, including in hPMM2 (*vide* **Figure 5A**): Motif I, -DX(D/T/Y)X(T/V)(L/V/I)-, which has been identified as a transient phosphorylation site; Motif II, comprising a serine or threonine placed near the active site, and Motif III, containing a conserved lysine that stabilizes an intermediate phosphorylated state^{45,46}. Furthermore, as the cap domain is positioned after the second motif, PMM2 is specifically classified as a *type II* HAD protein⁴⁴. As shown below in **Figure 5B**, the enzyme's HAD core domain (aa 1-83 and 189-246) has a $\alpha\beta\alpha$ sandwich fold, connected by a

flexible linker region, linker 1 (aa 84-86) and linker 2 (aa 185-188), to a cap domain (aa 87-184) consisting of a two β -hairpin sheet topped by four α -helices. Two magnesium ions bind the protein: A structural Mg^{2+} (S_Mg) is coordinated at the surface of the core with α 8- β 11 loop residues F221, D223 and T226, being thought to stabilize tertiary structure^{25,47}. Also, a catalytic Mg^{2+} (C_Mg) is present at the active site, coordinated to D12 and D14 from the β 1- α 1 loop, which belong to HAD motif I, and D209 from the β 10- α 8 loop. The C_Mg is accessible at the core domain's interfacial surface, partaking in the enzyme's catalytic mechanism⁴⁷.

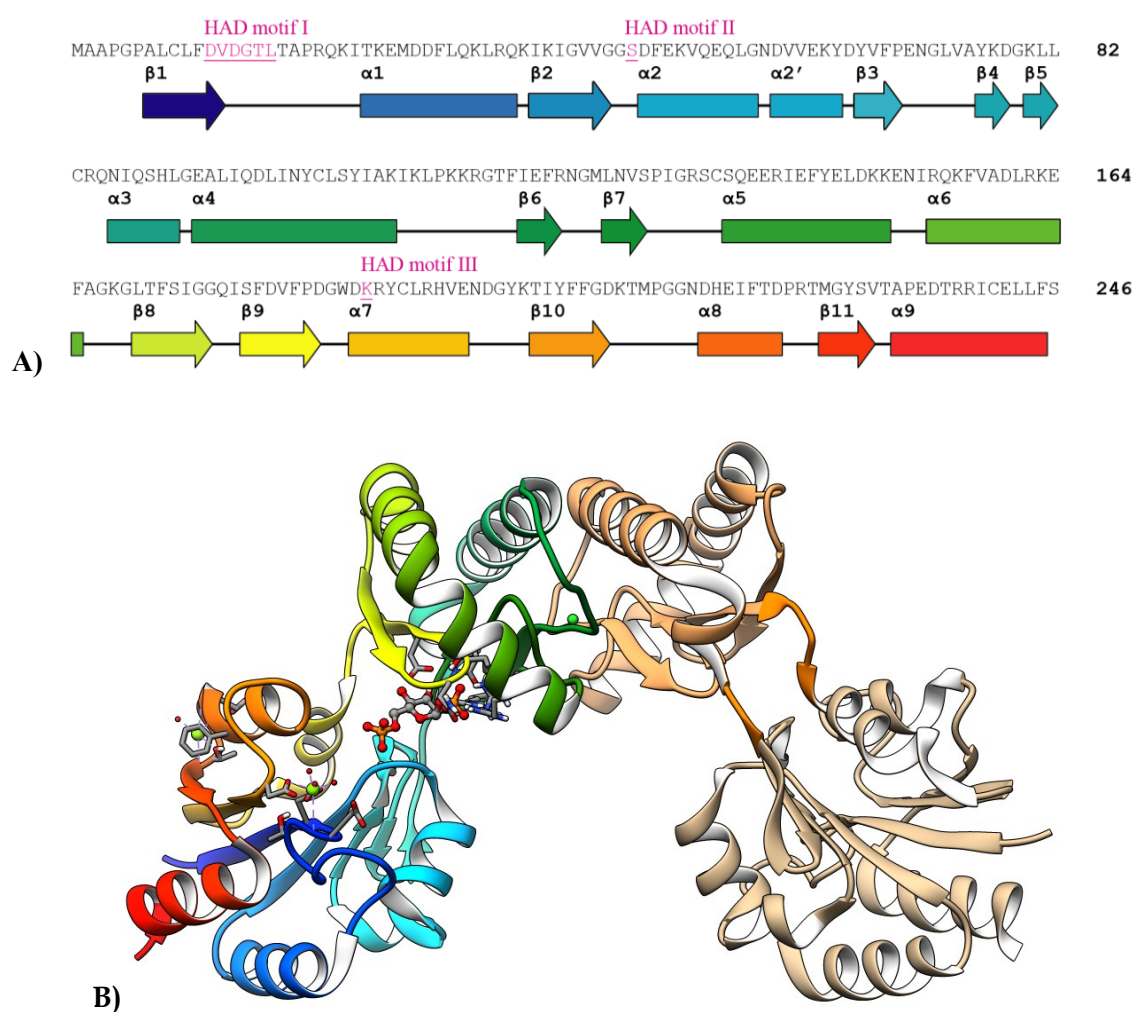


Figure 5. Structure representation of human phosphomannomutase 2 and primary sequence. **A)** hPMM2 primary sequence with secondary structure elements (β for β -sheet and α for α -helix) and highlighted in rainbow colouring, depicting relative residue positioning to N' and C' termini, and HAD motifs indicated in purple. **B)** Ribbon representation of hPMM2 dimer showing one monomer in rainbow colouring (left) bound with the activator Glc 1,6-BP in ball-and-stick representation and ion co-factors Mg^{2+} and Cl^- at their respective coordination sites, as green spheres. The other monomer (right) illustrates the protein's subdivisions with the core domain, in beige, the cap domain, in tan, and linkers 1 and 2 uniting both domains, in orange (PDB entry: 7O1B). Image generated using UCSF Chimera version 1.14.

It has been proven that hPMM2 is only catalytically active in its homodimer form²⁵, which features two subunits bound by their cap domains. Recent crystal structures showed that dimerization is established by interactions between the β 6 strand, α 4- β 6 loop and α 4 helix of each subunit and a Cl^- anion coordinates the dimer at the dyadic axis, strengthening its quaternary structure⁴⁷. As dimerization places the two active sites adjacent to each other, it is suggested that some sort of communication between both subunits happens during catalysis, being essential for enzyme function.

To become active, each monomer binds its bisphosphate sugar activator to hydrophilic residues from the exposed face of sheet β 6 and helix α 5 at the cap domain, as detailed in **Figure 5**. The precise orientation of the sugar is not completely clear: it most likely features the C1-phosphate bound to the cap, in agreement to models from related mutases⁴⁴ and the available crystal structures⁴⁷, although there may be two valid binding modes, where either the C1 or C6-phosphate interacts with the cap whilst the other projects towards the active site^{48,49}. Activator binding greatly affects structure: A conformational change is believed to take place, based on previous data from leishmanial models⁴⁴ and molecular dynamics modulations in hPMM2⁴⁹, where the two domains are drawn closer due to the mobility provided by the linker region, which is thought to accompany each catalytic cycle. This allows the co-factor's free phosphate to reach for a stable phosphate-binding site at the core comprising the C_Mg and neighbouring residues such as HAD motif III Lys189 and HAD motif II Ser47, as inferred from other HAD-like family members^{46,50}. This closed conformation is of extreme importance for enzyme dynamics, as the cap domain acts as a lid to seal the active site from the surroundings, creating a sheltered environment for catalysis to occur⁴⁸. Despite no actual experimental evidence of PMM2's closed form being available, it is an important aspect to consider when studying the intricate function of PMM2.

PMM2's catalytic action shares a similar mechanism as other functionally related HAD enzymes^{44,48,51}, involving a phosphorylated protein intermediate activated by its bisphosphate sugar co-factor. Catalysis may start as soon as the activator binds the protein, inducing its closure to approximate Glc 1,6-BP to the active site at the core domain. The beginning stage involves enzyme activation by transferring of the co-factor's phosphate leaving group at the active site to the HAD motif I aspartate nucleophile, D12 in hPMM2, through nucleophilic substitution ($\text{S}_{\text{N}}2$ mechanism), creating an aspartyl-phosphate intermediate, as depicted in **Figure 6**.

Once the Man 6-P substrate reaches the activated aspartate, phosphoryl group transfer occurs through another $\text{S}_{\text{N}}2$ reaction, forming a high-energy oxyphosphorane transition state, which features a pentavalent structure with D12 O δ 1 and Man 6-P C1O in the longer, weaker apical positions and the three phosphate oxygens in shorter, stronger equatorial positions of a trigonal bipyramidal geometry. As oxyphosphorane is intrinsically unstable, different active site residues, for example HAD motif III lysine (K189 in hPMM2), contribute to its geometric and electronic stabilization. In order for the bisphospho mannose intermediate to be formed, C_Mg provides shielding to the oxyphosphorane phosphorus centre by enhancing its electrophilicity, neutralizes the overall negative charge of the transition state and can also assist in positioning the nucleophile and electrophile⁵⁰⁻⁵³. C_Mg will also stabilize the D12 leaving

group by increasing aspartyl acidity through ionic interactions with its carboxyl oxygens. Nearby residues contribute as well to leaving group stabilization.

At last, the bisphosphomannose intermediate is released from the active site and it can either rebound to the same monomer to proceed with catalysis or travel across the dimer and bind at the other active site, in the unphosphorylated state. Note that, unlike other related phosphomutases^{54,55}, some hPMM1 and 2 form dimers in solution. This conserved feature corroborates the functional importance of dimerization, as this mechanism rejects the need for reorientation of the substrate within the active site, ensuring an efficient turnover⁵³. After rebinding, the Man 1,6-BP intermediate transfers its C6-phosphate moiety to the aspartyl nucleophile, in the same concerted reaction as before, assisted by the HAD motif I acid/base D14, to return the phosphorylated form of the enzyme, yielding the final Man 1-P product. The activator is also later replenished by the inverse of the initial reaction⁴⁴.

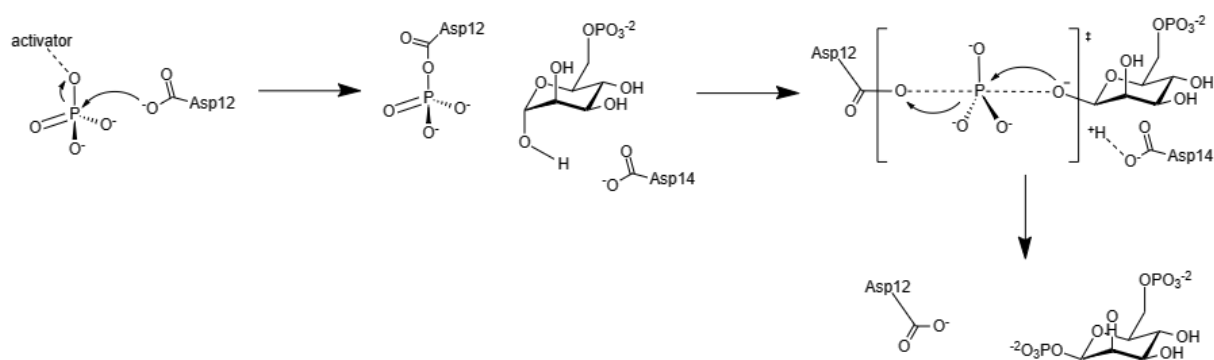


Figure 6. Simplified scheme of the reactional mechanism for catalysis in hPMM2. This image depicts the underlying mechanistic of catalysis in hPMM2, with enzyme activation by phosphate group transfer from the Glc 1,6-BP activator to HAD I motif Asp12 through a concerted S_N2 reaction. Next, the phosphate is transferred to the C1 of the Man 6-P substrate, forming an oxyphosphorane transition state in a similar nucleophilic substitution as before. The resultant Man 1,6-BP parts from the monomer's active site to be either re-oriented or re-bound to the active site of the other monomer for the opposite reaction to occur, detaching the phosphate group at its C6 to the enzyme (forming the activated intermediate) and then back to glucose to reconstitute the Glc 1,6-BP activator.

1.2.3 Impact of mutations in PMM2 enzyme function

As mentioned earlier, PMM2-CDG is a genetic condition resulting from mutations in the *PMM2* gene. As seen from the >900 reported cases, patients can be either homozygous for a hypomorphic mutation, meaning they inherited from both parents the same “milder” allelic variant encoding a protein with reduced function, or compound heterozygous, containing one hypomorphic and one inactivating variant, with the later producing a completely non-functional enzyme²⁷.

The first symptoms derived from PMM2-CDG manifest early on childhood and may include delayed cognitive development, failure to thrive, hypotonia and abnormal function of various organ systems. Nonetheless, milder clinical presentations have been observed in a reduced number of cases^{56,57}. The heterogeneity in disease manifestation found amongst patients and the profoundness of the symptoms

and the impact they have on the organism is in part dictated by the causative mutations that each patient carries. Identifying and understanding where within *PMM2* do different mutations occur and their effect on enzyme activity is essential given that disease-associated variants generally affect residue clusters of high structural importance. By performing a quick search for the *PMM2* gene on the Human Gene Mutation Database⁵⁸ (HGMD Public, latest public release as of February 2023, available at <https://www.hgmd.cf.ac.uk/>) it was found that a total of 144 mutations have been recorded, amounting to almost 140 pathogenic variants associated with PMM2-CDG, the majority ($\approx 83\%$) being missense mutations that lead to single amino-acid substitutions in 82 of the protein's 246 residues. Further below, **Figure 7** catalogues these known substitutions at the protein's sequence.

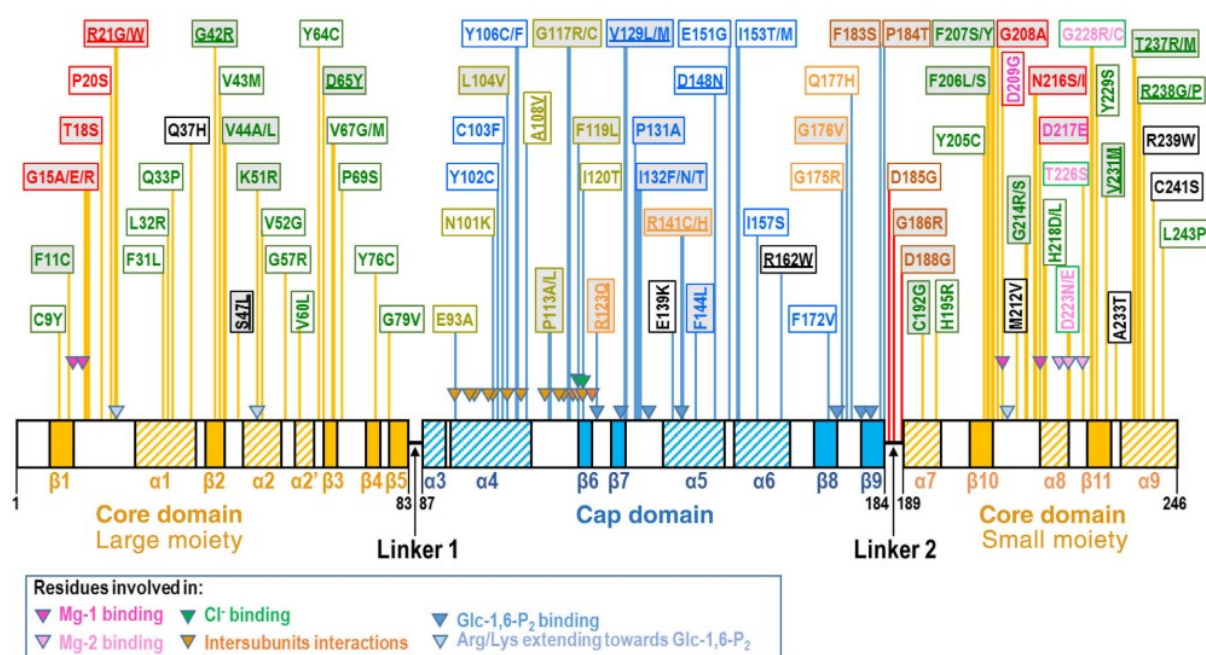


Figure 7. Distribution of amino acid substitutions found in PMM2-CDG patient pathogenic variants. The bar schematizes hPMM2's polypeptide chain with each domain indicated accordingly, along with secondary structure elements. All residue substitutions verified in their relative positioning are highlighted. Adapted from⁴⁷.

By analysing **Figure 7** it is inferred that the occurrence of patient-observed mutations along PMM2's primary sequence is, in fact, not homogeneous and the majority of pathogenic substitutions occurs in highly conserved amino acids/regions, indicating their relevance for the enzyme's structure and mechanism. A previous study summarized these mutations into groups according to their effect on the enzyme⁴⁷, with such being: effects on catalysis; activator binding; dimerization; cap and core domains folding; and linker flexibility.

All these different mutations ultimately have distinct effects on the catalytic function of the encoded enzyme. Thus, the residual activity of PMM2 is the remaining activity that results from the combined contribution of the mutated protein products derived from the expression of both pathogenic alleles.

Measurements for the residual activity of several PMM2-CDG mutants have been documented⁵⁹, revealing their activity patterns. Whilst all disease-associated mutations lead to a decrease in activity below 50% of the normal enzyme's, values greatly differ with a maximum around 47% and minimum around 0%, meaning complete loss-of-function, in relation to wild-type PMM2. Whilst patients presenting high residual activity variants tend to manifest milder phenotypes^{59,60}, a direct correlation between residual activity and symptom severity has not been verified and previous examinations confirmed that there is no evident genotype–phenotype relationship, adding that PMM2-CDG phenotypes can be also conditioned by other genes associated with glycosylation⁶¹.

Curiously, mutations in arginine 141, either p.R141H or p.R141M, are found to be rather common amongst PMM2-CDG patients, with paired heterozygous variants p.R141H and p.F119L being the most prevalent³⁴. Arg141 is crucial for the binding of the Glc 1,6-BP activator at the cap domain's $\alpha 5$ helix (*vide* **Figure 5**) and mutations at this residue are considered as loss-of-function, since no homozygous patients with R141 substitutions have been reported in all the clinical history of this disease⁵⁹. Indeed, null mutations encoding completely inactive hPMM2 products have never been detected in a homozygous fashion, most likely because total absence of activity is not viable³⁰.

It is also relevant to mention that the prevalence of a mutation is conditioned by geographical distribution, which is directly correlated to the mutational pool found within a certain population⁶². Therefore, variants found with an especially high frequency in a defined population, for example Portugal and Spain, are not as prevalent in the rest of European countries and vice-versa. In fact, the patient cohorts analysed in these two nations share a closely similar mutational spectrum and there are even variants, V44A and D65Y, thought to be original from the Iberian Peninsula, as they are exclusively seen in Iberian and Latin-American patients. By contrast, the very frequent European F119L variant (43% allele incidence in Danish patients) has not recorded in Iberian patients^{63,64}.

Genetic carriers also exist, as it is the case of parents of PMM2-CDG patients, bearing one wild-type allele and another with a mutated *PMM2* variant, either a fully inactivating or a hypomorphic mutation. Since the expression of the wild-type variant compensates for the defective allele, carriers retain 50% residual activity and do not suffer from any clinical symptoms as it has been observed so far even for the high number of carriers bearing the deleterious R144M/H mutation⁵⁹.

As an additional interesting remark, the processed pseudogene *PMM2* ψ , related to *PMM2*, has also been identified on chromosome 18p³⁵. Reportedly, processed pseudogenes may originate short interfering RNAs that contain a valuable regulatory potential for coding genes via the RNAi pathway and recent literature provides new insight on how these mechanisms are exerted and its contributions to cell function^{65,66}. Even though the function of *PMM2* ψ is not yet understood, its existence raises curiosity towards what might be its purpose in the genome and its implications in the pathological processes for PMM2-CDG.

1.2.4 Therapies for PMM2-CDG

As stated in **Section 1.1.3**, diagnostics for PMM2-CDG are mainly based on observing anomalies in the glycosylation pattern of serum transferrin. Recent MS strategies can be employed to examine the *N*-glycome expanding the diagnosis toolbox available but, regardless, these methods cannot distinguish specific genetic variants. Luckily, with the improvements in molecular diagnosis, namely WES and WGS approaches, it is becoming increasingly easier to perform single gene testing and next generation sequencing (NGS) panel testing to detect the underlying pathogenic mutations⁶⁷. Nevertheless, the number of biomarkers available to monitor the evolution of PMM2-CDG and progression of treatments is still scarce. In a recent effort attending this matter, the urinary sorbitol level was proposed as a novel biomarker, since it seems to be correlated with symptom severity in patients⁶⁸.

1.2.4.1 Current therapeutic options and perspectives

Presently, most treatment options proposed for PMM2-CDG are symptomatic, with a few emergent curative therapies in early stages of clinical investigation obtaining good preliminary results²¹. Some of these therapies are shortly mentioned below.

Phosphomannoisomerase (PMI) inhibition accompanied by mannose supplementation to upregulate the affected PMM2 pathway has been suggested, as well as bypassing this defect by administering membrane permeable Man 1-P directly³⁷. Liposomal targeting with the activated Man 1-P could also be an option despite its technical difficulties. These treatment routes raise important challenges such as the need for developing proper delivery systems and a tight control around sugar metabolic flux⁵⁹. Yet, sugar supplementation in different CDGs constitutes the oldest and one of the most employed successful therapies²¹.

Organ transplantation is also used, mainly in clinical presentations with serious cardiac and hepatic involvement, with many records of successful outcomes in patients²¹.

Enzymatic therapy has been described to be successfully used in another genetic metabolic disorder, lysosomal storage disease, by supplying cells with the wild-type variant of the defective enzyme. Whilst in theory this is possible for PMM2-CDG, complications related to blood–brain barrier permeability, cellular uptake and compartment targeting hinder its usage³⁷.

Another alternative to explore would be gene therapy, with the purpose of restoring a wild-type *PMM2* allele by means of transgene introduction. Despite its potential, no clinical trials have been performed in PMM2-CDG to date, only a research study using antisense therapy to treat pathogenic splice variants in patient-derived fibroblasts is documented⁶⁹.

Drug repositioning has also been developed as a potential therapeutic option for this disorder. This strategy involves the use of an approved drug in a new clinical setting, for a different application than the originally intended⁷⁰. Drug repurposing studies for PMM2-CDG are currently being developed with different drugs, including:

Acetazolamide (**Figure 8A**), which is a non-competitive carbonic anhydrase inhibitor whose action lowers cytoplasmic pH, increasing intracellular Ca^{2+} levels and ultimately interfering with transmembrane potential⁷¹. Acetazolamide administration is approved for retinal complications, epilepsy and

other nervous system diseases. It is hypothesised that some PMM2-CDG neurological manifestations, such as ataxia and stroke-like episodes, are prompted from defective glycosylation patterns in calcium channels⁷², additionally causing an adverse increase in Ca^{2+} , which inhibits PMM2. Thus, acetazolamide is currently under clinical trials (EudraCT number 2017-000810-44 at www.clinicaltrialsregister.eu) to further examine its therapeutic effects for cerebellar syndrome in PMM2-CDG patients. Up to now, early outcomes indicate adequate tolerability amongst most patients and progresses in motor and cognitive symptoms have been registered⁷³.

Other repositioning example is celastrol (**Figure 8B**), an anti-inflammatory drug used in cancer and inflammatory diseases⁷⁴ that is reported to function as a proteostasis regulator by triggering the heat shock response (HSR) in cells. Therapeutics employing proteostasis regulators for conformational diseases, such as PMM2-CDG, have been explored^{75,76} and recent celastrol testing in patient fibroblast lines showed augmented PMM2 levels and improved residual activity.

Epalrestat (**Figure 8C**), currently prescribed for diabetes-related neuropathies⁷⁷, has also been studied due to previous observations indicating increasing effects on PMM2 activity in patient-derived fibroblasts⁷⁸. This aldose reductase inhibitor disrupts the conversion of glucose to sorbitol (catalysed by aldose reductase) boosting intracellular Glc levels for the synthesis of PMM2 activator, Glc 1,6-BP. An ongoing phase 1 clinical trial aims to evaluate its tolerance in a PMM2-CDG child, already revealing promising results, with improvements in serum transferrin isoelectric focusing and body mass index⁶⁸.

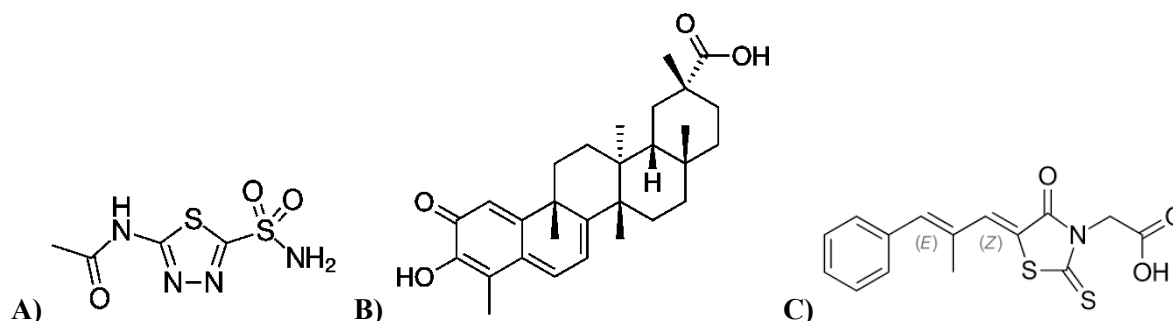


Figure 8. Repurposing candidates for PMM2-CDG. A) Acetazolamide (*N*-(5-sulfamoyl-1,3,4-thiadiazol-2-yl)acetamide), B) Celastrol (2R,4aS,6aR,6aS,14aS,14bR)-10-hydroxy-2,4a,6a,6a,9,14a-hexamethyl-11-oxo-1,3,4,5,6,13,14,14b-octahydronicene-2-carboxylic acid), C) Elpalrestat (2-[(5Z)-5-[(E)-2-methyl-3-phenylprop-2-enylidene]-4-oxo-2-sulfanylidene-1,3-thiazolidin-3-yl]acetic acid).

1.2.4.2 Pharmacological chaperoning applied to PMM2-CDG

Besides the aforementioned treatment routes, another option has been receiving particular focus in the later years. PMM2-CDG is fundamentally considered a conformational disease^{59,69}, meaning that its pathological phenotypes arise from errors in protein folding caused by destabilizing mutations. In conformational diseases, the cell's quality control machinery targets and clears the mutated protein, leading to a decrease in its population and total activity^{79,80}. Given that nowadays countless known diseases are

a result of protein misfolding, the proposition of applying a therapeutic compound with the potential to restore a target's native-like folding and consequently return its function to close-to-normal levels holds great promise^{81,82}. Pharmacological chaperones (PCs) formally appeared as a defined term around two decades ago⁷⁹. They are cell-permeable, small molecules⁸³ with the ability to selectively bind a misfolded protein target (such as an enzyme, ion channel, transporter or even a receptor^{80,81}) and stabilize it by triggering folding in a ligand-assisted manner, eventually increasing the correctly folded population, which can undergo trafficking for the intended cell compartment and perform its function^{79,81,83}.

So far, the majority of reported PCs are first-generation, reversible competitive inhibitors or antagonists of their targets. Though it may seem counterproductive to employ an inhibitor with the end goal of restoring proper function in an aberrantly folded protein, it is important to recall that, when an inhibitor binds to an enzyme's active site, it establishes several different interactions with the neighbouring residues to stabilize the proper conformation and, with it, partially rescue enzyme activity^{84,85}. Nevertheless, the main constraint of first-generation PCs is related to the intricate balance between activity enhancement and inhibition, the latter being, of course, not desirable. Designing novel, second-generation, non-inhibitory PCs may therefore be a likely alternative to overcome the risks of unwanted inhibition^{79,82}.

Due to the extensive applicability of PCs to rescue activity in pathogenic variants, pharmacological chaperoning has become a relevant tool to advance therapies for a range of conformational diseases^{21,86}. Some examples of conditions benefiting from these current efforts are the lysosomal storage disorders (including Fabry, Pompe, Krabbe and Gaucher's) which comprise the most popular targets; cystic fibrosis; nephrogenic diabetes insipidus; GM1-Gangliosidosis; Alzheimer's; amyloidosis; congenital erythropoietic porphyria; phenylketonuria; methylmalonic aciduria *cblB* type and even other cellular targets like gonadotropin-releasing hormone (GnRH) receptor; G protein coupled receptors (GPCRs) and ion channels⁷⁹⁻⁸³.

Pharmacological chaperoning is also endorsed as a suitable candidate for PMM2-CDG therapy, as suggested in **Figure 9**. A study assessing pathogenicity in several patient-observed missense variants described the reduced protein stability observed in certain PMM2 folding mutations to be the basis for partial loss of function, with the defective proteins retaining some residual activity⁵⁹. It is thus proposed that the stabilizing effect of a pharmaco-chaperone may be used to promote the native conformation of the protein, reducing the degradation of defective hPMM2 variants leading to an increase in residual activity. Yet, it is critical to understand that pharmacological chaperoning is not a "*one sizes fits all*" solution: As depicted in **Section 1.2.3**, whilst, in PMM2-CDG, over 80% of patients carry defective enzymes arising from missense alleles that disrupt protein conformation – which may be treatable with PCs – frameshifts and premature stop codon pathogenic variants originating a truncated product do not represent targets for PCs. Likewise, eventual mutations affecting conserved key residues essential for the catalytic mechanism, such as the HAD motifs, (*vide* **Section 1.2.2**) are certainly not mendable by chaperoning.

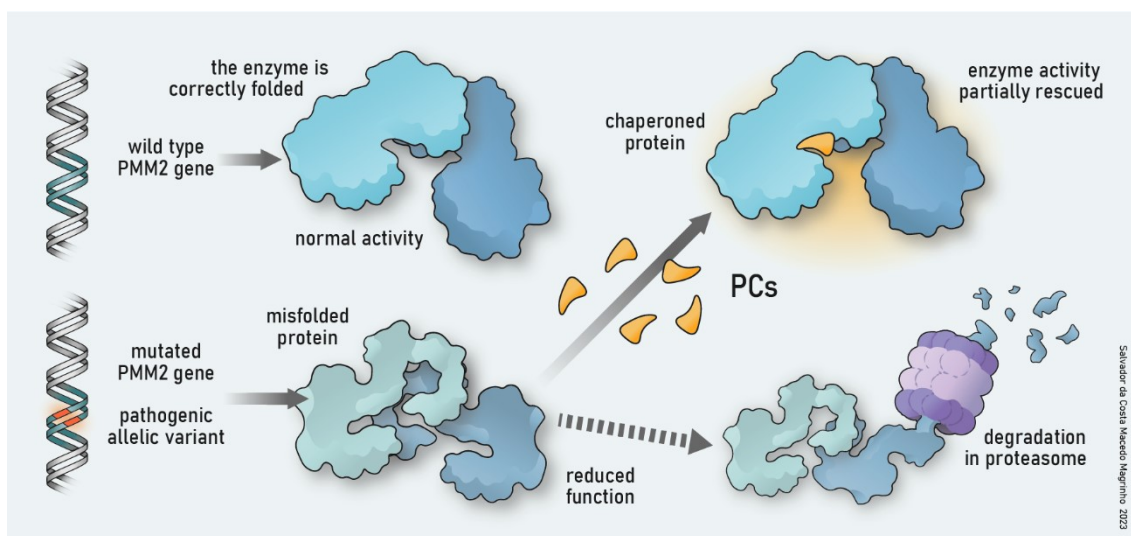


Figure 9. Simplified illustration of a pharmacochaperoning approach for PMM2-CDG therapy.

In later years, studies on PC development for PMM2-CDG reported by Yuste-Checa and collaborators included a high-throughput screening approach to find several chaperoning candidates from a ten thousand-compound library. Through thermal shift and enzymatic activity assays it was possible to identify a few promising compounds that enhanced structural stability in defective PMM2 variants. From the obtained outcomes, 1-(3-chlorophenyl)-3,3-bis(pyridin-2-yl)urea, known in literature as “compound VIII”, was considered as especially pertinent to proceed investigation regarding its chaperoning potential, to be followed by chemical optimisation⁸⁷. Other work by Monticelli and collaborators explored the rationale of using the β anomeric form of the enzyme’s own bisphosphate activator, β -Glc 1,6-BP, as a lead-like chaperone for stabilizing pathogenic protein variants. This catalytically inert ligand was seen through *in silico* experimentation to bind the enzyme, triggering a conformational change, similarly to the natural α -Glc 1,6-BP co-factor, that was validated *in vitro* by limited proteolysis and further studied with thermal shift assays to elaborate on its ability to stabilize both wild type and mutated PMM2 variants. Concerns regarding bioavailability of this potential chaperone are currently being investigated⁸⁸.

1.2.4.3 Overcoming obstacles to drug discovery in PMM2-CDG using drug repositioning allied to artificial intelligence

Conventionally, a standard drug research and development (R&D) process can take up to a decade or longer and require an enormous financial investment⁸⁹, which usually tend to be even greater for rare diseases⁹⁰. Over ten thousand diseases are identified as rare, many of them being multisystem and life threatening, like CDGs. Estimations indicate that less than 5% of these have an approved treatment and are widely underrepresented in drug R&D⁹¹, owing to several aspects such as: Underdiagnosis and low number of reported patients, which leads to underestimating the disease frequency; heterogeneity in

clinical presentations; insufficient biomarkers; scarcity of data related to pathophysiologic patterns and generally low commercial incentives and lack of funding^{92,93}. CDGs comprise a good example of all these hurdles.

Drug repurposing (DR) comes forward as an alternative approach to conventional drug R&D, particularly interesting for underrepresented diseases, with the potential to speed up this process and lower its financial weight, which is fundamental in the rare disease panorama⁹⁴. DR is based on redeveloping already approved drugs for a new clinical setting, different from the original application, by screening their biologic activity towards a different biochemical target, in search for potentially therapeutic effects⁹³. By employing these strategies, it may be possible to bypass assays regarding toxicity and pharmacological concerns, as these were contemplated in the original research⁹⁵.

The advancement of therapies for PMM2-CDG (and rare diseases in general) could benefit from pairing DR with artificial intelligence (AI)⁹⁶. The computational advances in AI have been expanding in the last years, with a growing number of potential applications for biomedicine being described⁹⁷. The role of AI in drug discovery has been the subject of recent studies providing new insights about current disease applications, types of input data and used algorithms⁹⁶. Adding to this, the increase in available data obtainable through new technologies like high throughput screenings and omics facilitates testing of large datasets of molecules to investigate their biochemical characteristics, target interactions and associated pathways to identify compounds with the desired therapeutic profile, making the process more efficient and less costly^{93,98}. Despite drug repurposing not being a novel strategy, its practice in conjunction with AI-assisted methods is far from being a standard, well-established approach⁹⁹. Nonetheless, it could be the much-needed push in current CDG drug development efforts.

As mentioned before, drug repurposing in PMM2-CDG is already being advanced with candidates like acetazolamide, celastrol and epalrestat. Yet, no pharmaco-chaperoning strategy for PMM2 has been associated with a comprehensive drug repositioning methodology, nor AI studies. This thesis explores the potential of artificial intelligence algorithms for unveiling repurposed drugs as prospective chaperones for stabilizing PMM2 structure and restoring enzyme activity.

The next chapter makes a thorough description of the essential aspects studied during this Master's project related to the application of cheminformatics and AI to the drug development thematic, presenting the underlying objectives of this study and its scope in developing a pipeline to advance potential pharmacological chaperones for PMM2-CDG through drug reprofiling.

1.3 Cheminformatics: A Computer-Aided Drug Design (CADD) approach to PMM2-CDG therapies

Quite simply, cheminformatics is the application of informatics methods to solve chemical problems¹⁰⁰. Since the start of its expansion a few decades ago, *in silico* chemistry has been broadly prospected for its potential applications in drug discovery. Indeed, in 1988, when Brown coined the expression "cheminformatics", he merely described the field as a combination of all information resources that a

chemist requires to optimise the properties of a ligand for it to become a drug¹⁰¹. Nowadays, the definition of cheminformatics is far wider than the original but, nonetheless, these tools hold great promise for accelerating therapies, especially in rare diseases, where the lack of data and disease models deems traditional drug R&D approaches not applicable^{102,103}.

Opting for massive virtual screening-based strategies became a possibility with cheminformatics, allowing researchers to computationally screen compound databases containing thousands, or even millions, of different molecules towards the pharmacological target in order to pinpoint structures of interest to further proceed to biological assays¹⁰⁴. Virtual screening is therefore a viable option to explore and identify pharmacophores for chaperoning⁸², as intended for the scope of this project, playing an increasingly important part in drug repositioning for rare diseases like CDGs¹⁰³.

1.3.1 Chemical Structure Representations & Molecular Descriptors

Since the last century, the amount of chemical information published by chemists around the globe has grown to such enormous dimensions that it is not possible to handle it manually: The only way for it to be accessible is through an electronic support, i.e, databases. In fact, chemical databases are an essential tool for anyone working in chemistry, containing collections of data for countless compounds, curated in a structured form to simplify extracting information about all kinds of properties according to a researcher's needs^{104,105}. For example, PubChem currently describes almost 114 million unique known compounds and this number grows approx. 4% every year¹⁰⁶. This amount of data created the necessity of using cheminformatics as a tool for visualise, manipulate and manage chemical information¹⁰⁷.

A key requirement to using computers for the input and output of chemical data is molecular representation¹⁰⁷, i. e., presenting molecular structures in a machine-readable manner, understood both by computer and operator. As computers have a unique way of handling data, coding is required to manipulate chemical structures: By using different symbols in a system under a set of observed rules, molecules can be characterized by its name, two-dimensional diagram or even three-dimensional model using scientific, artificial languages which consist of chemical nomenclatures and notation systems¹⁰⁵.

An important example of a language widely used in Cheminformatics is line notation, which represents chemical structures and reactions through single linear sequences of different letters, numbers and other characters. Their compactness and the fact that they can be easily processed makes them very appealing for computer science¹⁰⁵. Some examples of line notation used during this project include SMILES, SMARTS and InChI systems.

In 1986, David Weininger created the Simplified Molecular Input Line Entry System, or SMILES, which was inspired by the molecular graph theory. It allows structures and reactions to be represented by a string that employs a small character vocabulary as atom and bond symbols, with simple rules that provide information on double bonds, stereochemistry and chirality to allow an unambiguous notation. SMILES are widely used since its compacted textual coding does not require a graphical input, being

ideal for databases. On the other hand, SMILES cannot be used for three-dimensional representation^{105,108,109}. Along the years, SMILES were expanded to additional languages, such as SMARTS, SMILES Arbitrary Target Specification, an extension that introduced new rules and logical operators, useful for describing molecular patterns. SMARTS are predominantly used for substructure searching, based on finding a certain pattern (subgraph), like a functional group, in a molecule (graph), making it suitable for structure similarity analysis and drug design¹⁰⁸.

Another example of line notation is the International Chemical Identifier, InChI, developed about twenty years ago by the International Union of Pure and Applied Chemistry (IUPAC), who regulates the IUPAC systematic nomenclature broadly used by researchers. This canonical system is based on hierarchical layers where the main first layer denotes the compound's chemical formula, based on Hill notation, and other sub-layers represent atomic bonds. Following layers are added as needed to provide more information regarding charge, stereochemistry, isotopes and fixed hydrogens. By taking the InChI notation of a certain molecule, a hashed identifier called InChI Key can be generated, containing a standard key of 27 characters, coding chemical data in four layers¹¹⁰.

Whilst line notation is rather useful, it does not provide information about molecules in the three-dimensional space. For this, different file formats have been developed to efficiently serve as an exchange format for chemical structures, MDL files, created by *MDL Information Systems, Inc.* in the 1980s, being the most used and standardly accepted nowadays. MDL files comprise many format extensions, including Molfiles and SDfiles (Structure-Data file), the latter being used to describe any number of molecules, making it particularly convenient for processing large sets of compounds. In an SDfile, each compound is denoted by three components: Firstly, a header block, secondly a connection table. This table holds the complete description of all atoms indicating their Cartesian coordinates, which define the structure as two or three-dimensional; atomic symbol; mass; charge and other properties. Then, a bond block specifies which atoms are bonded, as well as multiplicity and stereo configuration. Lastly in the connection table, a properties block designates different structural properties. The third component of SDfiles compiles additional items describing non-structural properties like biological activity or other molecular descriptors^{105,111}.

One major task in Chemoinformatics is to correlate experimental observations like physico-chemical properties and biological activity with structural information and how this correlation can be employed for predicting new data¹¹².

A common way of doing such is to use Molecular Descriptors, mathematical depictions of a molecules' properties, created by a logical procedure that transforms the chemical information encoded by a symbolic representation into a value or a result of a standardized experiment^{100,105}. The content of a descriptor depends on the algorithm used for calculation and input (the representation). They can be structural descriptors, encoding properties of the input structure, or Fingerprints (FP), which are Boolean vectors (every bit is either true or false) that encode the presence of a certain structural fragment or

moiety^{105,107,112}. Molecular descriptors can be categorised as constitutional, topological, geometrical, electronic or quantum-chemical, depending on what they encode. Structural descriptors can additionally be classified in terms of dimensionality (0D, 1D, 2D, 3D or 4D structural descriptors): 0D descriptors are generated based on the simplest representation, the chemical formula, codifying atomic elements and their occurrences with no data on atom connections. 1D and 2D descriptors denote, in addition to atom composition, connectivity in terms of a molecular graph and its topological distances. Yet, as they do not signify the entire topology of a chemical structure, 3D descriptors were developed to add topographical information, including a set of spatial coordinates that describes the molecule as a three-dimensional object. 3D structures encode molecular conformation and configuration, which are required to calculate descriptors related to a molecule's geometry, spatial constraints and steric hindrance. Further 4D representations characterize additional chemical properties derived from, for example, the molecule's electronic distribution or potential interactions with a ligand-binding site at a target^{105,113}. Not all descriptors are interpretable, i.e., they do not have a perceptible physical meaning, although their predictive capacities may be extremely useful for exploring a certain experimental question. In contrast, descriptors with low predictive capacity may come in handy in models requiring an easily interpretable structural aspect¹¹². All in all, considering the wide diversity of descriptors available for coding different information, it is essential to comprehend the association between structure and experimental evidence.

In turn, molecular Fingerprints (FPs) are encoded in the form of binary bit strings (bits of 0s and 1s) that describe the presence or absence of a certain property, such as a fragment or a connectivity pathway at the input structure. Whilst almost all FPs follow this binary behaviour, counting FPs also occur. Different categories of Fingerprints exist, like structural keys, circular FPs and path-based (or topological) FPs. Structural key FPs encode the input molecule according to pre-defined features from a fragment library and each bit position is set as 0 or 1 depending on the presence of each feature. Examples of structural keys include MACCS and PubChem Fingerprints^{114,115}. Hashed FPs do not observe a fragment library and instead are generated by enumerating molecular fragments, converting them afterwards into numeric values employing a "hash" function for bit positioning. These include topological FPs, which account fragments of a structure by pursuing a linear path up to a specified number of bonds within the molecule, and circular FPs which are calculated by considering the surrounding "chemical neighbourhood" of each atom inside a defined maximum radius to generate the Fingerprint¹¹².

Today, several softwares developed for molecular descriptor calculation are freely available, including the one used throughout this project, PaDEL-Descriptor (Pharmaceutical Data Exploration Laboratory) version 2.21¹¹⁶. PaDEL-Descriptor is an open-source software developed in java language by Chun Yap, with the ability of calculating 1875 descriptors (1444 1D&2D descriptors and 431 3D descriptors) and ten types of Fingerprints, using the Chemistry Development Kit (a collection of modular Java libraries for processing chemical data). Extra descriptors and FPs encoding atom type electro-topological state, ring count, McGowan volume, amongst others, are also included in PaDEL-Descriptor.

1.3.2 Structure-Property and Structure-Activity Relationships

The final goal of a drug design strategy is to generate compounds with the desired properties or activities. To accomplish this, it is vital to understand which chemical structures can be associated with the intended property or activity. The practice of defining structure-property or structure-activity relationships, SPR or SAR, and describing them through quantitative methods, QSPR or QSAR, allows chemists to relate molecular structure to biological activity and pharmaceutical effects^{100,117}. QSAR models employ statistics to search for reliable mathematical correlations between the experimentally-described biological activities of a set of compounds towards a target and their calculated molecular descriptors, i.e., encoding their chemical structure. These relationships can be then used to evaluate the expected properties of new molecules, on the basis that structurally similar compounds yield similar activities towards the same target^{117,118}. The typical QSAR workflow encompasses numerous successive steps, namely: Dataset building and processing, descriptor calculation and selection, model building and validation, being that, the functionality of the final QSAR model deeply relies on how each step is carried out¹¹⁹.

QSAR modelling is intimately connected to the calculation of a large variety of molecular descriptors, since they encode the underlying chemical structures on which a model is built upon¹⁰⁵. Building a proper QSAR approach depends on selecting the descriptors that best describe the relationships that one is trying to predict in order to answer the experimental question as the model's independent variables. Whilst it may be possible, in some cases, to build models with all the calculated descriptors for the input dataset, great advantages are reportedly taken from performing molecular descriptor selection¹¹⁸.

Descriptor selection is a crucial step in model optimisation, centred on removing redundant descriptors from the model, as they do not contribute to its predictive accuracy, and thus reducing the model's data input size without losing information important for property prediction. This selection is far from senselessly removing descriptors from the dataset just because: When developing a QSAR system, it is expected for only a small subset of descriptors to denote the truly necessary information for the mathematical model to make its predictions. As so, additional descriptors generate noise by biasing the model into focusing some individual molecular features, causing it to lose sight of the "big picture" and ultimately affecting its predictability. Reducing this input data leads to a simpler and probably faster model, as dealing with fewer variables relieves the burden on the machine learning algorithm¹¹⁸. Currently, there are several sorts of available methods for descriptor selection, ranging from more classic, statistical methods to artificial intelligence-based approaches, amongst others. Furthermore, some machine learning techniques have their own metrics to measure descriptor importance^{118,120,121}.

A QSAR strategy must also contemplate a domain of applicability. Contrarily to early QSAR efforts, which were built with a relatively small set of compounds related to a certain chemotype, current QSAR models employ widely diverse, large datasets¹²². Whilst this produces a richer depiction of structure-activity relationships, the model's chemical space becomes ever more complex and disperse, which can negatively affect the prediction of new classes of compounds that are not well represented at the training

set used to build the model. To overcome this hurdle, it is crucial to define the model's domain of applicability, which is an estimation on the uncertainty in the prediction made for a particular query molecule based on how similar this input is to the training set compounds used to build the QSAR model. By quantifying the prediction error, it is possible to understand if the model is suitable to examine the query set (subject molecules of interest), this is, if such compounds fall within the applicability domain and will, therefore, return a reliable activity prediction^{118,122}.

Note yet that QSAR models can be grouped into classification or regression approaches: In a classification-based QSAR, the model will strive to establish a relationship between input descriptors and the categorical values of the property being studied (the variable). The output obtained in binary classification QSARs are predictions for the classification of the query into the predefined category, meaning that the model will generate an either "true" or "false" output (the submitted molecule is predicted to belong to the category or not). As the model building process comprises an internal evaluation through cross-validation¹²³ and, later, with the testing set partitioned from the original dataset (developed in **Section 1.3.3**), by comparing the model's categorical output predictions with the experimentally-obtained category, it is possible to classify the output for each instance with the following metrics: True Positives (TP) for compounds that belong to the variable category and are predicted as such; True Negatives (TN) for compounds that do not belong to the category and are predicted as such; False Positives (FP) for compounds not belonging to the variable category but being predicted as such; and False Negatives (FN) for compounds belonging to the variable category and not being predicted as such. In order to measure and evaluate the model's predictive performance, different statistical parameters can be used, including the Sensibility (SE), the Specificity (SP), the Matthews correlation coefficient (MCC) and the Overall Predictability (Q)^{124–126}. Their equations are described below.

$$\text{Sensibility (SE)} = \frac{TP}{TP + FN}$$

1

$$\text{Specificity (SP)} = \frac{TN}{TN + FP}$$

2

$$\text{Mathew's Correlation Coeff. (MCC)} = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}}$$

3

$$\text{Overall Predictability (Q)} = \frac{TP + TN}{TP + TN + FP + FN}$$

4

In a regression-based QSAR, the focus will be to find a quantitative correlation between the dependent variable (the target property or activity) and the input descriptors^{127,128}. A regression model will generate, for each instance, a numeric value that represents the “potency” of the response variable. These QSAR models can be evaluated by employing different parameters such as: The Pearson’s Correlation Coefficient (R), which measures the strength of a linear relationship between two variables (in this case the predicted activity value and the experimental one) generating a number between –1 and 1; the Mean Absolute Error (MAE), which quantifies the average deviation of the predicted value in comparison to the experimental; and the Root Mean Square Error (RMSE), which indicates the squared root of the mean squared error between the predicted variable and the experimental value^{129,130}. The equations used for calculating these parameters are showcased below:

$$\text{squared Pearson's Correlation Coeff. (R}^2\text{)} = 1 - \frac{\sum_{i=1}^n (y_i - x_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2}$$

5

$$\text{Mean Absolute Error (MAE)} = \frac{\sum_{i=1}^n |y_i - x_i|}{n}$$

6

$$\text{Root Mean Square Error (RMSE)} = \sqrt{\frac{\sum_{i=1}^n (y_i - x_i)^2}{n}}$$

7

Where y_i is the experimental activity value, $\bar{y}_i$ is the average of experimental activity values, x_i is the predicted variable value and n is the number of instances (molecules) in the dataset.

1.3.3 Machine Learning in CADD approaches

QSAR models can be designed using different well-established machine learning technologies. Machine learning (ML) is an evolving field of artificial intelligence devoted to building computational algorithms with the ability to mimic human intelligence by learning from massive datasets¹³¹. This learning process is called training and, to build a machine learning model, the first step is selecting the model’s dataset, such as a library containing compounds of interest, followed by its partition into a train and a test set. The training set comprises the majority of the total dataset and is effectively used by the ML algorithm to train from the examples contained within. Once a model is built, the quality of learning – which translates to the predictive accuracy in QSAR studies – is assessed using the test set¹⁰⁵. This set should represent at least 10% of the total dataset and must be kept from the model building process, so that

these structures, from whom the algorithm did not learn, can provide an unbiased trial of the model. Overall, the test set simulates a set of novel query compounds for which the algorithm generates a prediction for the property under study. Since the true experimental value for such property is known, it is possible to infer an estimation on how well the model will perform for new molecules^{104,105}. Using a test set to evaluate a model is crucial since the end goal of a ML algorithm is to be able to generalise, i.e., correctly predict an output. If the algorithm limits itself to just memorise the input data, it may provide a decent prediction for the internal training set but, if it does not effectively learn from it, it will not be able to generalise for the external test set and will yield an erroneous output¹⁰⁵.

Note that both train and test sets must be well balanced, meaning that, for whatever property being experimentally examined, an equally proportional representation of compounds from the same class must be observed. As a typical example in a drug design scenario, one can imagine a dataset containing molecules active towards a certain protein target, and others not active. In this case, the train and test sets will have proportionally equivalent amounts of both active and inactive ligands.

An important remark to bear in mind when working with ML algorithms in drug discovery, and even artificial intelligence in general, is that these approaches are intrinsically dependant on the quality and completeness of the datasets used, the so-called *garbage-in, garbage-out* philosophy. As so, the development of progressively more accurate ML technologies is only possible if accompanied by advances in compound screening methodologies to efficiently measure the information relevant for each study, whether it may be class identification, chemical reactivity, biological and pharmacological effects, amongst many other examples¹³².

Different ML strategies can be based in either unsupervised or supervised learning concepts. In this project only supervised learning (SL) was approached. This type of ML aims to make the model learn to correlate the input data provided to a set of known target outputs, with the objective of learning to generate the correct output result for every newly given input. This output is compared with the true correct output to obtain an error estimation, which is tendentially minimised by the algorithm as it progresses. Amongst the range of ML techniques that apply SL are the Random Forests (RFs), Support Vector Machines (SVMs) and Neural Networks (NNs) algorithms, which were the ones selected for conducting the present thesis project and are shortly described in the sections below.

The application of ML to a CADD workflow, specifically to drug repurposing efforts, has been reported in literature, with accounts of ML being used to successfully identify potential drug candidates for other rare diseases¹³³.

1.3.3.1 Random Forests

A Random Forests algorithm (RF) is a classifier based on the combination of decision trees. A decision tree, shown in **Figure 10**, is a graph consisting of nodes and branches. In order for an instance to be classified, it will be sorted down the tree from the root node to the final leaf node. Each node represents a defined attribute or choice and, depending on which attribute the instance fits into, classification will

proceed by taking either one of the node's branches. By reaching the edge of the tree, the leaf node, the final condition (or classification) is achieved¹⁰⁵.

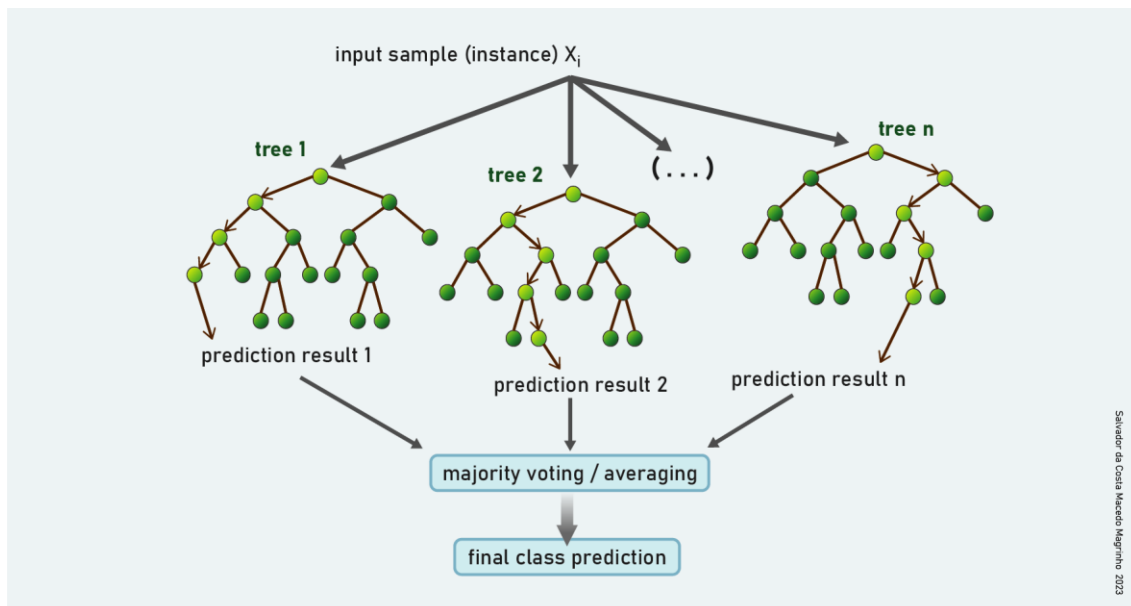


Figure 10. Schematic illustration of a Random Forest algorithm setup

As in the natural world, a random forest is created by growing an ensemble of decision trees, that will latter by applied in voting for the most popular class for each instance¹²¹. Not every tree in a forest grows the same, the RF algorithm employs random vectors to control the growth of each tree in the ensemble. The common principle of these vectors is that, for the tree number k , the random vector Θ_k is generated as independent of past random vectors for the other trees ($\Theta_1, \dots, \Theta_{k-1}$), but with the same distribution, and each tree grows by using the training set and its Θ_k , creating a classifier $h(\mathbf{x}, \Theta_k)$, where $\mathbf{x}$ represent an input vector.

These vectors can be generated by different methods like bagging, where Θ_k is created based on counting N darts thrown at random at N boxes, where N is number of examples in the input training set. When an instance from the training set reaches each node, the split (branch choosing) is made using the molecular descriptor that maximizes the separation of the set. To select this descriptor, a parameter called *mtry* narrows down the options to a set of best performing descriptors for the node. Another method is random split selection, where at each node in the tree, the branching split is randomly selected from amongst the K best splits¹³⁴.

During model building, each tree is grown using a distinct “out-of-bag” set, a random subset made out of a third of the total training set, being that the remainder set is used for training¹²¹. The final output prediction is conducted by cross-validation, which considers the average of the predictions from each out-of-bag set, hence from each tree, for each instance. In regression-type RFs the output is the average of each trees' predicted values and in classification RFs is the majorly voted class prediction¹²¹. The

generalization error of a RF model is associated to the sturdiness of each individual tree and their correlation: as the number of trees in the forest increases, this error converges to a limit.

1.3.3.2 Support Vector Machines

Support vector machines (SVM) were first developed by Vapnik and Chervonenkis in the 1960's and are typically used in binary property prediction models, although they may be used for either classification or regression QSAR approaches¹³⁵.

To start, the algorithm projects the input training set into a high-dimensional feature space where the different classes represented in the set become linearly separable. This projection is performed by using a kernel function, which could be either linear, polynomial, sigmoidal or radial basis functions^{136,137}. Afterwards, a linear boundary called hyperplane is generated using support vectors, which are a subset of examples from the input training set¹³⁶⁻¹³⁸. Once the test set is submitted to the model, the molecules contained within are mapped in the same feature space and a property class prediction is produced based on which side of the feature space, separated by the hyperplane, each test compound falls into (*vide Figure 11*). The distance of each test compound to the hyperplane, named "margin", is used to estimate a confidence level to the prediction accuracy: the wider the margin, the more confident the prediction is¹³⁶. Therefore, SVM algorithms will try to construct the hyperplane that maximizes the margin between the two classes. A soft-margin hyperplane is common when dealing with datasets containing inseparable classes.

For every SVM model, the kernel function parameters, and especially the error penalty factor or cost, C , is the most important parameter to optimise as it conditions the model's performance by regulating the balance between misclassification of training examples and margin maximization^{135,137}.

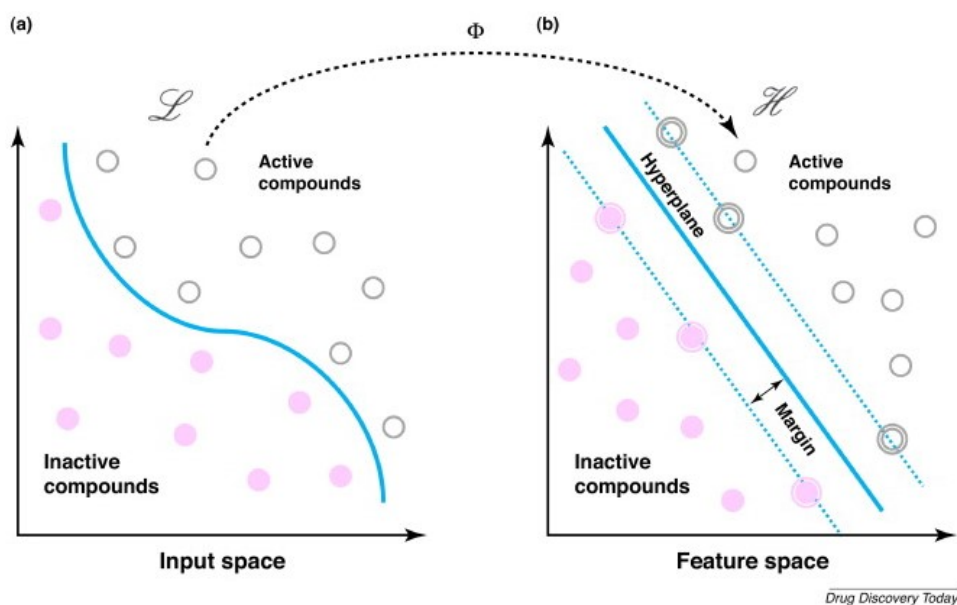


Figure 11. Illustrative example of a Support Vector Machine. The input set, which is represented in the left by active (empty circles) and inactive (pink circles) compounds, is not linearly separable in the low-dimensional input space. By applying a mapping function Φ , they are projected into the high-dimensional feature space to be linearly separated by the maximum-margin hyperplane. Circles intercepted by the dotted line are called ‘support vectors’. Adapted from¹³⁵.

1.3.3.3 Neural Networks

Neuron Networks (NN) are information processing algorithms that emulate the human nervous system and its mode of functioning. In a biological neuron, dendrites collect incoming signals from surrounding neurons and convey them as one to the cell body. If the signal surpasses a certain threshold, it will be transmitted across the axon to another neuron through the synapse. In analogy to its biological counterpart, an artificial neuron j receives input signals X_i , which are afterwards weighted, w_{ji} , and summed up to yield a net signal Net_j . All the weights w_{ji} of each neuron are coded together as a vector and they modulate the synapse strength, which is the essence of the learning process. Next, the net signal Net_j is converted by a transfer function, generally sigmoidal, and then transferred to other neurons (**Figure 12**). At a neural network, neurons are grouped into several layers. In each layer, neurons will simultaneously generate a set of outputs to be transferred for the next layers. NNs must contain, at least, a first layer for data input and a final output layer that transmits the calculated result, with optional hidden layers in between to process information. These networks can be applied for supervised learning by associating the input data vectors X_i with a specific outcome target. Once the network processes the signal, its output Out_i is compared with the known output to assess the predictive error δ . The model is then optimised to minimise the error by adapting the weights w_{ji} of the network¹⁰⁵.

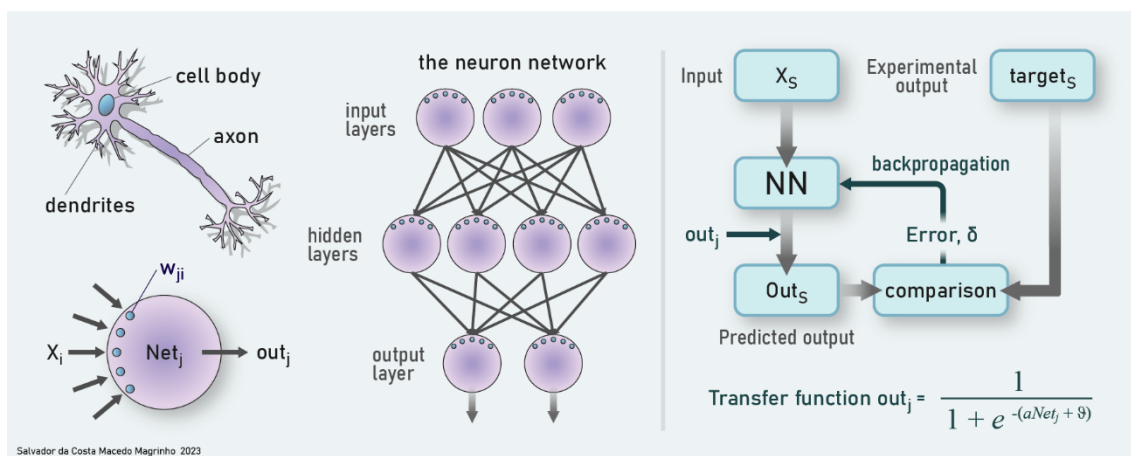


Figure 12. Neuron Network schematic in a supervised learning setting. An input vector is supplied to the neural network. Each neuron generates a weighted prediction which will be transformed by a transfer function into the final predicted output. The network output is compared with the target value to estimate an error that the model will try to iterately improve.

1.3.4 Molecular Docking

Currently, with the vast toolset that bioinformatics brought forward to the field of drug R&D, every CAAD effort will unavoidably, at some point, include molecular docking in its approach, as this computational instrument has become instrumental to the therapy development process^{139–141}.

In a brief definition, molecular docking is a computational modelling procedure used to predict how a protein interacts with small molecules at the atomic level. As the binding of ligands to a target deeply influences protein dynamics and ultimately biological function, docking represents an efficient method for describing the behaviour of ligands at a protein's binding site and elaborate on its possible effects. Docking provides two essential outcomes: A structural prediction on the most favourable binding mode(s), presenting the ligand's positioning, orientation and conformation at the protein surface (designated as “pose”) and an estimation for the free binding energy of the protein-ligand complex, which is a measure of affinity^{140,142}.

The standard docking protocol comprises several successive steps, as represented in **Figure 13** below. A further in-depth look onto each step will be offered later in this thesis by using the AutoDock Vina software¹⁴³ for the PMM2-CDG CAAD case study as an example on how to proceed.

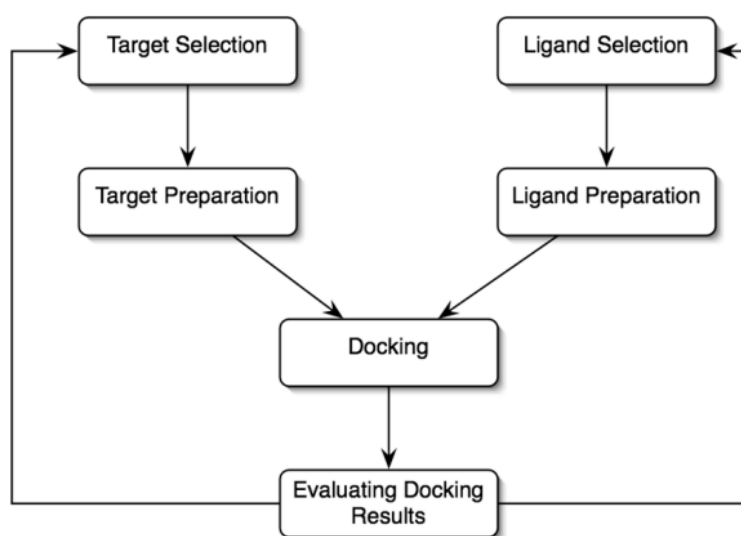


Figure 13. Typical docking workflow. This flowchart scheme shows the main steps in a docking protocol. The 3D structures for the target macromolecule and the ligand must first be chosen, and then each structure must be prepared in accordance with the requirements of the docking method being used. Following the docking, the results must be analysed, selecting the binding modes with the best scores.

When the concept of receptor-ligand binding was first formalised by Fischer in the end of the nineteenth century¹⁴⁴, it was based on a “key-and-lock” mechanism, meaning that both protein and ligand were imagined as rigid bodies. Almost a century later, in the 1960’s, Koshland suggested an “induced-fit” principle¹⁴⁵ where the conformation of the active site and ligand are continuously adjusted as a result of the established inter-molecular interactions. Similarly to both theories, whilst the earliest docking methods were only capable of performing rigid docking (in which both target and ligand are rigid), more recent and innovative approaches can explore a flexible depiction of both target and ligand to describe molecular interactions more precisely – this is called induced-fit docking. Unfortunately, though the latter strategy seems the most desirable, the processing capabilities of the normal computer at a laboratory are still subpar to the computing power required for induced-fit. For this reason, the most popular and effective method currently used is flexible docking, a compromise between the induced-fit concept and the present computational limitations, in which a flexible ligand binds to a rigid target¹⁴⁰.

Regardless of the method employed, identifying the binding site at the protein surface prior to docking is essential to assure an efficient performance. In cases where the binding site has not been previously described, it is possible to infer such information by comparing the target with available models for structurally related proteins. Binding pocket detection softwares may also be used, or even the actual docking software, as nowadays most of them contain an algorithm for this function. Without knowing where the binding site is, one can only performed what is referred as “blind docking”^{127,140}.

Docking requires two essential tools, related to its two outcomes: a search algorithm that explores the available conformational space to calculate the several binding poses for the ligand and a scoring function that encodes a force field (i.e., a computational method that describes the energy landscape, allowing to estimate the potential energy of a system of atoms¹⁴⁶) to rank the energetics of such poses accordingly. Generally, molecular docking utilizes simpler force fields as a trade-off to explore a broader area of the receptor's conformational space, although performance is intrinsically dependant on the softwares used, as these may exploit different search algorithms and scoring functions^{147,148}.

Search algorithms explore the ligand's binding modes by experimenting with its state variables: position coordinates, orientation (Euler angles, axis-angle) and conformation (torsion angles in rotatable bonds). When docking flexible ligands, it is crucial to use a structural representation format that encodes such variables, like the molecule's torsion tree, i.e., the description of rotatable and non-rotatable atomic bonds, which influences how the algorithm poses the ligand¹⁴⁹.

Search algorithms can be categorised as systematic or stochastic: A systematic method is deterministic and analyses the target's search space in defined intervals. In its turn, stochastic searches iteratively introduce random changes to state variables until reaching a user-defined criteria, this means that the outcome (i.e., binding modes) may vary from experiment to experiment, even if using the same setup. Search methods can be further classified by how broadly they explore the protein's conformational space, as either local or global.

The scoring function is used to calculate an estimation of the complex's free binding energy and can be classified as empirical, knowledge-based or force field-based: Force-field scoring functions determine binding energy by summing up non-covalent interactions (electrostatics and Van der Waals), H-bonding and solvation¹⁵⁰. Empirical scoring divides binding energy into various components that try to capture all the complex enthalpic and entropic contributions, such as binding entropy, H-bonds, ionic and hydrophobic interactions, and each component is multiplied by a coefficient before summing up the total score^{148,151}. Knowledge-based functions use structural analysis of other ligand-protein complexes to retrieve inter-atomic contact frequencies and distances. By operating under the principle that more favourable interactions occur more frequently, the function's score favours preferred contacts and penalises repulsive ones¹⁵².

Docking has the potential to cater to a range of applications in drug discovery studies, including as a final stage in a virtual drug screening, with the scope of generating a three-dimensional structural hypothesis of how a set of screened ligands is predicted to interact with the target of pharmacological interest. Truthfully, this is the intended use that was given to docking in this work (*vide* **section 1.5**).

1.4 Biological studies applied to PMM2-CDG drug development

Once a drug repositioning hit has been identified by an *in silico* strategy, its efficacy on the new protein target must be experimentally evaluated by conducting *in vitro* studies with appropriate disease models¹³³. Understanding what sort of biological systems best describe the behaviour of an organism in a disease setting is of key relevance, as *in vitro* models constitute the starting point in clinical research to provide insight into disease progression and therapeutic effects of lead compounds. Different types of models can be developed at different scales, being based on proteins, cells or tissue systems. Regardless of the model, it must always strive to effectively represent the underlying aspects of a disease (affected biological effectors, abnormal pathways, etc...) and, even when isolated from their natural biochemical environment, it should provide precise and complete predictions on the effects that therapy approaches may have on the organism¹⁵³.

1.4.1 Fibroblasts as disease study models for PMM2-CDG

The biological studies developed in this project were conducted using primary skin fibroblasts, i.e., cells collected from live tissue by means of biopsy, as an *in vitro* disease model. *Why fibroblasts?* – These mesenchymal cells are the most common amongst animal connective tissues, presenting a fusiform morphology. They play a role in connective tissue homeostasis by preserving its structural integrity and are crucial for various physiological processes in health and disease^{154–156}.

Tissue cells derived from patients are regarded as ideal sources to explore pathogenic phenotypes¹⁵⁷. In fact, the use of human fibroblasts to study the underlying mechanisms of a disease, its manifestations and even response to therapies has been documented for a large spectrum of diseases, being a well-accepted practice in cell biology¹⁵⁸. The reasoning why fibroblasts represent such a useful tool for disease studies is related to their abundance in the human body and the fact that they can be easily isolated from skin biopsies, requiring a minimally invasive harvesting procedure, with almost no discomfort to the donor. Also, with current culturing methodologies, fibroblasts can be grown and manipulated *in vitro* in a straightforward manner for a large period of time¹⁵⁵. Other relevant aspect is that fibroblasts are considered phenotypically stable¹⁵⁹ and, in a culturing setting, they are reported to maintain the patient's original genetic background, which is important when exploring genetic disorders¹⁶⁰. These remarks make patient-derived fibroblasts a very suitable model to study multi-system, metabolic diseases like CDGs. As so, fibroblast models have been reported by various research groups studying different CDGs, including PMM2-CDG¹⁶¹. The general availability of fibroblasts in commercial cell repositories and their relative affordability should also be accounted as another advantage.

1.4.2 Molecular biomarkers to study PMM2-CDG *in vitro* models

As mentioned before in **Section 1.2.4**, the availability of consistent biomarkers for PMM2-CDG disease progression is very limited. Currently, experimentation in therapy development studies using *in vitro* disease models rely on a relatively small handful of options, including monitoring PMM2 enzyme residual activity, which is explored more in depth in **Section 1.4.3** below. An additional important route pursues the analysis of expression/turnover patterns of specific protein markers that are known to be altered in the pathophysiological state of the disease, such as PMM2's, or other proteins affected by the defects in the glycosylation process, like ICAM-1:

The intercellular adhesion molecule 1 or ICAM-1, also known as CD54, is a ~90 kDa membrane-bound glycoprotein, member of the immunoglobulin superfamily, with a crucial role in intercellular signalling for numerous immune and inflammatory processes, including lymphocyte activation and leukocyte trafficking. ICAM-1 is reported to be constitutively expressed in low amounts in various cell types including fibroblasts, although pro-inflammatory stimuli like TNF- α , IL-1 β , IFN- γ and other cytokines are known to exert upregulation of its expression^{162–165}. Human ICAM-1 can be heavily glycosylated, accounting for eight *N*-glycosylation sites at its polypeptide sequence¹⁶⁴. The existence of different patterns of glycosylation, giving rise to different ICAM-1 glycoforms, has been characterised before, entailing that different degrees of glycosylation may be related to different specific functions^{162,166}.

The full-length transmembrane isoform of ICAM-1 is composed of five Ig domains and it can form a homodimer, though other quaternary structure arrangements are also documented (*vide* **Figure 14**). Furthermore, an unbound soluble form, sICAM-1, has also been detected, being proposed to have signalling properties in several cell types^{163,164}.

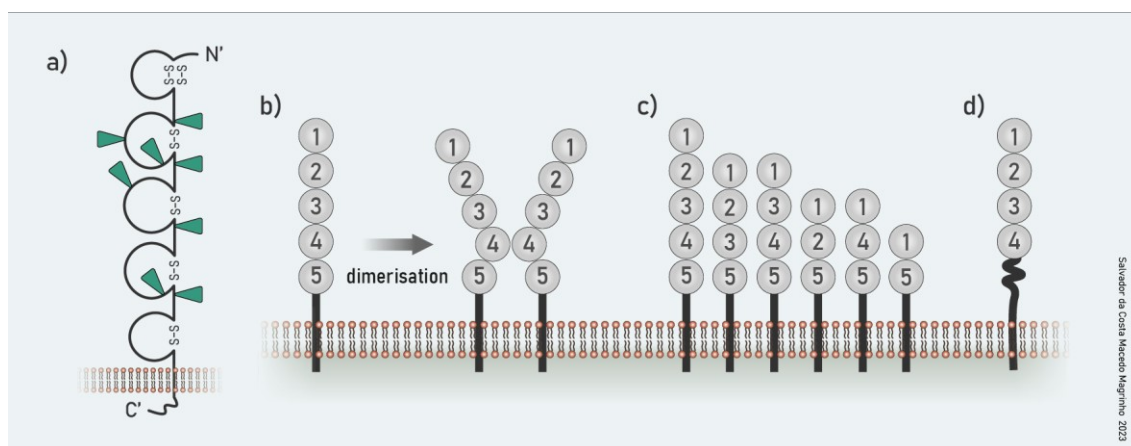


Figure 14. Schematic depiction of different ICAM-1 isoforms. **a)** In this simplified scheme of ICAM-1 each circle represents one of its Ig-like domains. Green triangles pinpoint the 8 glycosylation sites present in human ICAM-1. **b)** Full-length ICAM-1 isoform. This isoform is able to form a homodimer through interactions at Ig domain 4. **c)** Different ICAM-1 isoforms have been identified to be expressed at the membrane in endothelial and other cell types. **d)** Also, alternative splicing in exon 6 leads

to truncation of Ig domain 5 in the full-length isoform under inflammatory conditions. The number of each domain is based on their order in the full-length five domain isoform. Figure inspired by ^{162,164}.

Alterations in the glycosylation profile of certain proteins can be considered as suitable disease markers to be exploited⁶ and, as a matter of fact, due to the underlying impact that glycosylation ultimately has on ICAM-1 structure and biological function, previous studies have explored its application for evaluating cellular hypoglycosylation. ICAM-1 expression was examined in PMM2-CDG patient-derived fibroblast lines through western blotting experiments, reporting a significant decrease, or even total absence, compared to wild type controls, since the hypoglycosylated protein fails to localize at the membrane. In the same study, a therapeutic attempt involving *PMM2* gene complementation in these PMM2-deficient cell lines partially improved ICAM-1 glycosylation patterns, as inferred by the detected increase in its levels at the plasma membrane^{167,168}. This observation indicates that ICAM-1 can, in fact, be employed as a purposeful biomarker to further explore the therapeutic effects of potential drug leads in *in vitro* PMM2-CDG cellular models, as part of the intended drug discovery approach.

1.4.2.1 Western Blotting

Western blotting (WB) is a gold-standard technique widely employed for identifying and characterising specific proteins in complex biological samples. Its well-established methodology involves a size-based separation of proteins by polyacrylamide gel electrophoresis, followed by their blotting, i.e., their electrophoretic transfer to a membrane support where the protein(s) of interest will later undergo a selective immunodetection step by incubation with a specific antibody, as shown in **Figure 15**. As WB is centred on the high specificity of antibody-antigen interactions, it is a very powerful routine technique to detect the proteins under analysis within a complex mixture and monitor alterations in their expression patterns in a qualitative or even semi-quantitative manner. This makes WB a valuable resource in biomedical research and diagnostics for detecting specific proteins utilised as disease biomarkers, as it is demonstrated by its countless clinical applications described throughout literature^{169–172}. Detecting PMM2 and ICAM-1 through western blotting has been reported in past studies^{75,87,167,173} and it comprises an adequate tool to be explored in the present Master's project.

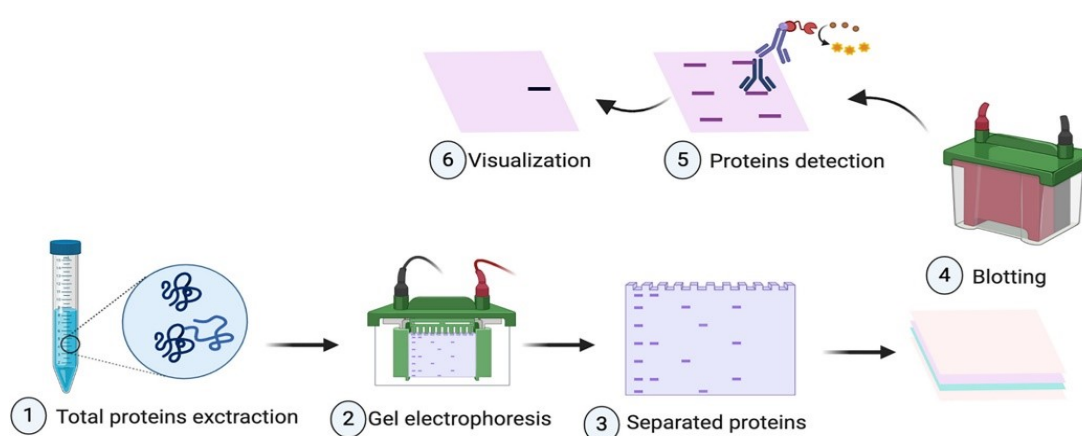


Figure 15. The typical western blotting workflow. Adapted from¹⁷⁰.

1.4.3 Enzymatic assays to measure PMM2 activity

When in 1995 Van Schaftingen & Jaeken¹⁷⁴ established the deficiency in hPMM2 enzymatic activity to be the underlying cause for PMM2-CDG (then known as carbohydrate-deficient glycoprotein syndrome type I), they developed a kinetic assay to effectively measure the catalytic activity of PMM2 in purified protein samples, as well as in extracts from patient-derived cell lines and tissues, including fibroblasts, leucocytes and liver biopsies. Different alterations to the original protocol have been gradually added by several research groups, mainly regarding the amount of protein extract employed in the assay and reagent concentrations^{25,59,175}, although the concept of the assay remains unchanged, being based on coupling the metabolic pathway in which PMM2 partakes (*vide* **Figure 3**) with NADP⁺ reduction:

By supplying Man 1-P as the enzyme's substrate, along with the Glc 1,6-BP activator and MgCl₂ to provide the required ion co-factors, the reaction yields Man 6-P, which will be converted from Man 6-P to fructose 6-phosphate (Fru 6-P), and then from Fru 6-P to Glc 6-P by ancillary enzymes phosphomannoisomerase (PMI) and phosphoglucosomerase (PGI), respectively. Next, glucose 6-phosphate dehydrogenase (G6PDH) catalyses NADP⁺ reduction to NADPH using the Glc 6-P metabolite and forming 6-phosphogluconate. The NADP⁺/NADPH conversion rate can be easily measured by UV-Vis spectrophotometry as the reduced NADPH species exhibits a distinct peak at 340 nm¹⁷⁶. Measurements are performed at a temperature around 30 to 37°C, during anywhere from 45 to 60 min. Fluorescence spectroscopy can be alternatively conducted as a way of improving assay sensitivity and reproducibility by measuring the NADPH production by fluorescence at 540/590 nm^{59,177}. Performing the assay in a 96-well microplate support may also bear some advantages over the traditional cuvette cells, as it renders the method more suited for high-throughput strategies.

The assay must be carried out in two setups, with and without the substrate, and the difference between the two values is described as the enzyme activity. Conventionally, 1 U of PMM2 enzyme is the activity that corresponds to the formation of 1 μmol NADPH/min under the assay conditions.

It is worth remembering that PMM2 has, to a lesser extent, a phosphoglucomutase activity (*vide* **Section 1.2.1**) though which it is also possible to examine PMM2 catalytic activity, performing a simpler assay with a reduced need for ancillary enzymes. Whilst this may seem a good option for purified protein samples, being reported to have produced feasible results in past works⁸⁸, it is not recommended for whole cell extracts.

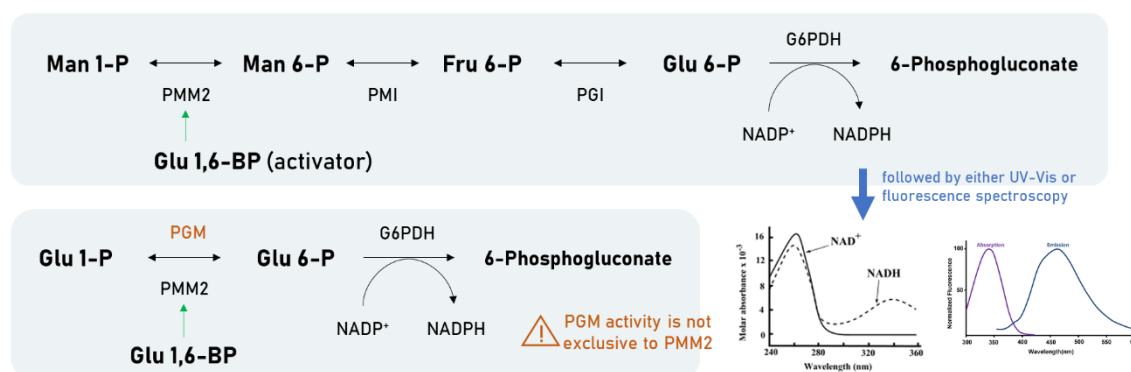


Figure 16. PMM2 phosphomannomutase kinetic assay pathway. Enzyme activity is measured, either by UV-Vis or fluorescence spectroscopy, by coupling the NADP⁺ reduction reaction catalysed by G6PDH to the pathway that produces its Glu 6-P substrate. PMI and PGI are required as ancillary enzymes. An alternative assay can be performed in purified protein samples to measure the lesser PMM2 phosphoglucomutase activity, reducing the need for ancillary enzymes.

1.5 Goals for this thesis

With the end goal of contributing to the recent efforts on developing new effective therapies for PMM2-CDG, the work described in this Master's thesis brings forward the possibility of repurposing already approved drugs as novel, first-generation pharmacological chaperones to restore PMM2 residual activity in disease-associated variants observed in patients.

This strategy was conceptually allied to a “dry lab” component, making use of innovative cheminformatics techniques such as artificial intelligence and drug screening methodologies to create a computational procedure that employs, in an initial phase, two machine learning QSAR models to predict the activity of a certain compound towards human PMM2:

The first model is a QSAR classification model meant to predict whether if a certain submitted input molecule would be able to establish a protein-ligand interaction with PMM2. By using a training set consisting of compounds that have been biologically tested towards PMM2 interaction in screening bioassays, the model's learning algorithm will be capable of generalising which chemical properties are associated with the ability to interact with the enzyme. Hence, when a new query molecule is submitted to the model, it will generate a categorical prediction which must be either a “positive” output, meaning

that such compound is predicted to be a ligand for PMM2, or “negative”, indicating that the molecule is not expected to be biologically active for the protein.

The second model is a QSAR regression strategy to estimate a theoretical half maximal inhibitory concentration value (IC_{50}) of a submitted compound towards the enzyme. IC_{50} is classically regarded as a widely informative measure of a compound’s efficacy on establishing an interaction with a protein target, exerting an inhibitory effect^{178,179}. As so, this model provides insight on how appropriate a compound is for chaperoning, since pharmaco-chaperones exhibit a mode of action very similar to that of a reversible inhibitor.

Subsequently, these QSAR models will be used to conduct a virtual drug screening to explore potential compounds of interest that present the desired pharmacological profile within a dataset of approved drugs retrieved from online chemical databases. Selected hits will then undergo molecular docking experiments to further characterise their complex drug-target interactions.

To conclude, experimental validation of the proposed CAAD approach will be pursued through the acquisition of some of the compounds identified as promising leads to carry out biological activity studies in *in vitro* disease models using PMM2-CDG patient-derived fibroblast cell lines, in a stage regarded as a final “wet lab” component. This phase includes performing a thorough characterisation of the cellular model and important disease markers, such as PMM2 and ICAM-1, by different techniques, including enzyme activity assays, to assess the therapeutic properties of hit compounds.

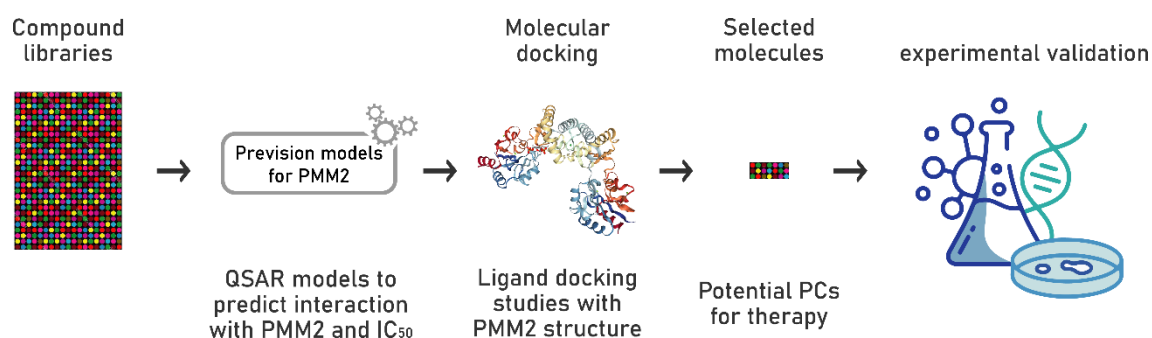


Figure 17. General workflow of the CADD approach

MATERIALS & METHODS

2.1 Cheminformatics component

2.1.1 Model 1: QSAR classification model to predict ligand interactions with PMM2

The first QSAR classification model was designed to predict compounds from between two activity categories: Compounds with potential to establish protein-ligand interactions with hPMM2, dubbed as “positive compounds”, and compounds not capable of interacting with the enzyme, dubbed as “negative compounds”. The concept for this model was previously explained in **Section 1.5**. A depiction of the general workflow that was used for constructing this QSAR model is represented below in **Figure 18**.

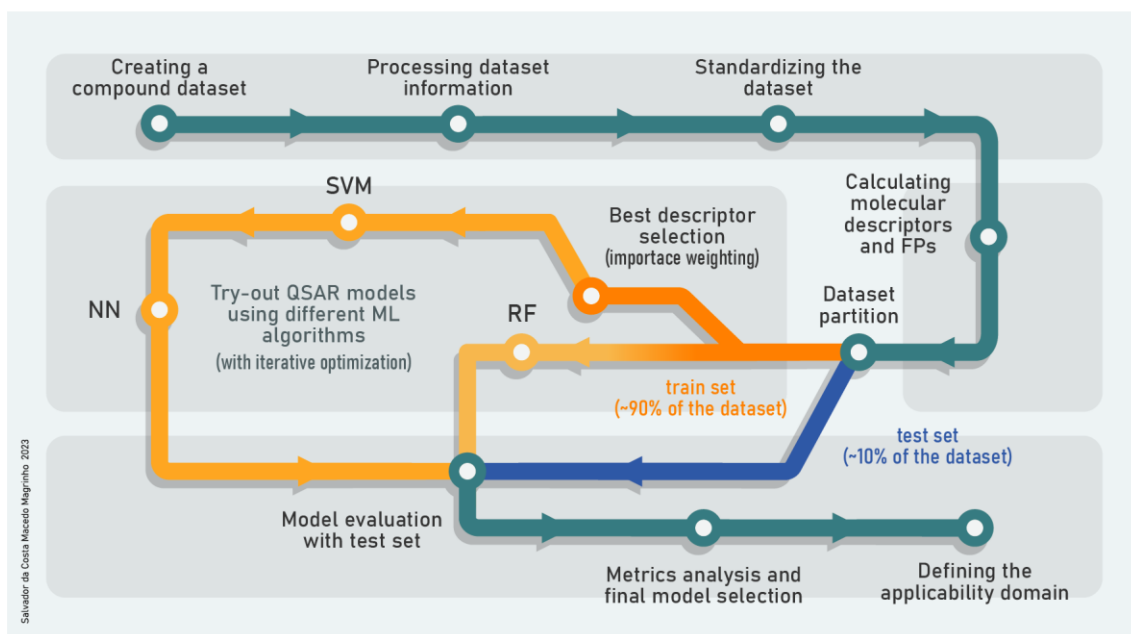


Figure 18. Subway map representing the general workflow employed to build the predictive QSAR models.

2.1.1.1 Building a compound dataset for model construction

For dataset construction, an *on-line* literature search was performed to extract all available published articles where a high-throughput screening (HTS) assay had been conducted to evaluate the interaction of different compounds with human PMM2. The article search was conducted using scientific databases such as the Web of Science¹⁸⁰ at <https://clarivate.com/webofsciencegroup/solutions/web-of-science/>, PubChem¹⁸¹ at <https://pubchem.ncbi.nlm.nih.gov/>, ChEMBL¹⁸² at <https://ebi.ac.uk/chembl/> and ZINC database¹⁸³ at <https://zinc.docking.org/>. Only academic papers dated from until June 2021 were considered. For model construction purposes, retrieved compounds were manually labelled with either one of two categories: “positive” or “negative”, based on a comprehensive reading of the assays’ associated literature. Relevant information from all compounds was compiled in an Excel file for the general dataset, including their original database identifiers; SMILES notation; molecular weight and experimental IC_{50} values, which were registered as logarithm by applying the function depicted in **Equation 8**, which was done to simplify model construction and analysis.

$$pIC_{50} = -\log_{10}(IC_{50})$$

8

The dataset’s original SMILES were standardized in Chemaxon Standardizer (JChem 15.4.6, 2015, Chemaxon Ltd., Budapest, Hungary) using an author-defined standardization setting list (*vide* **Table 19 in Appendix Section A.2**) to ensure equal representation for all compounds. As a considerable number of duplicates was found, a duplicate removal search within the dataset was performed using Excel’s

built-in *Remove Duplicates* function for the SMILES notation. In some cases, removal had to be done manually, requiring careful analysis: Whenever duplicates reported two distinct IC₅₀ values for the same compound, the lowest defined value was considered. From the final dataset that contained 10182 unique compounds, 60 were labelled as “positive” and 10122 “negative” (*vide Excel Appendix*).

2.1.1.2 Calculating Molecular Descriptors

Standardized SMILES were converted to SDF format files in ChemAxon Standardizer and submitted to PaDEL-Descriptor (Pharmaceutical Data Exploration Laboratory, Singapore) version 2.21¹¹⁶ (<http://www.yapcwsoft.com/dd/padeldescriptor/>, last accessed March 2023) to calculate molecular descriptors for the dataset. For this study, seven groups of FPs, either fragment or circular FPs, were calculated for each compound in the dataset: *Fingerprinter*; *ExtendedFingerprinter*; *EStateFingerprinter*; *GraphOnlyFingerprinter*; *MACCSFingerprinter*; *PubChemFingerprinter* and *SubstructureFingerprinter*. Similarly, all 1D&2D descriptors available in the software (55 descriptor subtypes amounting to 1444 individual attributes) (see the description made in **Section 3.1.2**) were also calculated for the dataset.

2.1.1.3 Model Building

2.1.1.3.1 Partitioning the dataset

For dataset partition, 9151 compounds were randomly selected from the entire input dataset for the training set, containing 53 “positives” and 9098 “negatives”, and 1021 compounds for the test set, with 7 “positives” and 1014 “negatives”, following a distribution ratio with similar order of magnitude amongst both sets (positive-negative ratio was 1:172 and 1:146 for the training and testing set, respectively).

2.1.1.3.2 Trial models to select input descriptors and optimise algorithm parameters

An essential previous step to model building was to select which types of Fingerprints/molecular descriptors would be the most suitable to represent the input dataset for algorithm calculations, as well as the best operating parameters for the ML algorithm. To do so, several preliminary trial models were constructed using the established training set employing a random forest (RF) algorithm. Model analysis was carried out by examining the prevision metrics obtained such as TP; FP; TN; FN; SE; SP; overall predictability (Q) and the Matthew’s Correlation Coefficient (MCC) (*vide Section 1.3.2*), the latter two being particularly relevant.

Firstly, to set the best parameters for the RF algorithm, trial models were made using the program WEKA (Waikato Environment for Knowledge Analysis) version 3.8.5¹⁸⁴. In WEKA’s *Explorer*, the

meta Cost Sensitive Classifier was selected, with RF set as the classifier algorithm. The Molecular Access System Fingerprints (MACCS) generated for the training set were chosen for this task due to their simplicity and small number of attributes (MACCS compiles 166 individual descriptor attributes), allowing faster calculations. Parameters under optimisation included the RF algorithm's *mtry* (number of features), *no. of iterations* (number of trees used in the forest ensemble), and the RF classifier's *cost matrix*, which balances the weight of each categorical class for model calculations. All other parameters were kept with their default values as defined in WEKA. *Out-of-bag* estimations for the training set were used and a *10-fold cross validation* testing option with the training set was employed to evaluate overall model performance.

Secondly, having set the best operating parameters, RF trial models were made for each one of the seven Fingerprint groups, plus the 1D&2D descriptors group, to analyse which one yielded the best predictive performance. For each group, model calculations were performed using a third and the square root of the total number of attributes as the algorithm's *mtry*.

The best model utilizing each descriptor type (fragment Fingerprint, circular Fingerprint and 1D&2D descriptors) was chosen to proceed to the next stage, which consisted of calculating individual attribute importance for each descriptor, i. e., how much every single attribute from each descriptor group contributes to the overall predictive capacity of the QSAR model. This was done with the R program¹⁸⁵ version 3.6.3. "Holding the Windstock" using the randomForest library¹⁸⁶. The same running script was used for the three groups with the required minimal changes (attribute tag names and *mtry* value) and a *class weight* classifier of 1:10 (negative:positive balance) was applied to balance the model. After analysing the resulting outcomes, the best individual attributes for each group of descriptors were listed by sorting them according to the Random Forest library's built-in *Mean Decrease Accuracy Value* metric.

2.1.1.3.3 Descriptor attribute selection

Having the individual attributes of each of the three best-performing descriptors organized by order of importance, a model optimisation step was performed in WEKA as done in the previous section to see how the model's prediction capability varies with the number of attributes used in its construction. For each descriptor, the model was subsequently built using the 250 best (most importance weighted) attributes; then the 200 best; the 150 best; the 100 best; and the 50 best.

2.1.1.3.4 Building Random Forest QSAR models

Different QSAR classification models were attempted, using different ML algorithms for calculation, namely Random Forests, Support Vector Machines and Neural Networks, as explained earlier.

RF models were trained in WEKA employing the same methodology used in the previous optimisation steps. Models were built for the two best-performing Fingerprints and the 1D&2D descriptor groups, with no attribute selection and applying the earlier pre-defined operating parameters. As, in general, RF

models are not affected by noisy descriptors¹⁸⁷ the RF building methodology did not contemplate descriptor attribute selection, i.e., RF models were built using the whole Fingerprint/descriptor group, unlike the subsequent SVM and NN models.

2.1.1.3.5 Building Support Vector Machine QSAR models

SVM models were also built in WEKA, employing two different possible approaches for class balancing: In a first strategy at the WEKA's *Explorer*, the *LibSVM* classifier algorithm was employed for model building, selecting the *C-SVM classification* SVM type, a radial basis function as the kernel function, and algorithm normalisation was turned on. Optimisation rounds were performed by varying the classifier's *cost* and *weight* parameters. The other strategy that was experimented used the *meta Cost Sensitive Classifier (mCSC)* as before for RF, now with the *LibSVM* algorithm, and fine-tuning of the model was performed by adjusting the class balance *cost matrix* of the *mCSC* classifier menu instead of using the SVM's own parameters. In both approaches, the SVM algorithm's ϵ value was constantly kept at 0.0001 and the *loss value* parameter 0.01. The same *10-fold cross validation* testing option as before was used for evaluation.

2.1.1.3.6 Building Neural Network QSAR models

Neural network algorithms were trained and applied using the software library Keras¹⁸⁸ version 2.2.5, based on Tensorflow numerical backend engine¹⁸⁹. Using a specific script to build the network, optimisation of the NN architecture was manually conducted by adjusting the algorithm's hyper-parameter settings, namely the *no. of hidden layers* in the network; the *no. of neurons*; the *optimiser*; and the *class weight*. Other parameters such as the *initializer*, *learning rate*, *decay* and *epochs* were not explored, remaining with default values. Model building was iteratively tested by 10-fold cross validation as well.

2.1.1.3.7 Model evaluation with test set

For each one of the three ML strategies explored, the models that indicated the best predictive performances according to the defined prevision metrics calculated from the 10-fold cross-validation process were evaluated using the established test set to verify prediction accuracy and overall performance with an external set of compounds. For the models trained in the WEKA program, the built-in testing option for *supplied test set* was chosen to validate the model, loading the test set input descriptors. For the NN algorithm in Keras, the script contained a specific line to perform test set running.

2.1.1.4 Setting the model's applicability domain

To describe the domain of application for this QSAR model, certain selected compounds of interest, regarded as experimentally-verified "positive" leads, were defined as templates (*vide* **Section 3.1.2.6**)

for calculating the chemical structural similarity shared amongst such templates and the model’s input dataset. The Tanimoto similarity coefficient, TC, (baptised after the Japanese mathematician Takeo Tanimoto, who developed it in the 1950s) was used, as it generates a similarity score between its two arguments: a provided template (or target) compound structure and a query set of molecules. As the coefficient generates a value between 0 and 1, the higher this threshold is the closer the structural identity between query and target structures is. TC can be calculated through the **Equation 9** presented below¹⁹⁰.

$$TC_{AB} = \frac{N_{AB}}{N_A + N_B - N_{AB}}$$

9

The TC is calculated by generating a *hashed binary Fingerprint* for the query and target molecule (A and B), with N_A signifying the sum the number of "on" features ("1" bits) in structure A, and N_B the ones in structure B. N_{AB} depicts the total "on" features common to both.

The Tanimoto coefficient may be calculated by considering as input either two- or three-dimensional structure representations. Generally, 2D and 3D TC scores yield similar results, although 2D similarity measures may neglect important structural features leading to over- or underestimation in similarity metrics¹⁹¹. By including calculation for both 2D and 3D scores, a more comprehensive and unbiased analysis will be achieved.

To obtain the two-dimensional structural Tanimoto coefficient, the RDKit software¹⁹² was used by running a script that required providing the template molecules in SMILES format, standardized identically as earlier described, and the query set (the model’s train and test sets), also as SMILES. The software’s own RDKFingerprints were generated and used for this calculation.

In order to calculate the three-dimensional Tanimoto coefficient, firstly the dataset compounds’ SMILES were converted to three-dimensional molecular conformers. This was done using a conformer optimisation script in RDKit that converts the input SMILES data into SDF files containing, for each submitted molecule, a specified number of the lowest-energy conformers generated. In this case, the script was set to generate the five best conformers for each compound. Then, using yet another script for the 3D Tanimoto coefficient in RDKit, the SDF files were used to conduct the calculation of the similarity between the template conformers and the query set. This script generates two different similarity scores based on the Fingerprint type that was used for calculation. In this experiment, E3FP Fingerprints and Morgan Fingerprints were used. Both these Fingerprints are binary rule-based FPs, containing 1024 individual features. Both E3FP and Morgan’s are circular 3D FPs¹⁹³.

2.1.2 Model 2: QSAR regression model to predict theoretical IC50 value

As mentioned in **Section 1.5**, this QSAR regression model was trained with the intent to predict a theoretical pIC_{50} value towards hPMM2 inhibition for every submitted compound.

2.1.2.1 Building a compound dataset for model construction

To build the dataset used for constructing this QSAR regression model, the dataset from the previous Model 1 was repurposed by only considering, from the total 10182 compounds, the 56 compounds that reported in literature an experimental measure for the IC_{50} value towards hPMM2. The standardized SMILES of the entire dataset were converted to SDF format with ChemAxon Standardizer in order to submit the dataset for structural optimisation.

The dataset SDF file was submitted to JChem CXCALC tool (JChem 15.4.6, 2015, Chemaxon Ltd., Budapest, Hungary) to set the best three-dimensional structure conformers for the entire dataset using CXCALC's *lowestenergyconformer* calculation script (<https://docs.chemaxon.com/display/docs/cxcalc-calculator-functions>, lastly accessed in March 2023), that employed the *MMFF94* molecular force field¹⁹⁴, without prehydrogenize and by considering a very strict conformer generation optimisation limit for different environments. These and other command parameters can be checked at the developer's webpage. This action produced a three-dimensional SDF file containing the lowest energy conformer for each molecule at the dataset.

2.1.2.2 Calculating Molecular Descriptors

1D&2D descriptors and Fingerprints were calculated in PaDEL-Descriptor for the entire dataset, in the same manner as for the QSAR classification strategy. As 3D descriptors were also required for this model, the optimised 3D conformers in SDF file were also submitted to the program to calculate seven sets of distinct 3D descriptors: *Auto-correlation 3D*; *CPSA* (Charged Partial Surface Area descriptors); *Gravitational Index*; *Length Over Breadth*; *Moment of Inertia*; *Petitjean Shape Index* and *WHIM* (Weighted Holistic Invariant Molecular descriptors). Radial Distribution Function (RDF) 3D descriptors were also calculated for the dataset. RDF descriptors describe the probability for one atom, amongst A atoms, being found in a spherical volume with radius r ¹¹². Calculation of such descriptors requires the partial atomic charge of each molecule in the dataset to be known beforehand. Partial atomic charges can be regarded as a simpler depiction of the electronic configuration of a molecule, as they are closely associated to polarizability¹⁹⁵. To calculate partial atomic charges for the entire dataset, the JChem CXCALC tool was used, employing the *myCHARGE* calculator function script (<https://docs.chemaxon.com/display/docs/cxcalc-calculator-functions>, lastly accessed in March 2023). Afterwards, the partial atomic charges were used to generate the RDF descriptors for the dataset in a Java *in-house* application, utilizing the **Equation 10** depicted below.

$$g(R) = f \times \sum_{i=1}^{A-1} \sum_{j=i+1}^A w_i \times w_j \times e^{-\beta \times (R-r_{ij})^2}$$

10

Where f is a scaling factor, w is the calculated partial atomic charge for atoms i to j and A is the number of atoms. The exponential term describes r_{ij} , the distance between the i^{th} and j^{th} atoms, which was set to 12.8 Å, the radius R of the spherical volume and a constant temperature parameter β that accounts for atomic vibration¹¹².

2.1.2.3 Model building

2.1.2.3.1 Partitioning the dataset

For dataset partition, 51 compounds were selected from the entire input dataset for the training set with reported pIC₅₀ values for hPMM2 ranging from 6.172 to 3.547, and five compounds for the test set, with pIC₅₀ values between 4.054 and 4.257.

2.1.2.3.2 Trial models to select input descriptors and optimise algorithm parameters

Following the same methodology as the earlier QSAR classification approach, it was conducted a preliminary training phase to select which types of Fingerprints/molecular descriptors best describe the input dataset for this model, as well as the best operating parameters for the ML algorithms. This involved training Random Forests in the WEKA program, where different trial models were built using each one of the seven Fingerprint sets generated for the dataset, the whole 1D&2D descriptor set, and each one of the eight 3D descriptors sets available. The performance of regression models was assessed using metrics like the Correlation Coefficient (R^2), the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE), as previously described in **Section 1.3.2**, plus the Relative Absolute Error (RAE) and the Root Relative Square Error (RRSE).

At this stage, the best-performing descriptor from the Fingerprint and 3D sets was chosen, in addition to the whole 1D&2D set, to further carry out model optimisation. Next, descriptors were combined to train different trial models using bundles of Fingerprints+3D descriptors and 1D&2D+3D descriptors as input for the algorithm. This allowed to define which bundles yielded the best predictive performances. Descriptor attribute importance was again weighted in the RF algorithm in R software as before.

2.1.2.3.3 Descriptor attribute selection

After weighting individual descriptor attribute importance for each one of the defined bundles. Algorithm optimisation was carried out as before, training RF models for each bundle with the 250 best attributes and so on, reducing the amount of attributes.

2.1.2.3.4 Building Random Forest QSAR models

RF models were trained identically as for Model 1, optimising for each descriptor bundle the *mtry* and *no. of iterations* that provided the best prevision capacity according to the evaluation metrics.

2.1.2.3.5 Building Support Vector Machine QSAR models

SVM regression models were also built in the WEKA program by selecting at the Weka's *Explorer* the *LibSVM* classifier algorithm operating with the *epsilon-SVR regression* SVM type. The SVM's ϵ and *loss* parameters were kept constant as 1×10^{-4} and 1×10^{-2} respectively. Several attempts to achieve the best possible performance were made by optimising algorithm's parameters such as *input normalisation* (either with or without it), the *cost* and the kernel function (tested options included linear, radial basis and polynomial functions). Non-mentioned parameters were kept with the default setting.

2.1.2.3.6 Building Neural Network QSAR models

Neural network models were again built in Keras as already detailed. Adjusted hyper-parameters included the *no. of hidden layers* in the network, the *no. of neurons*, and the *optimiser*.

2.1.2.3.7 Model evaluation with test set

Assessment of the regression models with the established testing set and conducted equally to Model 1 as formerly described.

2.1.3 Virtual screening of approved drugs

To perform the virtual screen of approved drugs towards hPMM2 as proposed in **Section 1.5**, a systematic search in the ZINC database was conducted, as illustrated by **Figure 19**. In this search, the results (drugs) were obtained by filtering the searched substances with the *world* subset from the *Bioactive and Drug* settings, meaning that, according to the database's description, only compounds approved in major jurisdictions, including the FDA, i.e., DrugBank approved, are presented. Another filter was added: the *for sale* subset from the *Availability* settings, allowed to filter the database for all substances that are for sale, including in-stock, on demand and boutique. The final search set was downloaded in SMILES format and standardization was performed in Chemaxon Standardizer, as previously stated, followed by

a similar duplicate search to remove matching entries. From the originally extracted 4914 drugs in the dataset, only 3306 unique molecules were kept. In the same way as the earlier QSAR approaches, the dataset's SMILES file was converted to SDF format to calculate the required Fingerprint/descriptors. The drug dataset was then submitted to the classification QSAR model 1 to attain a protein-ligand hPMM2 interaction prediction.

In parallel to prediction calculation, the drug dataset was analysed to understand if all contained molecules fell within the QSAR models' applicability domain. This task was performed by employing the two- and three-dimensional Tanimoto coefficients as a chemical similarity filter to compare the drug dataset with the defined template compounds for the model. Tanimoto calculations were carried out identically as before. For a certain drug from the virtual screen set to successfully "pass" the Tanimoto similarity filter, and hence being considered in the CAAD strategy, it had to achieve, for a majority of the template compounds, a similarity score equal or higher than that predefined by the input training set employed to build the QSAR models.

The drugs that passed similarity analysis for the majority of templates were considered for submission to the regression QSAR model 2 alongside the drugs predicted as "positives" in model 1, and an estimation to the theoretical pIC_{50} value for each ligand was predicted. To proceed for further examination by molecular docking studies, certain drug compounds from the screen dataset were selected, plus the template molecules.

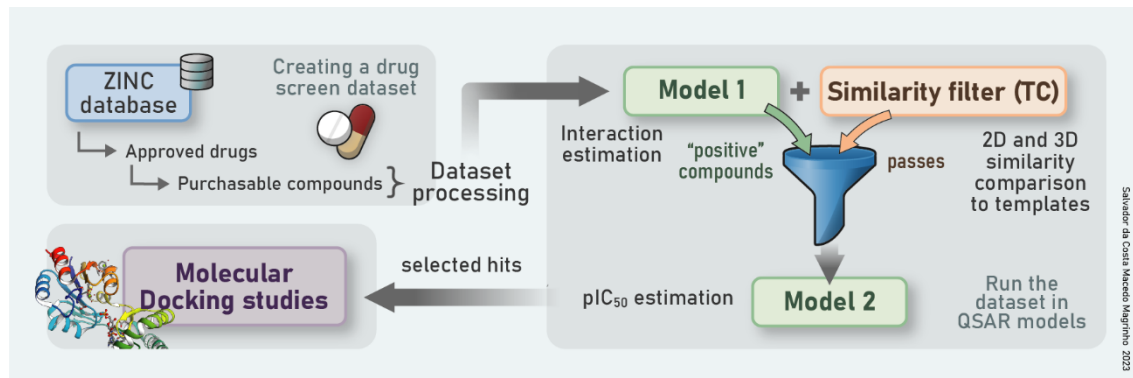


Figure 19. Simplified scheme of the workflow employed for the virtual drug screening approach.

2.1.4 Molecular Docking studies

Molecular docking experiments were carried out in the AutoDock Vina software version 1.12¹⁹⁶ using the recent crystal structural model of human PMM2, deposited in RCSB's Protein Database¹⁹⁷ (<https://www.rcsb.org/>) under the entry ID 7O1B⁴⁷ (wild-type hPMM2 co-crystallized with its activator Glc 1,6-BP). The model's PDB file was extracted and processed in AutoDockTools version 1.5.6¹⁴⁹ (<http://mgltools.scripps.edu/>) to prepare a PDBQT format for docking. This format is an extension from the original PBD file that includes additional data on atomic partial charges, atom types and torsional

degrees of freedom, which is required since AutoDock is parameterized to operate using protein and ligand models that include polar hydrogen atoms.

Protein receptor treatment included a series of pre-processing steps using the toolset available in AutoDock Tools' *Edit* and *Grid* menus to:

- Remove crystalline water molecules;
- Remove ligands prevenient from crystallography (crystal structures often contain small molecules with no biological function in the protein, such as glycerol or polyethylene glycol that are used as cryo-protectants and additives in the diffraction experiment¹⁹⁸);
- Add hydrogen atoms to the receptor model;
- Proceed to generate the PDBQT file by deleting non-polar hydrogens (hydrogens from the protein's aliphatic backbone) and merge their charges with the carbon atoms; compute partial charges and assigning atom types.

The protein's co-factor magnesium and chlorine ions were kept in the model.

Molecules submitted to docking were converted from their standardized SDF files to 3D-optimised SDF files using the *lowestenergyconformer* calculator function script from JChem CXCALC, just as done previously. The optimised three-dimensional files were then converted to PDBQT format using the OpenBabel software version 3.1.1¹⁹⁹, which contains important data on rotatable bonds for the torsion tree that describes the degree of flexibility each ligand possesses.

The docking experiment was conducted in AutoDock Vina using a search grid box with dimensions x = 40.00; y = 40.00; z = 40.00, centred near PMM2's active site at the interface region between the cap and core domains. For its search algorithm, AutoDock Vina used a stochastic global optimisation approach that operates using a genetic algorithm called Iterated Local Search global optimiser¹⁹⁶.

As scoring function, the program used a semi-empirical free energy force field to score the pose conformations by estimating the energetics for the binding using pair-wise terms to evaluate receptor-ligand interactions and an empirical method to account the contribution from the water solvent environment. The equation used by the software is indicated below:

$$\Delta G = (V_{bound}^{L-L} - V_{unbound}^{L-L}) + (V_{bound}^{P-P} - V_{unbound}^{P-P}) + (V_{bound}^{P-L} - V_{unbound}^{P-L} + \Delta S_{conf})$$

11

As shown in **Equation 11**, free binding energy is given by summing the differences of six pair-wise evaluations (V) that describe the energies of the protein (P) and ligand (L) in both unbound and bound states and the energy of the complex, plus an estimation of the conformational entropy lost upon binding (ΔS_{conf}) which is directly proportional to the amount of rotatable bonds in a molecule. Thus, binding is evaluated in two steps: Firstly, it is calculated the intra-molecular energetics of the transition from the unbound to bound states for ligand and protein separately and then combining the inter-molecular energetics of the protein-ligand complex¹⁴⁸. Each pair-wise V term considers the contribution of different effects on the molecule, as shown in **Equation 12**:

$$V = W_{Vdw} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right) + W_{Hbond} \sum_{i,j} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) + W_{Elec} \sum_{i,j} \frac{q_i q_j}{e(r_{ij}) r_{ij}} + W_{Sol} \sum_{i,j} (S_j V_i + S_i V_j) e^{(-r_{ij}^2 / 2\sigma^2)}$$

12

In this equation, each one of its four terms includes a weighting constant W , which are optimised to calibrate the final empirical free energy based on experimentally-reported binding constants. The first term accounts for the dispersion/repulsion potential and it is based on the Amber force field²⁰⁰. The second term is a directional hydrogen bonding interaction term that includes a $E(t)$ function to provide directionality based on ideal H-bond geometry²⁰¹. The third term describes the Coulomb potential for electrostatic interactions²⁰². The fourth term defines the desolvation potential through the volume of atoms (V) that protect a given atom from the water solvent, weighted by the solvation parameter (S) and an exponential term based on atomic distance²⁰³.

The program was set to calculate, for each docked compound, the ten best ligand conformer poses and their free binding energies. As AutoDock Vina employs a non-deterministic search algorithm, replicates were made for every assay and the median values of the free binding energies calculated for each ligand were considered.

Upon completion, docking pose outputs were examined in PyMOL Molecular Graphics System version 2.4.1. Schrödinger, LLC and UCSF Chimera version 1.14²⁰⁴ to visualize the protein structure with the docked ligands in all calculated conformers. The software LigPlot+ version 2.2.5²⁰⁵ was also used to further examine protein-ligand dynamics for each docked compound by generating a schematic diagram that indicates all predicted interactions between the receptor and ligand.

2.2 Biological studies

2.2.1 Fibroblast culturing

The biological studies described in the following sections were conducted using two human primary fibroblast cell lines: Control line GM00969 (wild type *PMM2*), collected from a healthy female, age 2 at biopsy (Coriell Institute for Medical Research, NIGMS Human Genetic Cell Repository, Camden, New Jersey) and *PMM2*-CDG line GM27386 (c.422G>A/c.415G>A, p.R141H/p.E139K), collected from a female patient, age 5 at biopsy (also purchased from Coriell).

Throughout these experiments and unless mentioned otherwise, fibroblast cell lines were cultured in low-glucose Dulbecco's Modified Eagle Medium, DMEM (Gibco™, Life Technologies) supplemented with 1% (v/v) L-glutamine, 10% (v/v) fetal bovine serum (FBS) and 1% (v/v) penicillin-streptomycin (all from Gibco™, Life Technologies) and kept incubated in a sterile, humidified environment at 37°C in a 5% CO₂ atmosphere. Maintenance-wise, the medium was changed every ~5 days and cell lines were

standardly sub-cultured every 7-10 days, or whenever 100% confluency was achieved, by incubating cultures with a trypsin-EDTA solution (Gibco™, Life Technologies) for 5 min at 37°C, followed by addition of medium to stop trypsinization, and then centrifuged at 200 ×g for 5 min at RT before re-suspending the pellet in fresh DMEM in order to plate a new passage.

Every cell stock used in these experiments was screened for *Mycoplasma spp.* contamination using the Lonza MycoAlert® Detection kit operating in a Lonza Lucetta™ 2 luminometer, following the manufacturer's protocol.

2.2.2 Preparation of test compounds for the biological studies

Four test compounds were acquired for this study:

- [5-(3-Bromo-phenyl)-2,4-dihydro-[1,2,4]triazole-3-thione] (Santa Cruz Biotechnology, SC317500), hereafter mentioned as "BPTT".
- Flurbiprofen [2-(3-fluoro-4-phenylphenyl)propanoic acid] (Sigma-Aldrich, F8514), kindly provided by Prof. Maria Madalena Dionísio.
- Ozenoxacin [1-cyclopropyl-8-methyl-7-[5-methyl-6-(methylamino)pyridin-3-yl]-4-oxoquinoline-3-carboxylic acid] (LGC Labor GmbH, Dr Ehrenstorfer™, DRE-C15825000).
- Riboflavin [7,8-dimethyl-10-[(2S,3S,4R)-2,3,4,5-tetrahydroxypentyl]benzo[g]pteridine-2,4-dione] (Sigma-Aldrich, R4500).

To prepare compounds for *in vitro* testing, the following stock solutions were formulated according to literature or manufacturer's instructions: 10 mg/mL BPTT in 100% DMSO (Sigma-Aldrich, D2438); 10 mg/mL flurbiprofen in 100% DMSO; 5 mg/mL ozenoxacin in 100% DMSO and 10 mg/mL riboflavin in 0.1 M NaOH aqueous solution. Stocks were kept protected from light at 4°C.

2.2.3 Resazurin assays

2.2.3.1 Defining a working cell density for the assays

In a preliminary step to optimise the appropriate cell density for conducting resazurin assays, growing fibroblasts from lines GM00969 and GM27386 were handled as in a typical sub-culturing procedure and the resulting cell suspension was used for seeding 200 µL/well in a sterile, flat bottom, tissue culture 96-well plate at three different densities: 6.0×10^4 , 3.0×10^4 and 1.5×10^4 cells/well. Technical triplicates were used for each condition. After incubation at 37°C for 24 hours, the medium in each well was exchanged for 200 µL of 44 µM resazurin solution (Alfa Aesar, B21187) in DMEM. The plate was incubated at 37°C for 30 min before measuring absorbance at $\lambda = 600$ nm and $\lambda = 570$ nm every 30 min during a three and a half-hour timespan on a Molecular Devices SpectraMax 190 Microplate Reader, heated to 37°C. Data acquisition was performed using the manufacturer's software SoftMax Pro version

6.5.1. (<https://www.moleculardevices.com/>), followed by data processing in Excel, where a viability curve was traced by plotting absorbance difference $Abs_{570\text{ nm}} - Abs_{600\text{ nm}}$ vs. time. Statistical analysis was conducted in GraphPad Prism version 8.0.2. (263) for Windows (GraphPad Software, San Diego, California USA, <https://www.graphpad.com>). Three independent biological triplicates ($n = 1, 2, 3$) were performed at different cell passages: P12, P3, P4 and P13, P4, P5 for GM00969 and GM27386, respectively.

2.2.3.2 Studying the effect of DMSO in fibroblast viability

The effects of DMSO in fibroblast viability were studied by performing 24 hour-exposure resazurin assays. Similarly from before, 200 μL /well of cell suspension from each line were seeded in a 96-well plate at 3.0×10^4 cells/well. After incubation at 37°C for 24 hours, the medium was exchanged for 200 μL of DMSO solution at seven different concentrations: 0%; 0.1%; 0.5%; 1%; 2%; 10% and 20%, prepared by serial dilutions in DMEM. Technical triplicates were used for each condition. After a 24-hour incubation, the DMSO solution was discarded and 44 μM resazurin solution was added to the wells and incubated for half hour prior to the colorimetric assay. Data acquisition and analysis were performed identically as done previously. Statistical analysis was conducted in GraphPad Prism. Three independent biological triplicates were performed for different passages, P5; P6; P7 and P6; P7; P8 for GM00969 and GM27386, respectively.

2.2.3.3 Analysing the effect of test compounds at different concentrations in fibroblast viability

To carry out the biological assays and evaluate *in vitro* the potential chaperoning effect of the four chosen test molecules, defining suitable working concentrations at which none of the fibroblasts' viability would be compromised was an essential first step. Therefore, a set of six test concentrations, plausible to be used in cell systems, ranging from 1 to 200 μM was selected based on literature to perform the assays. With the aim of examining if any of the test compounds would flag a significant loss in cell viability at any proposed concentration, a 48 hour-exposure resazurin assay was conducted. For both fibroblast lines, 200 μL /well of cell suspension were seeded in a 96-well plate at 3.0×10^4 cells/well. Solutions at concentrations of 1 μM ; 5 μM ; 10 μM ; 20 μM ; 60 μM and 200 μM were prepared by serial dilution from stocks in DMEM for all four test compounds (BPTT, flurbiprofen, ozenoxacin and riboflavin). After incubating the seeded plate at 37°C for 24 hours, the medium was exchanged for 200 μL of each test compound solution at all six concentrations. Technical duplicates were made for each condition. After 48-hour incubation, solutions were discarded and 44 μM resazurin was added and incubated before the colorimetric assay as earlier. Data acquisition and analysis were again conducted as stated above. Due to the time frame of this Master's project it was only possible to perform a single assay using passages P5 and P6 for GM00969 and GM27386, respectively.

2.2.4 Assays to evaluate PMM2 enzyme activity in fibroblast extracts

As previously elaborated in **Section 1.4.3**, different reported methodologies on how to measure PMM2 activity, all branching from the original assay by Van Schaftingen & Jaak¹⁷⁴, may be considered. Bearing in mind the type of equipment and reagents available in the laboratory at that time, it was decided that the present assays would be carried out by evaluating PMM2's complementary phosphoglucomutase activity (Glc 1-P $\leftrightarrow$ Glc 6-P) using fibroblast extract samples for analysis instead of purified protein samples, measuring the NADP⁺ reduction reaction through a colorimetric assay performed in a microplate support.

To do so, cultures from both GM00969 and GM27386 lines were grown till $\approx 100\%$ confluent and cells were collected by trypsinisation. The resulting suspensions were washed with PBS 1 $\times$ and centrifuged at 200 $\times g$ for 5 minutes at RT. Soluble extracts were produced from the pellets by adding 50 μL of a lysis solution consisting of *PierceTM IP lysis buffer* (Thermo Scientific) and a protein inhibitor cocktail diluted 1:10 from a stock made by dissolving one tablet of *cOmplete Tablets, Mini EDTA-free, EASY-pack* (Roche) in 1 mL of milli-Q water. Samples were then subjected to thermal shock cycles by flash freeze in an ethanol-ice slurry for 5 min and immediately placed in a dry bath at 37°C for 5 min, repeating the cycle three times to promote cell lysis. Finally, lysates were centrifuged at 17000 $\times g$ for 10 min at 4°C and the resulting supernatants were used for total protein quantification by a bicinchoninic acid assay (BCA) with the *PierceTM BCA Protein Assay Kit* (Thermo Scientific) following the manufacturer's protocol to perform a colorimetric assay that measured absorbance at $\lambda = 562$ nm (BCA absorption max) in a *SpectraMax 190 Microplate Reader*. Data analysis was carried out in Excel and a standard curve was traced by plotting the average blank-corrected measurement for each BSA standard vs. concentration. The standard curve equation was used to calculate total protein concentration for all extracts. Samples containing 20 μg of BCA-measured protein extracts were added to a clean 96-well plate placed on ice and the required volume of enzymatic reaction mix was added to each well in order to complete 200 μL /well. The reaction mix contained 20 mM HEPES buffer pH 7.5 (at 20°C) + 5 mM MgCl₂ (*vide Table 20 in Appendix Section A.2*); 0.1 mM Glc 1-P (Sigma-Aldrich, G7000); 27 μM Glc 1,6-BP (Sigma-Aldrich, 49225); 2,66 U/mL G6PDH enzyme (Sigma-Aldrich, G6378) and 0.25 mM NADP⁺ (Roche). The protocol included a blank condition where no PMM2 activator was supplied and so, a reaction mix prepared without Glc 1,6-BP was added to the blank wells. Technical triplicates were used for each condition. Immediately after adding the reactional mix, plates were placed on a *SpectraMax 190 Microplate Reader* heated to 37°C to measure absorbance at $\lambda = 340$ nm every minute for one hour. Data acquisition was performed as for the resazurin assays and a curve was traced in Excel by plotting the amount of produced NADPH (calculated from Abs_{340 nm} with a molar extinction coefficient, $\epsilon = 6.22$ mM⁻¹cm⁻¹ for NADPH²⁰⁶), as a function of the reaction time in minutes.

2.2.5 Western blotting studies with fibroblasts extracts

2.2.5.1 hPMM2 western blotting

PMM2 western blotting was performed with samples from GM00969 and GM27386 lines by collecting the content of a fully confluent TC flask with growing fibroblasts by trypsinisation. The resulting cell suspensions were washed with PBS 1× and centrifuged at 200 ×g for 5 minutes at RT, after which the pellets were kept. Soluble extracts were produced from the obtained pellet samples by adding 50 µL of lysis solution and samples were subjected to three cycles of thermal shock in EtOH-ice slurry for 5 min and 37°C dry bath for 5 min to stimulate lysis. Finally, lysates were centrifuged at 17000 ×g for 10 min at 4°C and the supernatants recovered. Total protein quantification from the extracts was carried out with the *PierceTM BCA Protein Assay Kit* followed by UV-Vis spectroscopy at $\lambda = 562$ nm just as reported above. SDS-PAGE was performed using the *Mini-PROTEAN Tetra cell* system (Bio-Rad).

For each extract, a 30 µL sample containing 30 µg of BCA-measured protein and 5 µL of 5× *SDS-PAGE Sample Loading Buffer* (NZYTech) was boiled in a dry bath at 95°C for 10 min, centrifuged at 17000 ×g for 1 min and then loaded onto a 1.5 mm-thick gel (stacking and separating gel: 4% and 10% acrylamide 37.5:1, respectively). The *Prestained Protein Ladder - Broad molecular weight (10-245 kDa)* (Abcam) was used as molecular weight marker. Electrophoresis ran for approximately an hour and a half at low temperature with a fixed voltage of 80 V for the stacking, and 140 V for the separating gel, supplied by a *PowerPacTM 300* power unit (Bio-Rad). (For SDS-PAGE buffers and gels recipes, see **Table 20 in Appendix Section A.2**). Following SDS-PAGE, blotting was conducted in order to transfer proteins at the gel matrix onto a *Amersham Hybond P 0.2 PVDF* membrane (GE Healthcare) with a 0.2 µm pore. Wet transferring was performed with the *Mini Trans-Blot* module (Bio-Rad) using the same electrophoresis system and power unit as above, ran at a fixed amperage of 400 mA for one hour at low temperature.

Next, the membrane was blocked with a solution of 5% BSA (Sigma-Aldrich, A3294) in TBS-T 0.1% for one hour at RT under gentle shaking. Afterwards, blocking solution was removed and the membrane incubated O/N at 4°C with the primary antibody, a polyclonal rabbit anti-human PMM2 antibody (Epitomics, ref S2985) diluted 1:1000, kindly provided by Diana Gallego-Martinez. The following day, the primary antibody solution was recovered, the membrane washed with TBS-T 0.1% and a secondary antibody solution containing a horseradish peroxidase (HRP)-conjugated polyclonal donkey anti-rabbit antibody (Jackson ImmunoResearch, 711-035-152) diluted 1:10000 in TBS-T 0.1% was added, incubating for one hour at RT under gentle shaking and protected from light.

The membrane was later washed and signal detection was conducted in a dark chamber by enhanced chemo-luminescence (ECL) using the *Lumi-Light Western Blotting Substrate* kit (Roche), following manufacturer's instructions. The membrane was incubated with the peroxidase-based (POD) substrate for about 1 min before exposing a piece of *Amersham HyperfilmTM ECL* film (GE Healthcare) to the luminescent membrane for 5 min, after which the film was developed using *Carestream® autoradiography GBX* (Sigma-Aldrich) developer and fixer solutions. To ensure that the loaded samples contained equal amounts of total protein, the membrane was later re-probed for α -tubulin as loading control. For

the new probing, the membrane was washed and stripped from the previous staining with *Restore™ Western Blot Stripping Buffer* (Thermo Scientific) for 30 min at RT with gentle shaking. The stripped membrane was washed and the same probing procedures as afore mentioned were employed, with the difference that a polyclonal mouse anti-human α -tubulin primary antibody (Sigma-Aldrich, T6074) diluted 1:50000 in TBS-T 0.1% was used, followed by a polyclonal goat anti-mouse secondary antibody (BD Biosciences, #SS4002) diluted 1:2000 in TBS-T 0.1%. Signal detection was also made by ECL, using a film exposure time of 5 min.

Finally, to perform a quantitative analysis of the results, films were imaged in a *Gel Doc™ EZ* system (Bio-Rad) operated with the built-in Image Lab™ software, version 6.0.1.34. (<https://www.biorad.com>). A stain-free blot imaging protocol for intense bands was selected for image acquisition, band analysis was performed by quantifying the raw volume of each band of interest after background subtraction (raw volume = gray intensity $\times$ no. of pixels within the band area) and normalization was made by calculating a normalizing factor from α -tubulin's loading control bands and applying it to PMM2 band data. At least three replicates using different sub-culturing passages from both the control and wild-type lines were analysed by western blot. Statistical analysis was conducted in GraphPad Prism.

2.2.5.2 ICAM-1 western blot

ICAM-1 western blotting was performed following the same experimental methodology employed for PMM2 minding the following alterations: To prepare soluble extract samples, fibroblasts grown to a $\approx 100\%$ confluency in TC25 flasks were collected, the media were removed and cultures washed with PBS 1 $\times$ before placing them in ice, adding 50 μ L of lysis buffer and scrapping the adherent cells with sterile cell scrappers. The lysates were then collected, submitted to thermal shock cycles and centrifuged to obtain the final extracts. For SDS-PAGE, a 1.0 mm-thick gel was used instead of the previous 1.5 mm gel for PMM2. For the probing step, the primary antibody utilized was a polyclonal rabbit anti-human ICAM-1 antibody (Santa Cruz Biotechnology, SC-7891) diluted 1:1000 in TBS-T 0.1%, also provided by Diana Gallego, and the secondary antibody the same HRP-conjugated polyclonal donkey anti-rabbit antibody as before diluted 1:10000 in TBS-T 0.1%. Signal detection was also performed by ECL using a film exposure time of 45 min. The same loading control detection step was made for α -tubulin.

RESULTS & DISCUSSION

3.1 Cheminformatics component

3.1.1 Building a compound dataset for QSAR model construction

From the *on-line* literature search that was performed at the Web of Science Core Collection, very few bioassays were retrieved where hPMM2 had been tested. No restrictions on what type of chemical/biological outcomes intended for the assays were imposed, and so, different studies reported distinct types of results according to their experiments, such as IC_{50} values and melting temperatures (T_m) for the protein. A total of three different bioassays, associated to three publications were collected.

In the work of Yuste-Checa and collaborators⁸⁷ a high-throughput screening (HTS) was performed using the ten thousand compound library MyriaScreenI Diversity Collection (<https://www.myriascreeen.com/index.php/analytics/diversity>) to identify small molecules capable of thermally stabilizing wild-type hPMM2 produced by heterologous protein expression. HTS was carried out by differential scanning fluorimetry (DSF), which measures the protein's thermal denaturation through monitoring the changes in the fluorescence intensity of an extrinsic fluorophore²⁰⁷. The stabilising effect of each compound is given by the difference from T_m values (ΔT_m) in the presence and absence of the compound. From the whole library, only eight molecules stabilized hPMM2 by $\geq 1^\circ\text{C}$, signifying a possible interaction (**Figure 20**). These were further studied and, from eight compounds, only compounds "IV", "VI", "VII" and "VIII", as nicknamed by the authors, were submitted to degradation time course analysis and PMM2 enzyme activity assays as described somewhere else, using wild-type and patient-observed pathogenic variants of PMM2 to understand the effect of such hits. The assay allowed to calculate an experimental IC_{50} value for these compounds towards wtPMM2 and quantify the effect they have on the residual activity of mutated variants. As a result, 1-(3-chlorophenyl)-3,3-bis(pyridine-2-yl)urea (aka compound VIII) was deemed as especially interesting as a potential lead for pharmacological chaperoning, showing no inhibitory effect under assay conditions and adequate physicochemical and ADME properties.

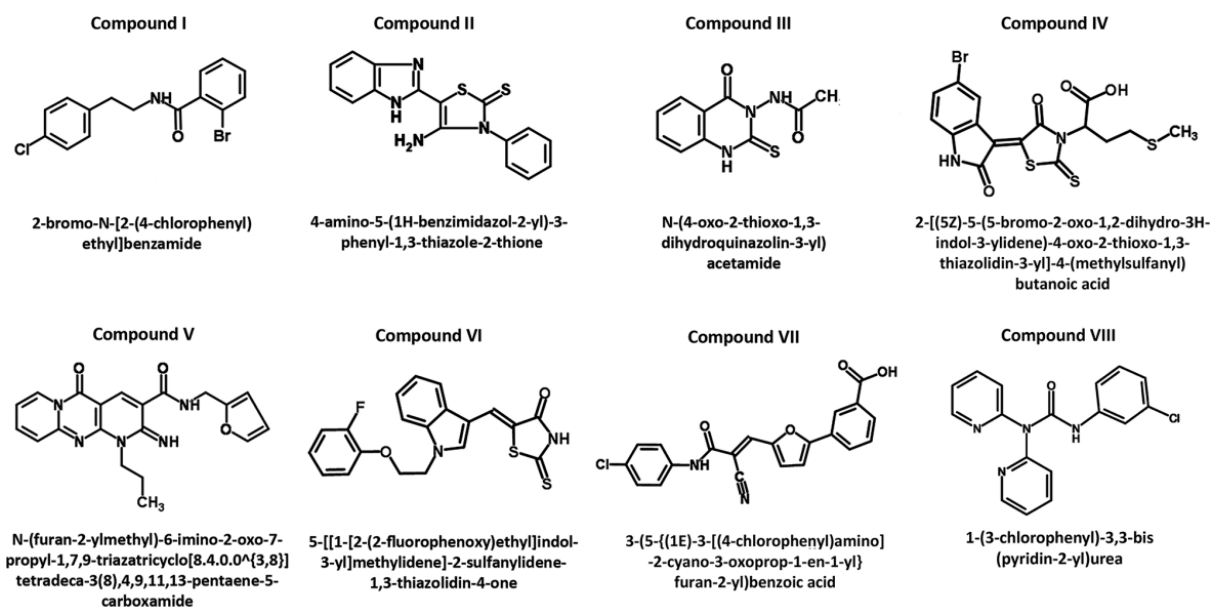


Figure 20. Representation of hit compounds described in the work of Yuste-Checa and collaborators. Adapted from⁸⁷.

Another assay was related to the investigation of Dahl and colleagues²⁰⁸, which explored the application of benzisothiazolinone (BIT) derivatives, identified as a promising scaffold in a previous HTS study, as potent inhibitors for phosphomannoisomerase (PMI) in an effort to develop therapies for PMM2-CDG. As PMI and PMM2 compete for the same substrate, inhibiting PMI (Man 6-P $\leftrightarrow$ Fru 6-P, *vide* **Figure 3**) allows to divert the metabolic flux of mannose by excess to PMM2.

Several BIT derivatives were tested towards PMI and PMM2 inhibition using concentration-response assays to measure enzyme activity as described earlier. IC₅₀ values for different substituted derivatives were determined.

The third and final assay was a follow-up study from the previous article, where Bravo and collaborators²⁰⁹ reported another functional characterization of novel BIT inhibitors, this time for PHOSPHO1, a soluble phosphatase involved in skeletal mineralization and soft tissue ossification processes in a wide number of osteopathologies. Since PHOSPHO1 belongs to the HAD superfamily just as human PMMs and PMI, sharing a similar structural architecture, it was prospected that the previous HTS for PMI inhibition with BIT compounds could be of great interest for developing curative therapies for diseases associated with this phosphatase. A further screen of newly synthesized BIT compounds was evaluated using the same kinetic assays to determine enzyme activity and with it the compounds' inhibitory effects for the three proteins hPHOSPHO1, hPMI and hPMM2. IC₅₀ values were obtained as before.

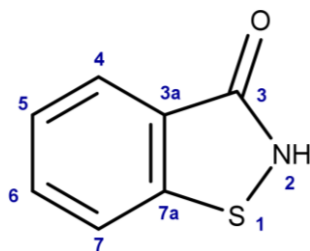


Figure 21. Benzisothiazolinone 2,3-dihydro-1,2-benzothiazol-3-one scaffold. The BIT moiety consists of a bicyclic benzene and thiazolinone rings.

The scarcity of published data and bioassays for hPMM2 is a prime example of the hurdles and consequences of the underrepresentation in rare diseases and comprised an obstacle to bear in mind throughout this work. Note that the limited number of information available for QSAR modelling made it almost mandatory to mix the data collected from different assays, in order to compile the sturdiest and completest dataset possible.

Mixing in the same dataset published biological data from different sources can be problematic, as an assay's results are always specific for the experimental conditions, being comparable only under certain circumstances.

In this work, mixed data was mainly based on IC_{50} results for PMM2. Whilst for large scale examination it would not be viable to check each experiment manually, the reduced number of assays facilitated this aspect: It was possible to read and analyse the information on each assay's conditions and ultimately it was considered that the different IC_{50} values could be interpreted together. Furthermore, by exploring current review literature on the topic, it was understood that mixing IC_{50} data will only contribute to a moderate amount of noise in the final dataset, which can even be surely minimised by performing a careful analysis of experimental conditions and outcomes to reduce chances of errors and variability in the data²¹⁰.

As many chemical databases are semi-automatically curated and therefore, not exempt from error and redundant entries, a considerable amount of duplicates was found within the compound sets retrieved. This situation was promptly addressed by performing a duplicate search to obtain the final duplicate-free 10182 unique molecules for the model building dataset.

3.1.2 QSAR classification Model 1 for interaction prediction

In the dataset curation stage for Model 1, each compound in the dataset received an activity label based on its experimentally-measured ability to interact with PMM2. Every compound that was associated with a defined IC_{50} value and/or that produced an increase in the protein's T_m by $\geq 1^\circ C$ was labelled as "positive", *i.e.*, able to establish a protein-ligand interaction with the enzyme, regardless of its potency. Compounds to which no IC_{50} value was recorded or that have not shown to produce an effect in protein stabilization by DSF were labelled "negatives", deemed not to interact in the enzyme. The dataset contained 60 "positive" and 10122 "negative" molecules, which corresponded to around 0.59% and 99.41% of the total dataset, respectively (*vide* **Excel Appendix**). Since these two categorical classes were widely

disproportionate, with the “negative” class being almost 169 times larger than the “positive” class, the machine learning experiments that were conducted had to account this imbalance and strategies to compensate it had to be used.

3.1.2.1 Calculation of Molecular Descriptors

Descriptor calculation in PaDEL-Descriptor was successful and, for each instance (compound in the dataset) the different FPs and 1D&2D descriptors described in **Table 1** below were generated.

Table 1. Brief descriptive summary of some of the molecular Fingerprints and descriptors available for calculation in PaDEL-Descriptor that were selected to be utilised in this study.

type	Description
CDK Fingerprinter	Contained within the Chemistry Development Kit (CDK) software library there is a function for generating molecular Fingerprints. CDK Fingerprints are circular FPs that represent the input molecule as a vector of binary values, where each value corresponds to the presence or absence of a particular substructure in a certain instance (the input compound). CDK Fingerprints are calculated using a method designated as extended-connectivity Fingerprint (ECFP), which considers the connectivity of atoms in the molecule up to a certain defined radius away from each other ²¹¹ . This FP comprises 1024 individual features (attributes) and, by generating the CDK Fingerprints for an input dataset, it provides for each instance, an either "on" (1 bits) or "off" (0 bits) outcome for every attribute, encoding if the corresponding substructures are observed at the molecule.
CDK Extended Fingerprints	The Extended CDK FPs are an extension of the standard CDK FPs that include supplementary bits describing ring features to obtain additional information about the input instance. Extended CDK FPs use a calculation method named “graph only signature” to encode information about the connectivity of atoms and bonds in the molecule. Similarly to the standard CDK FPs, the extended CDK FPs comprise 1024 individual features represented as binary vectors ²¹¹ .
CDK Graph Only Fingerprints	CDK Graph Only FPs derive from the molecular graph representation of a chemical structure, where the vertices of the graph correspond to the atoms in the molecule and the edges of the graph to the atomic bonds established. Unlike the other circular FPs that encode the presence/absence of certain substructures within a defined distance, Graph Only FPs encode the topology of the molecule's graph without considering bond orders. This Fingerprint is calculated by repeatedly traversing the molecular graph to encode the topological features of each atom and bond in a compound. By doing so, Graph Only FPs can retain information about the connectivity of the atoms and bonds in the molecule, as well as identify the presence of the functional groups that are comprised in its 1024 individual features ²¹¹ .
EState Fingerprints	Electrotopological state FPs are fragment-based FPs that provide an electronic and topological characterization of a molecule. EState FPs first assign a quantitative value for each atom in a molecule based on its electrotopological state, a measurement that contemplates electronic and topological properties. This intrinsic state of each atom in the molecular graph is depicted by an E-state variable, calculated by considering aspects such as element properties, electronic valence, topology of the molecule and connectivity, and perturbations by the electronic influence of neighbouring atoms. EState FPs analyse the molecule in search for the presence or absence of 79 specific E-state substructures (which are the individual attributes) defined by authors Hall & Kier ²¹² in each instance. The final binary output obtained indicates if each E-state is observed or not in the instance.

Substructure Finger- prints	Substructure FPs perform a functional group classification as they code for the presence/absence of particular moieties in a compound. Unlike circular FPs, which work by considering connectivity information or molecular graph properties, Substructure FPs are simply based on the observation of the specific chemical substructures that are defined by its 307 individual attributes on the input molecule. Each bit in the Substructure FPs is encoded by a SMARTS format pattern and, to generate the output, the substructures identified within each instance are calculated and tallied up. The resulting Fingerprint englobes the 307 bits, represented as binary vectors, each identifying a certain substructure in the molecule, as defined by Christian Laggner.
MACCS Finger- prints	The Molecular Access System FPs encode information regarding the presence or absence of certain key substructures in the input molecule. MACCS FPs comprise a defined library set of 166 structural keys (its attributes), each representing a particular molecular substructure commonly involved in the chemical properties explored in drug development, as they are associated with functional groups and molecular fragments found in bioactive compounds and drugs. Some examples of such keys include the presence of certain heteroatoms, aromatic ring structures and many more ¹¹⁴ . When this FP is generated, a binary outcome is provided for every single one of its 166 keys, encoding if the corresponding fragments are present or not.
PubChem Finger- prints	PubChem FPs are another fragment-based FP that are generated by the American National Centre for Biotechnology Information (NCBI) PubChem database ¹¹⁵ . Similarly to the Substructure FPs, they can be employed to identify chemical substructures and functional groups within the input molecules. This Fingerprint represents an 881-bit binary string that calculates, for each instance, the presence or absence of the specific substructures denoted in its predefined set of SMARTS patterns, covering a vast array of moieties, associated to different biochemical properties and designed to be broadly applicable to a variety of research settings such as drug discovery approaches.
1D&2D	The 1D&2D descriptor set accounts for several attributes from different subtypes (a total of 1444 individual features in PaDEL-Descriptor) encoding for various aspects that provide a wide range of information regarding the overall size, complexity and topology of an input molecule, such as atom counters and atom type indexes; atomic bond counters; H-bond counters; various functional group counters; indexes for topological properties like charge, electronic state, surface area, non-covalent interactions and many other aspects ¹¹² .

3.1.2.2 Partitioning the dataset

During dataset partition, the training set was constituted by 9151 compounds with 53 of them being labelled as “positive” and including the eight hit compounds identified in the work of Yuste-Checa *et al.*⁸⁷, purposefully added to the training set given their importance to predict the intended biological activity. The training set also included 9098 “negatives” to a final positive-negative ratio of 1:172. The test set contained 1021 compounds, with 7 “positives” and 1014 “negatives” (1:146 ratio).

3.1.2.3 Trial models to optimise algorithm parameters and select input descriptors

Designing and training a QSAR model requires a comprehensive optimisation round where the ML algorithm is iteratively fine-tuned by tweaking its different hyperparameters and running several preliminary models to define the best conditions, under which the model’s predictive performance (generalisation capacity) is gradually improved to be maximal^{213,214}. The evaluation metrics mentioned earlier (TP; FP; TN; FN; SE; SP; Q and MCC) are an essential tool to monitor the progress in optimisation.

The first step consisted of training RF models in WEKA to explore adequate algorithm parameters to perform the optimisation rounds ahead. Random Forests provide suitable models for this exercise, as

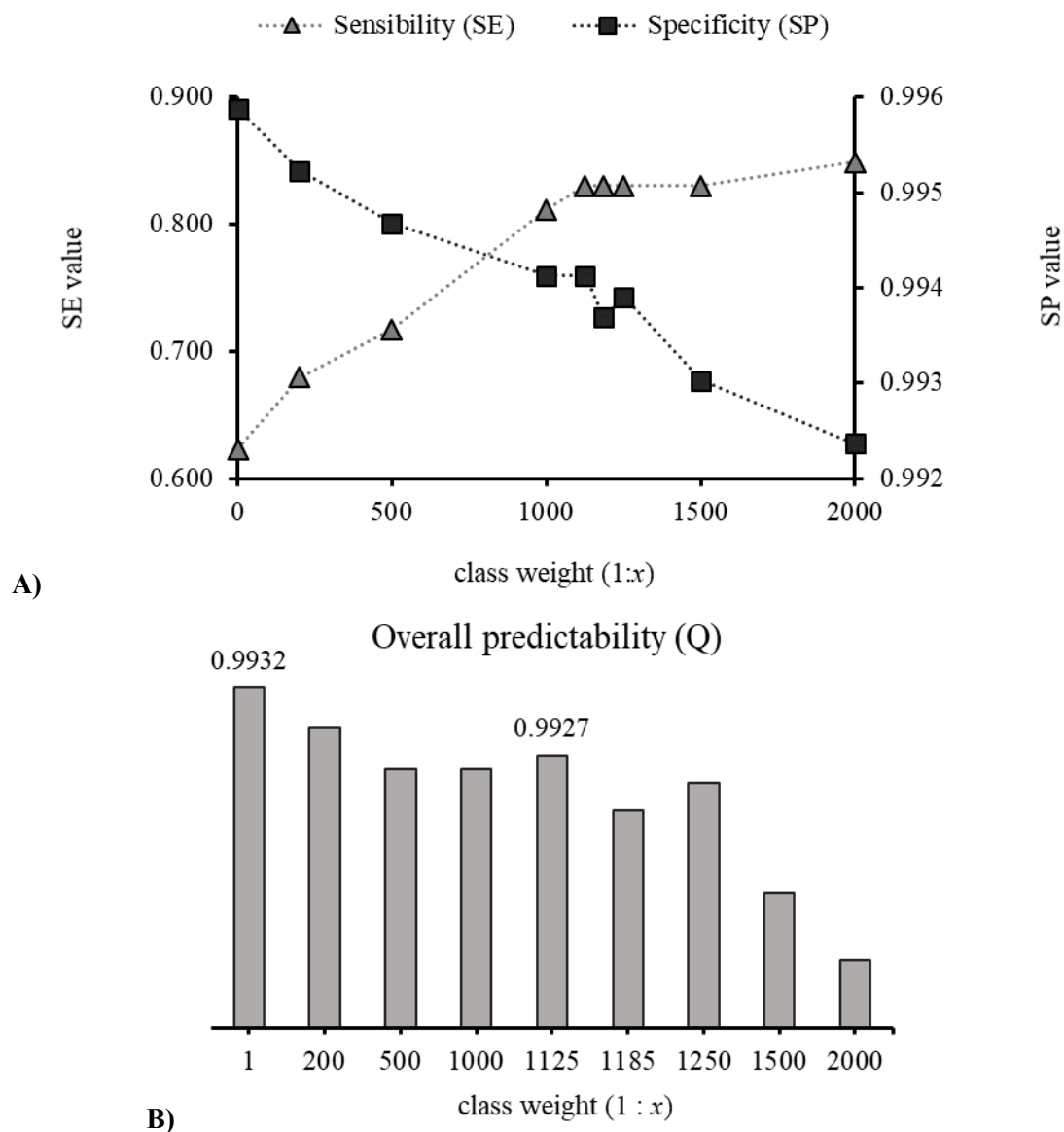
this ML algorithm is relatively simple, robust, does not require descriptor selection and allows for fast calculations²¹⁵, the latter being an important advantage in optimisation steps, which tend to involve training several trial models taking over a wide timespan. Models were trained with the training set's MACCS keys FPs, mainly due to its small attribute number (166) and simplicity, contributing for speeding up calculations.

One of the parameters that was subjected to optimisation was the RF algorithm's *no. of iterations*, which is number of trees used in the forest ensemble as set by the user. There is not a defined rule for the number of trees to be used to compose a forest and sometimes it is uncertain if it should simply be the biggest computationally feasible number or whether a smaller number suffices. Following the bagging principle, the more trees there are, better the model is, although in practice model performance tends to reach a minimum, *i.e.*, a threshold, from which increasing the number of trees does not improve the classification error and only increases computational cost. This means that model performance is not directly proportional to tree number and, therefore, augmenting it does not necessarily improve the RF^{216,217}. In order to find an optimal number of trees for this QSAR model, the *no. of iterations* was varied from 100; 150; 200; 250; 300; 350; 400; 450 and 500 trees. As increasing numbers of trees have not provided a better result than the model with 100 trees, the *no. of iterations* was kept with 100 trees for simplicity's sake and to alleviate the computational load.

Other parameter that was tweaked was the algorithm's *mtry*, that is, the number of features sampled. It controls the maximum number of features (the individual variables/attributes from the input descriptors) that are considered for splitting at each node of each decision tree in the forest. Setting of the number of features greatly influences predictive performance: Whilst using a lower value can result in reducing over-fitting, it may also decrease the algorithm's ability to generalise, affecting its accuracy. On the other hand, opting for higher values may improve accuracy, but also increase the risk of over-fitting. During this optimization phase, different *mtry* values were tested, including a third, the square root, and the $\log_2$ of the total number of attributes. These specific values were employed because they have been reported to work well for RF strategies. The rationale behind using them is that they provide a good balance between considering a sufficient number of features for splitting and, at the same time, these numbers are reduced enough to prevent over-fitting.

Another crucial factor to keep in mind in this QSAR strategy was class balancing. As mentioned above, the input dataset that was used to train the ML models is heavily imbalanced, *i.e.*, the number of instances in the "negative" class is much larger than those in the "positive" class. This situation may result in biased predictions, as the model can much more easily predict the major "negative" class over the minor "positive" class. A few ways of overcoming this issue exist, including cost-sensitive learning. Quite simply, cost-sensitive learning relies on balancing the weight that each of the two classes in the training set has to enhance model performance by making the algorithm allocate a bigger weight to the minority class of "positive" compounds. By doing so, the algorithm assigns a heavier penalty cost on misclassifying the minor "positive" class than the major "negative" class, which helps to contradict the biased tendency of the model towards the majority class.

To do this, the WEKA's RF classifier's *cost matrix* parameter was gradually adjusted, balancing the weight of each categorical class. **Figure 22** below shows the effect that class weight balancing has on the evaluation metrics used to estimate performance. Different class weight balances were used, initiating at an equal cost (1:1) and increasing the weight for the minor class until 1:2000. From the wide range of values that were experimented, the present data only represents the results obtained for the following class weights: 1:1; 1:200; 1:500; 1:1000; 1:1125; 1:1185; 1:1250; 1:1500 and 1:2000.



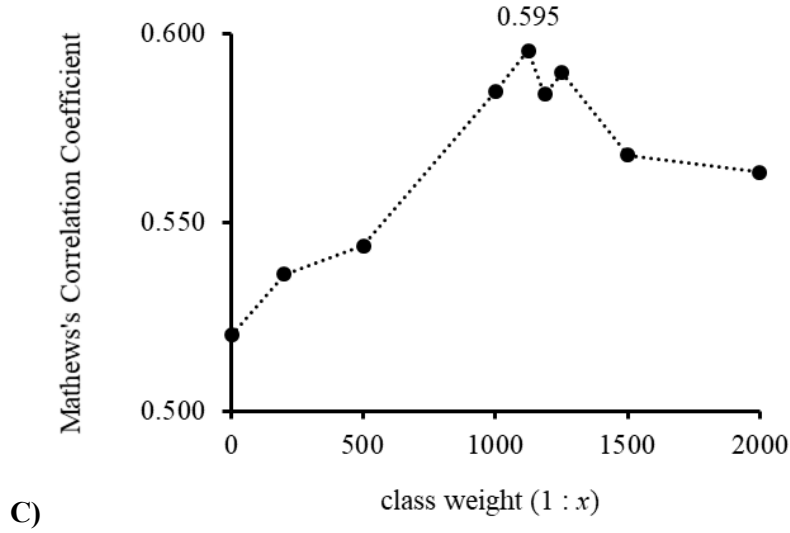


Figure 22. Variation of model performance evaluation metrics with different class weights. RF models were trained in WEKA using the *meta Cost Sensitive Random Forest* classifier varying only the built-in class weight cost matrix hyperparameter. *Out-of-bag* estimations for the training set were used and a *10-fold cross validation* testing option with the training set was employed to obtain the evaluation metrics for the model. The plots represent the values obtained for the models built with the 1:1; 1:200; 1:500; 1:1000; 1:1125; 1:1185; 1:1250; 1:1500 and 1:2000 negative-positive weights. **A)** Scatter plot for Sensitivity (SE) and Specificity (SP) as a function of class weight. **B)** Bar chart for Overall Predictability (Q) as a function of class weight. **C)** Scatter plot for the Matthews's Correlation Coefficient (MCC) as a function of class weight.

By analysing the data in **Figure 22A** it is possible to observe an increase in Sensitivity (SE) and a decrease in Specificity (SP) with the rise of the class weight cost for the “positive” category. This is easily explained by the fact that whilst SE is a measure of the true positive (TP) rate, indicating how well the model classifies TPs, SP measures the true negative (TN) rate, that is, how accurately are TNs classified²¹⁸. As, by increasing the weight of the positive class, the algorithm favours positive classification, the SE rises, as more positive instances are getting correctly identified, and the SP lowers as more negative instances are gradually getting misclassified as false positives (FP) (*vide Equations 1 and 2*).

The overall predictability (Q) seen in **Figure 22B** slowly decreases from an initial value of 0.9932 for a 1:1 weight onwards, being considerably lower for the higher weights of 1:1500 ($Q = 0.9916$) and 1:2000 ($Q = 0.9910$). This happens because the Q's numerator (*vide Equation 3*) $TP+TN$ reduces in value, as the increase of TPs is far lower than the rate of TNs that are being misclassified, resulting in a higher number of FPs. The higher class weights show a more pronounced decrease, which implicates that the model's accuracy for predicting the negative class is compromised, meaning that these weights should not be further used. Whilst this decrease is important and should not be disregarded, note that Q values do not vary greatly for both extremes, in fact $Q_{1:1} = 0.9932$ and $Q_{1:2000} = 0.9910$. This happens because the number of negative instances is immensely greater than the number of positive instances, even for the heavier weights, meaning that the Q's numerator, for this case, is almost insensitive to TP contribution: $TP+TN \approx TN$.

The Matthews's Correlation Coefficient variation is depicted in **Figure 22C**, rising from an initial value of 0.520 for a weight of 1:1 to a peak value of 0.595 at a weight of 1:1125 and then decreasing to 0.563 at a 1:2000 weight balance. The mountain-like profile of the plot suggests that the class weight of 1:1125 acts as an optimal threshold for balancing, where lower weights fail to improve the classification of the minor "positive" class and heavier weights start to critically ignore the misclassification of negative instances.

This analysis shows the importance of considering the entire set of evaluation metrics available to interpret the model as a whole, since metrics such as SP and Q are only suitable to depict the behaviour of the model towards the "negative" class prediction, incorrectly indicating a worsening of model performance along the rise in weight balance. SE is also not ideal by itself, as it only portrays the TP rate and fails to describe the impact that an overly positively weighted cost has on "negative" classification. The MCC, in conjunction with the former metrics, allows to properly understand how class weighting improves performance.

All in all, the class weight is a true balancing effort that must strive to provide an optimised representation of the minority class in the dataset and concomitantly consider the misclassification cost on the majority class. For this case, the best model for further optimisation steps was trained using 100 trees, an *mtry* of a third of the total features (55) and a class weight of 1:1125.

With the best-performing operating hyperparameters being successfully defined, preliminary RF trial models were trained in WEKA using each one of the seven Fingerprint groups, plus the 1D&2D descriptors group calculated in PaDEL-Descriptor to analyse which one best describes the chemical properties that allow to accurately differentiate compounds between both classes. The predictive performance of such trial models was assessed using the same metrics as before and the best-performing circular and fragment-based FPs were chosen (**Table 2**).

But *why* was it desired to train the model using only one descriptor over the whole set? As different descriptor types encode distinct properties and have diverse ranges and units of measure, which make comparing the relative importance of each FP/descriptor an unbearable task, selecting only one descriptor type is considered to be the best practice. By choosing only one descriptor type to pursue model construction, the ML algorithm's complexity can be kept at bay, resulting in a model with an improved capacity to focus on the relationship between chemical structure and property prediction (i.e., generalisation).

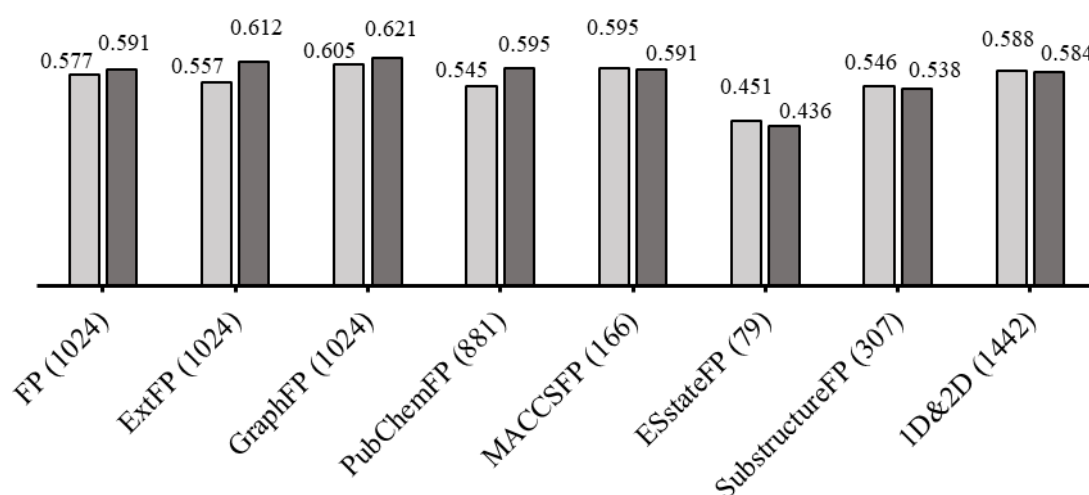


Figure 23. Bar chart for Matthews's Correlation Coefficient values for trial models using different FP/descriptor types. The graph represents the MCC values obtained for the models described below in Table 2. Models trained with an *mtry* of one third of the total attribute number are depicted in light grey whilst models using a square root *mtry* are illustrated in a darker grey.

Table 2. Preliminary analysis of trial RF QSAR models to evaluate the predictive performance of each FP/descriptor group. Trial models were trained in WEKA with the *meta Cost Sensitive Random Forest* classifier for the circular Fingerprinter, Extended FPs and Graph Only FPs; the fragment-based PubChem FPs, MACCS FPs, Substructure FPs and the whole 1D&2D descriptors set, using all attributes. RF parameters were kept as the predefined optimal hyperparameter settings, except for the *mtry* (no. of features), which was varied from a third and the square root of the total attribute number. Bold lines indicate the best-performing model for each FP type and descriptor set according to the evaluation metrics generated during training.

type	FP/Descriptors (no. of attributes)	<i>mtry</i> [*]	TP ^a	FP ^b	TN ^c	FN ^d	SE ^e	SP ^f	Q ^g	MCC ^h
Circular FPs	Fingerprinter (1024)	1/3	41	53	9045	12	0.774	0.994	0.993	0.577
		√	41	49	9049	12	0.774	0.995	0.993	0.591
	Extended FP (1024)	1/3	40	56	9042	13	0.755	0.994	0.992	0.557
		√	42	46	9052	11	0.792	0.995	0.994	0.612
	Graph Only FP (1024)	1/3	44	55	9043	9	0.830	0.994	0.993	0.605
		√	44	50	9048	9	0.830	0.995	0.994	0.621
Fragment FPs	PubChem FP (881)	1/3	42	69	9029	11	0.792	0.992	0.991	0.544
		√	42	51	9047	11	0.792	0.994	0.993	0.595
	MACCS FP (166)	1/3	44	58	9040	9	0.830	0.994	0.993	0.595
		√	43	56	9042	10	0.811	0.994	0.993	0.591
	EState FP (79)	1/3	39	99	8999	14	0.736	0.989	0.988	0.451
		√	37	96	9002	16	0.698	0.989	0.988	0.436
	Substructure FP (307)	1/3	44	77	9021	9	0.830	0.992	0.991	0.546
		√	41	67	9031	12	0.774	0.993	0.991	0.538
whole 1D&2D set (1442)	1/3	38	40	9058	15	0.717	0.996	0.994	0.588	
	√	36	35	9063	17	0.679	0.996	0.994	0.584	

^{*}The *mtry* was set either as a third (1/3) or square root (√) of the total attribute number for each FP/descriptor set.

a) TP = true positive. b) FP = false positive. c) TN = true negative. d) FN = false negative. e) SE = Sensitivity. f) SP = Specificity. g) Q = Overall Predictability. h) MCC = Matthews's Correlation Coefficient.

Examination of the obtained results indicated that the best-performing models trained with the circular and fragment-based Fingerprint types were the Graph Only FPs with an MCC = 0.621 and the MACCS FPs with an MCC = 0.595, respectively. The 1D&2D model yielded an MCC = 0.588. Testing different *mtry* options produced results with very slight differences, the most notorious discrepancy being observed for the PubChem FP model where the $MCC_{1/3} = 0.545$ and $MCC_{\sqrt{}} = 0.595$ ($\Delta MCC = 0.05$), indicating that both the third and square root values are appropriate feature selection number for model training, concordant to what has been verified across literature.

3.1.2.4 Descriptor attribute selection

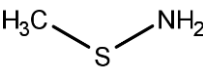
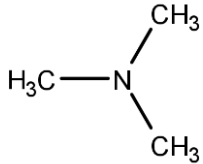
The three models (Graph Only FP, MACCS FP and the 1D&2D set models) proceeded to the next stage, where individual attribute importance for each FP/descriptor was calculated in the R software using the R's RF algorithm *Mean Decrease Accuracy* tool.

Mean Decrease Accuracy is a measure of feature (attribute) importance that can be employed to determine which features are the most relevant for the generalisation capacity of the algorithm. This is calculated by randomly permuting the values of each feature and comparing the decrease in the accuracy of the model with and without each feature. The difference in accuracy is then averaged over all the trees in the forest, providing the final output: the larger the decrease in accuracy, the more important a certain feature is. This tool allows to highlight the most important attributes within a descriptor and discard the less important ones with the scope of improving the accuracy of the forest²¹⁹.

For each FP/descriptor, the R software provided a list of attributes sorted by order of importance. While it would be hardly possible to enumerate in detail the entire attribute list for the three FP/descriptor sets explored, the top ten best attributes for MACCS FPs and 1D&2D descriptors are depicted in **Tables 3 and 4** below. Note that the same analysis could not be done for Graph Only FPs since the features that this Fingerprint encodes are not easily interpretable in a physical sense.

Table 3. Top ten MACCS FP attributes measured by Mean Decrease Accuracy. This table displays the ten most important MACCS FP individual features as calculated by the R software's RF library Mean Decrease Accuracy metric. For each attribute, its hierarchy, corresponding key fragment in SMARTS format and chemical structure representation is depicted.

Hierarchy	Attribute no.	Key SMARTS	Key description*
1	MACCSFP62	A\$A!A\$A	atom is at a ring/chain boundary
2	MACCSFP33	NS	H_2N-CH_2-SH
3	MACCSFP97	NAAAO	$H_2N-CH_2-CH_2-CH_2-OH$
4	MACCSFP46	BR	Br
5	MACCSFP47	SAN	$HS-CH_2-NH_2$

6	MACCSFP103	CL	Cl
7	MACCSFP32	CSN	
8	MACCSFP67	QS	HS—Q
9	MACCSFP85	CN(C)C	
10	MACCSFP73	S=A	S=A

* A = an atom of any element; Q = any heteroatom (no C or H)

According to **Table 3**, the most important attribute is related to the atomic bonds observed in a cyclic structure, which is not surprising since the dataset encompasses a wide number of molecules containing ring systems. In fact, all hit compounds from Yuste-Checa's work, as well as the BIT scaffold, contain at least one ring in its molecular structure.

Furthermore, ring moieties are a very common observation in CADD approaches since over 90% of known drugs possess at least one ring system. Rings comprise the core scaffold of most drugs, dictating their shape and three-dimensional conformation, therefore deeply influencing their biological activity²²⁰. Literature reports over 350 known ring systems with additional six new rings being introduced in the drug space every year^{221,222}.

From this top ten it was also noted the presence of several nitrogen-containing fragments (2nd, 3rd, 5th, 7th and 9th places) and sulfur-containing fragments (2nd, 5th, 7th, 8th and 10th places) which, again, are expected observations considering that 84% of all FDA-approved drugs have one or more nitrogen atoms, with 59% of drugs including some nitrogen heterocycle system²²³ and 249 drugs containing a functional group with at least one sulfur atom²²⁴. The importance given to such key fragments may also be an indicative sign of the overrepresentation of the BIT scaffold in the dataset, highlighting its structural features.

Table 4. Top ten 1D&2D attributes measured by Mean Decrease Accuracy. This table displays the ten most important 1D&2D individual features, as calculated by the R software's RF library Mean Decrease Accuracy metric. For each attribute, its hierarchy, descriptor name and typology, as well as attribute description is indicated.

Hierarchy	Descriptor name	Descriptor type	Attribute	Attribute description
1	MATS7m	topological	Autocorrelation	Moran autocorrelation - lag 7 / weighted by mass
2	SpMax1_Bhp	topological	Burden modified eigenvalues	Largest absolute eigenvalue of Burden modified matrix - n 1 / weighted by relative polarizabilities

3	minaaS	count	Atom type electrotopological state	Minimum atom-type E-State: aSa (S atom bearing two aromatic bonds)
4	SpMax1_Bhv	topological	Burden modified eigenvalues	Largest absolute eigenvalue of Burden modified matrix - n 1 / weighted by relative van der Waals volumes
5	SpMin2_Bhi	topological	Burden modified eigenvalues	Smallest absolute eigenvalue of Burden modified matrix - n 2 / weighted by relative first ionization potential
6	nAtomP	count	Largest π system	Number of atoms in the largest π system
7	JGI4	topological	Topological Charge Descriptor	Mean topological charge index of order 4
8	nHBint4	count	Atom type electrotopological state	Count of E-State descriptors of strength for potential Hydrogen Bonds of path length 4
9	GATS5m	topological	Autocorrelation	Geary autocorrelation - lag 5 / weighted by mass
10	minsF	count	Atom type electrotopological state	Minimum atom-type E-State: -F

By examining **Table 4** above it is seen that, amongst the most important 1D&2D descriptor attributes, lie the Burden Modified Eigenvalue features (which occupy the 2nd, 4th and 5th places) and the Atom type Electrotopological State features (3rd, 8th and 10th places).

Atom Type Electrotopological State features work similarly to EState FPs, but considering the two-dimensionality of the input molecule to encode the electronic environment.

In their turn, Burden Modified Eigenvalues represent the eigenvalues (scalar values) of the burden matrix (a matrix that encodes the distance between pairs of atoms in a molecule), coding for information regarding electronic configuration, symmetry, charge distribution and polarity, ionisation potential, non-covalent interactions and even detailing certain functional groups²²⁵.

With the individual attributes of each FP/descriptor set organized by order of importance, model optimisation proceeded to define what number of attributes used for training provide the best prediction accuracy. As shown in **Table 5** below, for each FP/descriptor, try-out RF models in WEKA were built using all attributes and then sets of the 250; 200; 150; 100 and 50 best features (most importance weighted attributes).

Table 5. Analysis of try-out RF QSAR models to evaluate the predictive performance of different sets of descriptor attributes. Trial models were trained in WEKA with the *meta Cost Sensitive Random Forest* classifier for MACCS FPs, Graph FPs and 1D&2D sets built with all attributes, and then gradually downsizing to the 250, 200, 150, 100 and 50 best attributes as weighted by importance in the R software. RF parameters were kept as the predefined optimal hyperparameter settings. Only the best models obtained for each set and attribute number are represent, further optimisation is not shown. Bold lines indicate the best-performing model for each FP/descriptor set according to the evaluation metrics generated during training.

FP/Descriptor (no. of attributes)	TP ^a	FP ^b	TN ^c	FN ^d	SE ^e	SP ^f	Q ^g	MCC ^h
MACCS FP (166)	44	58	9040	9	0.830	0.994	0.993	0.595

MACCS FP (50)	46	62	9036	7	0.868	0.993	0.992	0.605
Graph FP (1024)	44	50	9048	9	0.830	0.995	0.994	0.621
Graph FP (250)	43	49	9049	10	0.811	0.995	0.994	0.613
Graph FP (200)	43	49	9049	10	0.811	0.995	0.994	0.613
Graph FP (150)	43	49	9049	10	0.811	0.995	0.994	0.613
Graph FP (100)	44	50	9048	9	0.830	0.995	0.994	0.621
Graph FP (50)	41	47	9051	12	0.774	0.995	0.994	0.597
1D&2D (1442)	38	40	9058	15	0.717	0.996	0.994	0.588
1D&2D (250)	38	35	9061	15	0.717	0.996	0.995	0.608
1D&2D (200)	36	34	9062	17	0.679	0.996	0.994	0.588
1D&2D (150)	39	31	9065	14	0.736	0.997	0.995	0.638
1D&2D (100)	41	33	9063	12	0.774	0.996	0.995	0.652
1D&2D (50)	40	38	9058	13	0.755	0.996	0.994	0.620

a) TP = true positive. b) FP = false positive. c) TN = true negative. d) FN = false negative. e) SE = Sensitivity. f) SP = Specificity. g) Q = Overall Predictability. h) MCC = Matthews's Correlation Coefficient.

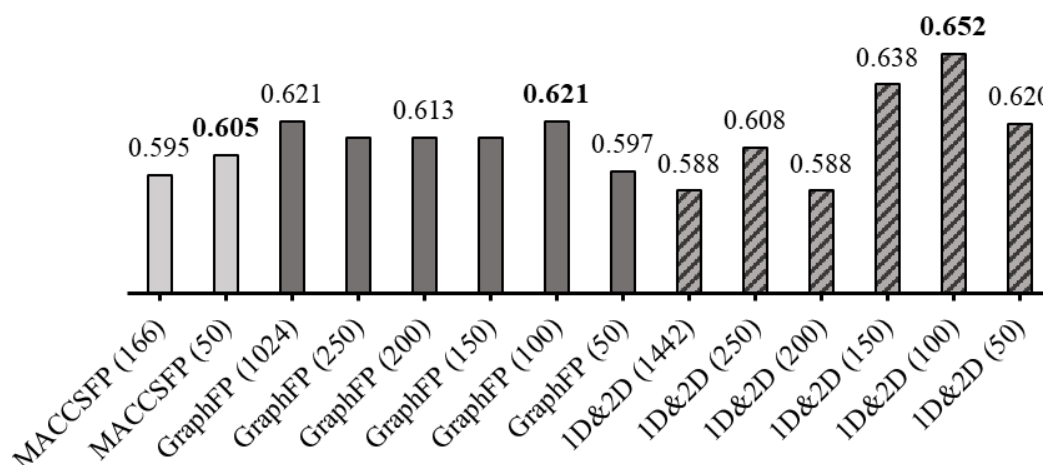


Figure 24. Bar chart for Matthews's Correlation Coefficient values for the try-out models. The graph represents the MCC values obtained for the models described above in Table 5. Models trained with MACCS FPs are represented in light grey, Graph Only FPs in dark grey and the 1D&2D set in striped grey. The attribute number used for training is indicated in parenthesis. Bold values indicate the MCC for the best performing model from each set.

The results obtained show that the best models were produced using the MACCS FP set with the 50 most important attributes, $MCC_{MACCS\ FP} = 0.605$, the Graph Only FP set with the 100 most important attributes, $MCC_{Graph\ FP} = 0.621$ and the 1D&2D descriptors set with the 100 most important attributes, $MCC_{1D\&2D} = 0.652$.

3.1.2.5 Finalised QSAR predictive models with RF, SVM and NN algorithms with evaluation from test set

The final QSAR models were trained for each ML algorithm and with each FP/descriptor set, as mentioned in **Section 2.1.1.3**. The following paragraphs describe the outcomes of these models and the final results are compiled in **Table 6**.

For the RF models built in WEKA, the MACCS FP model was trained with 100 trees, an *mtry* of a third, and a *cost matrix* = 1:1125 resulting in a $MCC_{MACCS\ FP} = 0.605$. The Graph Only FP model was trained with 100 trees, an *mtry* of a third, and a *cost matrix* = 1:1125 resulting in a $MCC_{Graph\ FP} = 0.621$. The 1D&2D model was trained with 100 trees, an *mtry* of the square root, and a *cost matrix* = 1:1125 resulting in a $MCC_{1D\&2D} = 0.652$. Other hyperparameters were kept as default.

SVM models were also built in WEKA using the two approaches described earlier. After a few optimisation rounds and fine-tuning the classifier's parameters (data not shown), it was concluded that the best approach for SVMs consisted on using the WEKA's *Explorer LibSVM* classifier algorithm with *C-SVM classification* type, a radial basis kernel function and algorithm normalisation turned on, varying the classifier's *cost* and *weight* parameters.

When training SVM models it is important to contemplate input normalisation (also designated as feature scaling). Since the algorithm is sensitive to the scale of values of the input descriptors used for training, if its attributes use different unit scales, the model gives more weight to the attributes with larger values, regardless of their performance. Normalization scales all features between 0 and 1 to ensure that the contributions from each descriptor are evenly weighted.

Another important hyperparameter in SVMs is the *cost*. It allows the user to regulate the trade-off between achieving a low training error and a low testing error. A larger cost value makes the algorithm strive for a smaller margin in order to correctly classify as many training set instances as possible. However, this makes the model more prone to overfitting. Employing a smaller cost translates into a larger margin where misclassification of train instances may be more frequent, although this can also lead to underfitting.

In its turn, the *weight* parameter is used to assign different weights to different categorical classes, just like the class weight parameter in the RF algorithm.

After many iterative optimisation steps, the final MACCS FP model with the best 50 attributes was trained with a *cost* = 12 and a *weight* = 1:1125 resulting in a $MCC_{MACCS\ FP} = 0.581$. The Graph Only FP model with the best 100 attributes was trained with a *cost* = 1 and a *weight* = 1:25 resulting in a $MCC_{Graph\ FP} = 0.605$. The 1D&2D model with the best 100 attributes was trained with a *cost* = 5 and a *weight* = 1:17 resulting in a $MCC_{1D\&2D} = 0.558$. Other hyperparameters were kept as default.

The NN algorithm models were trained in Keras using a specific script, tweaking the network hyperparameters such as the *no. of hidden layers* in the network; the *no. of neurons*; the *optimiser*; and the *class weight*.

To optimise NN models, the number of hidden layers and the number of neurons in each layer should be tweaked to achieve the best predictive accuracy possible. In the present approach, the starting architecture of the NN try-out model was simpler, with a low no. of layers and neurons. Then, the model's complexity was gradually increased to assess its effect on the performance (data not shown).

Another parameter that was tested was the network's *optimiser*. The optimiser is used to determine how the network's weights are adjusted in response to the error gradient δ generated during backpropagation²²⁶ (vide **Figure 11**). Keras provides some options of different optimiser algorithms, including the Adam, Adadelata, Adagrad, SGD and RMSprop. Throughout this work, the Adam algorithm (Adaptive Moment Estimation) proved to be an efficient optimiser.

After the optimisation stage, all models were trained with 1 layer (initial + final, no hidden layer), the *Adam* classifier and a *class weight* = 1:7 resulting in a $MCC_{MACCS\ FP} = 0.765$, $MCC_{Graph\ FP} = 0.851$ and $MCC_{1D\&2D} = 0.929$. Other parameters were kept as default.

Table 6. Analysis of the final models trained for different ML algorithms. The final models using Random Forest, Support Vector Machine, and Neural Networks were built with the optimized hyperparameters for MACCSFPs, GraphFPs and 1D&2D sets with the predefined amount of attributes. Each line in the table indicates the final model evaluation metrics generated during training.

ML algorithm	FP/Descriptor (no. of attributes)	TP ^a	FP ^b	TN ^c	FN ^d	SE ^e	SP ^f	Q ^g	MCC ^h
<i>mCSC</i> RF	MACCS FP	46	62	9036	7	0.868	0.993	0.992	0.605
<i>mCSC</i> RF	Graph FP	44	50	9048	9	0.830	0.995	0.994	0.621
<i>mCSC</i> RF	1D&2D	41	33	9063	12	0.774	0.996	0.995	0.652
SVM	MACCS FP (50 best)	44	63	9035	9	0.830	0.993	0.992	0.581
SVM	Graph FP (100 best)	46	62	9036	7	0.868	0.993	0.992	0.605
SVM	1D&2D (100 best)	45	76	9020	8	0.849	0.992	0.991	0.558
NN	MACCS FP (50 best)	49	28	9070	4	0.925	0.997	0.997	0.765
NN	Graph FP (100 best)	53	20	9078	0	1.000	0.998	0.998	0.851
NN	1D&2D (100 best)	52	7	9089	1	0.981	0.999	0.999	0.929

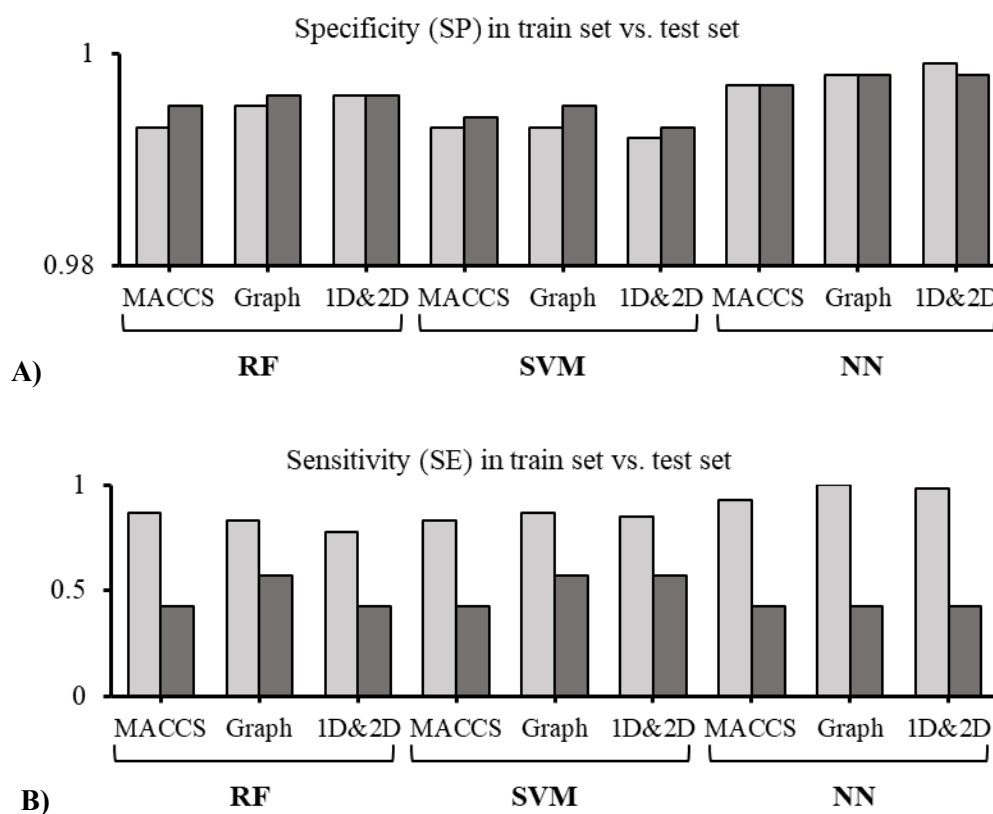
a) TP = true positive. b) FP = false positive. c) TN = true negative. d) FN = false negative. e) SE = Sensitivity. f) SP = Specificity. g) Q = Overall Predictability. h) MCC = Matthews's Correlation Coefficient.

These QSAR models were then tested with the 1021-molecules external test set isolated from the original dataset to evaluate their generalisation ability for a novel set of instances. **Table 7** compiles the test set evaluation results obtained for each final model, illustrated as well in **Figure 25**.

Table 7. Evaluation with test set of the final models trained with different ML algorithms. The prediction accuracy of final RF, SVM and NN models was evaluated with the external test set. Each line in the table indicates the performance metrics generated.

ML algorithm	FPs/Descriptors (no. of attributes)	TP ^a	FP ^b	TN ^c	FN ^d	SE ^e	SP ^f	Q ^g	MCC ^h
<i>mCSC</i> RF	MACCS FP	3	5	1009	4	0.429	0.995	0.991	0.396
<i>mCSC</i> RF	Graph FP	4	4	1010	3	0.571	0.996	0.993	0.531
<i>mCSC</i> RF	1D&2D	3	4	1010	4	0.429	0.996	0.992	0.425
SVM	MACCSFP (50 best)	3	6	1008	4	0.429	0.994	0.990	0.373
SVM	GraphFP (100 best)	4	5	1009	3	0.571	0.995	0.992	0.500
SVM	1D&2D (100 best)	4	7	1007	3	0.571	0.993	0.990	0.451
NN	MACCSFP (50 best)	3	3	1011	4	0.429	0.997	0.993	0.459
NN	GraphFP (100 best)	3	2	1012	4	0.429	0.998	0.994	0.504
NN	1D&2D (100 best)	3	2	1012	4	0.429	0.998	0.994	0.504

a) TP = true positive. b) FP = false positive. c) TN = true negative. d) FN = false negative. e) SE = sensitivity. f) SP = specificity. g) Q = overall predictability. h) MCC = Matthews Correlation Coefficient.



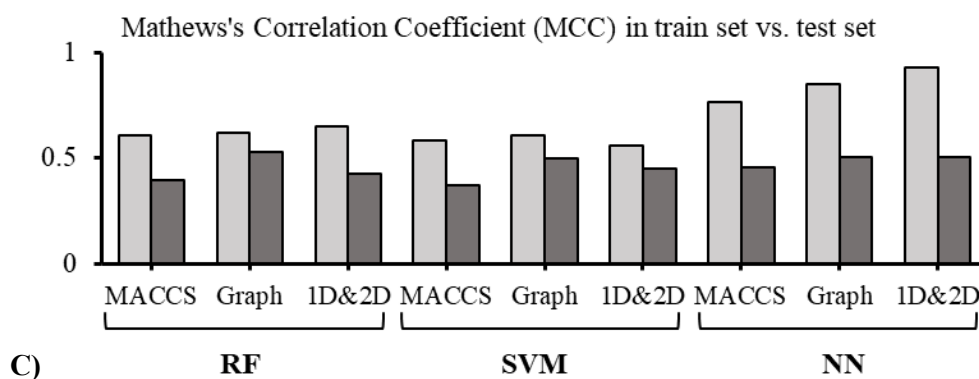


Figure 25. Bar charts with comparisons between Sensitivity (SE), Specificity (SP) and Matthews's Correlation Coefficient values. The SP (chart A), SE (chart B) and MCC (chart C) values were obtained for the training set during model training (in light grey) and the external test set evaluation (in darker grey) for each QSAR model built using the three RF, SVM and NN algorithms for the different FP/descriptor sets optimised.

Testing different ML algorithms to build the same QSAR model allows to elaborate on which may be the most suitable candidate to describe the properties under exploration. As distinct algorithms employ different mathematical architectures, the best option for each QSAR prediction task greatly depends on the problem at hand and on the input dataset supplied for training, requiring a thorough experimentation stage to optimise and fine-tune the ideal hyperparameters to achieve the most accurate performance.

The exercise of evaluating and comparing train and test set performance revealed that, as expected, the MCC metric tends to decrease from the cross-validation estimation made for the training set to the external test set prediction. Another noticeable difference occurs for the SE metric: Whilst the SP metric, which measures the accuracy of the model in identifying a “negative” class compound is essentially maintained in the test set (**Figure 25A**), the sensitivity metric describing the ability for the models to correctly predict the positive class is greatly inferior for the test set. This observation suggests that the models may probably be slightly overtrained.

In few words, overtraining, or overfitting, happens when the ML algorithm fits the training set data too closely and, instead of learning how to generalise the fundamental structure-property relationships described by the dataset, it limits itself to memorise the training molecules. When this occurs, the model will still be able to satisfactorily predict the categorical classes for the training set, as seen by the values obtained in **Table 6**, although the algorithm's accuracy for the external test set is worst, as it struggles to generalise for the new variations that the unseen test instances present²²⁷. This may explain why the MCC values decrease from train to test set. Note that, as illustrated in **Figure 25B and C**, the loss on performance is more accentuated for the Neural Network models, as overtraining is reported to be an especially prominent problem for this type of ML algorithm^{136,228}.

Once more, the great imbalance in both categorical classes is an underlying factor in model performance: The learning ability of the models is much more oriented towards correctly predicting the negative class over the positive. Additionally, the fact that most positive input molecules are sourced from the BIT

HTS bioassay (almost 87% of the total “positive” compounds number) may be contributing to make the “positive” class prediction to be biased to a more specific subset of chemical structures and scaffolds. Ideally, to ameliorate this lack of diversity and representation of the “positive class” in the chemical space of the model, more positive templates should be introduced in the dataset, although none are publicly available as seen during the literature search and dataset curation stages – yet again reminding the fundamental hurdle in rare disease drug development: lack of data¹³³.

Another noteworthy aspect to be taken from this analysis is that, regardless of the ML algorithm and descriptor type used, model building was deeply conditioned by the effort of iteratively balancing the negative-positive class weight to compensate for the disparities in class representation. There are other techniques that can be used to address this issue, such as over-sampling, which consists on increasing the number of instances in the minority class by generating duplicates or new synthetic molecules extrapolated from the input instances, and under-sampling, based on randomly removing majority class molecules to reduce the instance amount in hopes of balancing out classes²²⁹. In the context of this Master’s thesis, given the limitations on the dataset and the fact that under-sampling would reduce even more the chemical space of the model, these methods were not explored and class balancing was resumed to tweaking the hyperparameters included in each ML algorithm.

3.1.2.6 Applicability domain

Finally, after building the model, its domain of applicability was determined. As discussed earlier, the applicability domain of a model denotes the range of compounds for which an accurate prevision can be confidently obtained. Since the generalisation capacity of the algorithm is intrinsically limited to the diversity (and quality) of the training structures, delineating the applicability domain allows to prevent computing an inaccurate property prediction for compounds outside this domain during the drug screen up head.

To determine the domain, “template” compounds were selected from the total dataset (**Figure 26**). It was desired for these compounds of interest to be experimentally-verified “positive” leads, meaning that they can establish an interaction with PMM2. Following a logical rationale, the four hit compounds reported by Yuste-Checa and colleagues, identified in the figure below as A), B), C) and D), were selected as positive templates, alongside training set compound E), 2-(3-chloro-4-fluorophenyl)-4,6-dinitro-2,3-dihydro-1,2-benzothiazol-3-one, originated from the BIT HTS bioassay, which was chosen for being the TP-predicted compound with the highest prediction probability value obtained in the final QSAR Model 1.

It is important to keep in mind that the model not only serves to predict “positive” class compounds but also “negative” ones. For this reason, negative templates were also included. Chosen negative templates were the MyriaScreenI library compound F), 2,6-di-tert-butyl-4-(4-hydroxybutan-2-yl)phenol, and the BIT HTS bioassay compound G), 3-(3-bromophenyl)-4,5-dihydro-1H-1,2,4-triazole-5-thione. These two molecules were selected for being the TN-predicted compounds with the highest prediction probability value obtained in the final QSAR Model 1 from the two different sources used to build the dataset.

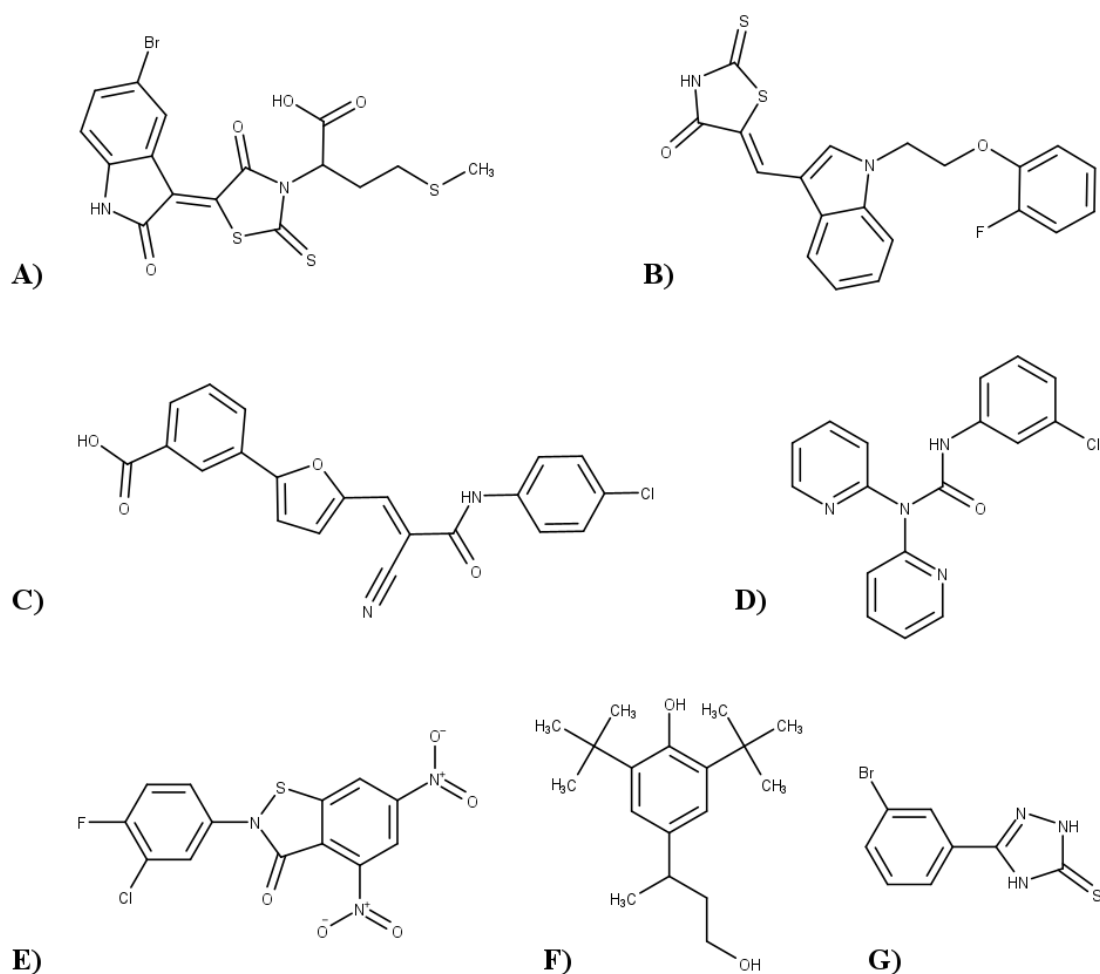


Figure 26. Template compounds used for calculation of the model's applicability domain. Positive templates **A)** 2-[(5Z)-5-(5-bromo-2-oxo-1,2-dihydro-3H-indol-3-ylidene)-4-oxo-2-thioxo-1,3-thiazolidin-3-yl]-4-(methylsulfanyl)butanoic acid; **B)** 5-[[1-[2-(2-fluorophenoxy)ethyl]indol-3-yl]methylidene]-2-sulfanylidene-1,3-thiazolidin-4-one; **C)** 3-(5-{(1E)-3-[(4-chlorophenyl)amino]-2-cyano-3-oxoprop-1-en-1-yl}furan-2-yl)benzoic acid; **D)** 1-(3-chlorophenyl)-3,3-bis(pyridine-2-yl)urea; **E)** 2-(3-chloro-4-fluorophenyl)-4,6-dinitro-2,3-dihydro-1,2-benzothiazol-3-one and negative templates **F)** 2,6-di-tert-butyl-4-(4-hydroxybutan-2-yl)phenol and **G)** 3-(3-bromophenyl)-4,5-dihydro-1H-1,2,4-triazole-5-thione.

Having defined the template molecules, the next step consisted of calculating the Tanimoto coefficient (TC) in Keras for the entire dataset in relation to each template. As previously developed, the TC is a chemical similarity metric used to compare the structural resemblance shared between query and template molecules by calculating the ratio of the number of features present in both molecules to the number of features present in either molecule. The features that TC calculation observes are provided by certain Fingerprints. In three-dimensional TC calculations, E3FP (Extended 3-Point Fingerprint) and Morgan FPs were employed. The attributes within the E3FP encode a set of pharmacophore fragments consisting of three points and Morgan FPs denote the topology of a molecule by generating a set of circular substructures around a radius r from each atom.

After calculating the two- and three- dimensional TC for the whole dataset, the last step for setting the applicability domain was to define a TC value threshold for each template compound. This means that, by analysing the similarity scores obtained, a cut-off value is set, below which a query molecule is considered to fall out of the applicability domain.

A threshold value was described for each positive and negative template by examining the average and the maximum value of the 2D and 3D TC similarity scores obtained for the FN, FP, TN and TP compounds.

For the positive templates, the threshold was set below the maximum TP score and above the maximum TN, FP and FN scores to favour the inclusion of compounds resembling the positive instances of the model. **Table 8** below showcases the cut-off values that were defined for each template for the different similarity scores.

Table 8. Tanimoto Coefficient threshold values calculated for the QSAR model’s dataset compounds defined for each one of the positive and negative template molecules.

Template compounds		A)	B)	C)	D)	E)	F)	G)
2D TC		0.50	0.42	0.35	0.26	0.43	0.25	0.29
3D TC	E3FP	0.14	0.14	0.13	0.09	0.26	0.13	0.26
	Morgan	0.16	0.18	0.12	0.17	0.42	0.13	0.41

The applicability domain of Model 1 will also be later applied for Model 2, as both of them were generated from the same dataset.

3.1.3 QSAR regression Model 2 for IC₅₀ estimation

As mentioned in **Section 1.5**, Model 2 is a QSAR regression model trained with the intent to predict a theoretical pIC₅₀ value towards hPMM2 inhibition for every submitted compound.

But then again, why the IC₅₀?

The half maximal inhibitory concentration, IC₅₀, is the concentration of a certain compound that is required to achieve a 50% inhibition effect on the response (activity) of the treated biochemical target, under the assay’s experimental conditions. This value is widely applied to measure the effectiveness of a molecule in inhibiting a target, describing the potency of the non-covalent interactions that are established between the molecule and its target.

In a drug research and development scenario, the IC₅₀ is a precious tool to understand a drug’s efficacy towards a pathological target, which is generally a protein or protein complex.

Besides this, other reasons behind why the IC₅₀ is such a good mean for exploring the extent of a protein-ligand interaction include: The ability to compare the values obtained for different ligands targeting the

same protein, allowing to rank their effectiveness in an on-target lead optimisation strategy; The fact that, by being a value of concentration (normally reported in molarity), the IC_{50} indicates if the inhibitory effect that is observed happens at a biologically relevant concentration, i.e., a concentration that is amenable to be used in the organism to produce the intended therapeutic effect with the least adverse effects possible. If a compound is too weak to inhibit the target at such concentration ranges it most likely will not be a good drug candidate^{178,210,230,231}.

Overall, the IC_{50} is a reliable measure to study protein-ligand interactions, being of the utmost importance in drug R&D for lead optimisation and can be used, just like in the project, as input data to create QSAR capable of predicting an IC_{50} estimation for novel compounds towards hPMM2.

3.1.3.1 Calculation of Molecular Descriptors

Just as before, the FPs and 1D&2D descriptor set contained in PaDEL-Descriptor were successfully calculated, alongside the three-dimensional descriptors desired for this model (the autocorrelation 3D; CPSA; GI; LOB; MOMI; PSI and WHIM) for each instance in the dataset and RDF descriptors were generated separately.

Table 9. Brief descriptive summary of some of the molecular three-dimensional descriptors available for calculation in PaDEL-Descriptor that were selected to be utilised in this study.

type	Description
Autocorrelation 3D	The autocorrelation 3D descriptors relate to the spatial distribution of generic molecular properties in the input compound and measure the strength of a relationship between atoms at a predefined distance, that is, the descriptor calculates the correlation between the properties of a given atom and the properties of neighbouring atoms. This calculation is based on Euclidean interatomic distances. The 80 attributes contained within this descriptor type encode for various different properties like atomic charges, atomic radius and molecular weight, amongst others ¹¹² .
CPSA	The Charge Polar Surface Area 3D descriptors are designed to represent the three-dimensional structure of the input molecule through the electrostatic potential at its surface. It comprises 29 individual attributes that encode information about the chemical features involved in polar intermolecular interactions and the electronic charge distribution on the surface in the 3D space ²³² .
Gravitational Index	The Gravitational Index descriptors encode information about molecular structure by taking the notion of gravitational potential energy, which measures the energy of an atom in a gravitational field. The nine features described by this Gravitational Index are generated by considering the distance between a pair of atoms, then the potential energy between the pair is calculated through a gravitational formula and the sum of the potential energy values for all the pairs of atoms established in the input molecule is used to generate the final output ²³³ .
Length Over Breadth	The Length over Breadth descriptor represents, for a certain instance, the relationship between the longest and shortest dimensions of such molecule, inferring on the degree of elongation or compactness of the compound. The LOB descriptor encompasses only two features: the length of the longest dimension in the input molecule and the length of the shortest dimension, which can both be generated employing different calculation methods (for example, using the principal moments of inertia or the radius of gyration of the compound).

Moment of Inertia	The Moment of Inertia (MOMI) descriptor denotes the rotational motion and the distribution of mass within the input molecule to calculate the resistance offered by the molecule to the rotational motion around its principal axes, which are defined by its seven features (MOMI-x, MOMI-y, MOMI-z, MOMI-xy, MOMI-xz, MOMI-yz, MOMI-r), based on the distribution of mass to ultimately provide information on the shape and symmetry of the instance.
Petitjean Shape Index	The Petitjean Shape Index descriptor, designated as PSI, is calculated by considering molecular convexity and distribution of distances between pairs of atoms in the input molecule to describe the deviation of the overall molecule's shape from a perfect sphere. The final PSI descriptor englobes information of its three features, the geometrical radius, geometrical diameter and geometrical shape, generated by dividing the surface area of the instance by the surface area of a perfect reference of equal volume ²³⁴ .
WHIM	The Weighted Holistic Invariant Molecular descriptor combines information on the topological and three-dimensional structure of an input compound without taking into consideration the connectivity amongst atom pairs, it only accounts for the atoms' positioning in the 3D space. The WHIM descriptor is generated by a set of 91 features that encode geometric and topological properties obtained using different weighting schemes to reveal information on size, shape, symmetry, and atomic property distribution, being also invariant to alterations in the orientation and translation of the molecule ¹¹² .
RDF	The Radial Distribution Function (RDF) descriptor considers an input molecule containing a defined number of n atoms placed in a three-dimensional space divided into various spherical shells, where the probability distribution of finding an atom in a spherical volume at a given distance of radius r from another reference atom is the RDF function. This descriptor set encodes 210 individual attributes and it can be calculated for a single atom or the entire input compound to supply information about interatomic distances, bond distance, ring types, amongst other aspects ²³⁵ .

3.1.3.2 Partitioning the dataset

For implementing the train and test set partitions, only the compounds from the dataset with a defined IC_{50} value were considered. This corresponded to a total of 56 compounds with reported pIC_{50} values for hPMM2 that were divided into a training set of 51 instances with pIC_{50} values ranging from 6.172 (0.672 μM) to 3.547 (283.7 μM), and a test of five compounds, with pIC_{50} values between 4.054 (88.3 μM) and 4.257 (55.3 μM).

3.1.3.3 Trial models to optimise algorithm parameters and select input descriptors

Following the same rationale as applied in Model 1, model building was initiated by training regression RF models in WEKA with the goal of selecting the best-performing circular and fragment FPs, as well as the best 3D descriptor sub-type, to build the try-out models.

The evaluation metrics used to assess performance in regression models were the Correlation Coefficient (R^2), the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) and, moreover, WEKA also provides the Relative Absolute Error (RAE) and the Root Relative Square Error (RRSE) metrics. RAE measures the difference between the experimental pIC_{50} values and the predicted values, normalized by the variability observed within the experimental values, the lower the RAE is, the better the predictive accuracy of the model will be. In its turn, the RRSE evaluates the difference between the experimental values of the training set and the predicted values for the same instances normalized by

the range of predicted values. The lower the RRSE, the better the predictive accuracy is. Both these metrics are expressed in percentages as the **Equations 13 and 14** below indicate.

$$Relative\ Absolute\ Error\ (RAE, \%) = \frac{MAE}{\frac{1}{n} \times \sum_{i=1}^n |y_i - \bar{y}_i|} \times 100$$

13

$$Relative\ Root\ Square\ Error\ (RRSE, \%) = \frac{RMSE}{\max y_i - \min y_i} \times 100$$

14

Where y_i is the experimental activity value, $\bar{y}_i$ is the average of experimental values, x_i is the predicted variable value, n is the number of instances in the dataset and $\max y_i$ and $\min y_i$ represent the maximum and minimum experimental values observed in the dataset, respectively.

In a starting point, preliminary RF models were trained using the optimised default parameters for the algorithm and utilising each one of seven Fingerprints and the whole 1D&2D descriptor set, plus the eight 3D descriptors previously calculated with all of their individual attributes, as shown below in **Table 10**.

Table 10. Analysis of preliminary RF QSAR models to evaluate the predictive performance of each FP and 3D descriptor type. Trial models were trained in WEKA for the circular Fingerprinter, Extended FPs and Graph Only FPs; the fragment-based PubChem FPs, MACCS FPs, Substructure FPs, the entire 1D&2D set and each one of the eight 3D descriptors sets (autocorrelation; CPSA; GI; LOB; MOMI; PSI and WHIM) using all attributes. RF parameters were kept as the predefined optimal parameters. Bold lines indicate the best-performing model for each FP/descriptor according to the evaluation metrics generated during training.

type	FP/Descriptors (no. of attributes)	R ^a	MAE ^b	RMSE ^c	RAE (%) ^d	RRSE (%) ^e
Circular FPs	Fingerprinter (1024)	0.417	0.379	0.513	80.089	93.393
	Extended FP (1024)	0.403	0.382	0.504	78.657	87.836
	Graph Only FP (1024)	0.407	0.381	0.536	80.469	89.275
Fragment FPs	PubChem FP (881)	0.429	0.375	0.521	79.178	90.647
	MACCS FP (166)	0.401	0.389	0.526	82.147	91.651
	EState FP (79)	0.486	0.347	0.497	73.302	86.565
	Substructure FP (307)	0.479	0.360	0.494	76.157	86.976
	1D&2D (1442)	0.476	0.391	0.499	82.654	86.923
3D	Autocorrelation 3D (80)	0.504	0.397	0.490	83.823	83.356
	CPSA (29)	0.166	0.454	0.581	95.918	101.122**
	Gravitational Index (9)	0.117	0.483	0.613	102.073*	106.804**
	Length Over Breadth (2)	0.235	0.481	0.606	101.690*	105.586**
	Moment of Inertia (7)	0.080	0.502	0.613	106.132*	106.778**

Petitjean Shape Index (3)	0.153	0.475	0.614	100.454*	106.884**
RDF (210)	0.305	0.433	0.549	91.472	91.515
WHIM (91)	0.127	0.481	0.584	101.656*	101.634**

a) R = Pearson's correlation coefficient. b) MAE = mean absolute error. c) RMSE = root mean squared error. d) RAE = relative absolute error. e) RRSE = relative root squared error. * A RAE value > 100% means that the MAE of the predicted pIC₅₀ values is larger than the mean absolute deviation of the experimental data. ** A RRSE value > 100% means that the RMSE of the predicted pIC₅₀ values is larger than the range of the experimental data.

Examination of the results allowed to select the best-performing models for the circular and fragment-based Fingerprint types: The CDK Fingerprinter with a $R^2 = 0.417$ and the EState FP model with a $R^2 = 0.486$, respectively. The 1D&2D model yielded an $R^2 = 0.476$.

From these initial outcomes it is important to highlight that the error obtained for the models is generally very high, indicating a poor prediction capacity. Next, in order to further carry out the optimisation process, the influence of adding supplementary three-dimensional information to Fingerprints and 1D&2D descriptors was explored by training different try-out models using bundles of Fingerprints + 3D descriptors or 1D&2D + 3D descriptors as input for the RF algorithm, to infer on the improvement that these combinations may have on the predictive performance. The results are shown in **Figure 27** and **Table 11**.

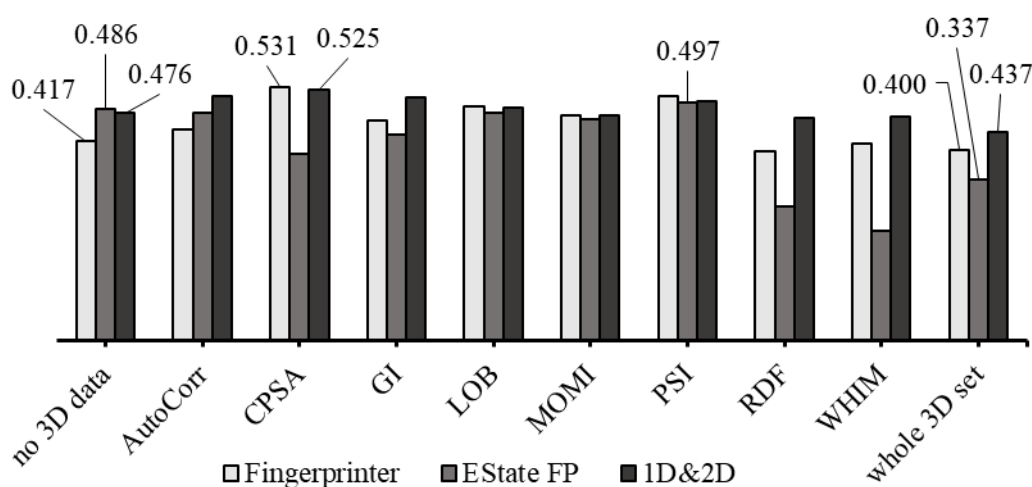


Figure 27. Bar chart representing the Correlation Coefficient for trial models built with the addition of different 3D descriptors. This figure illustrates the results reported in Table 11 below, showcasing the R value obtained for each model built with the Fingerprinter FP set (light grey), the EState FP set (medium grey) and the 1D&2D set (dark grey) adding each 3D descriptor and also the entire 3D set. Indicated values represent the model with the highest R value reported and the models built without the 3D set and with the whole set for comparison.

Table 11. Performance evaluation for RF QSAR models built with different FP/descriptor bundles. Trial models were trained in WEKA using different bundles of the best-performing Fingerprinter, EState FPs and 1D&2D descriptors sets with each one of the eight 3D descriptors (autocorrelation; CPSA; GI; LOB; MOMI; PSI and WHIM) and the entire 3D set. RF parameters were kept as the predefined settings. Bold lines indicate the best-performing model for each FP/descriptor + 3D descriptor bundle, according to the evaluation metrics generated for training.

3D descriptor added to bundle (no. of attributes)	R ^a	MAE ^b	RMSE ^c	RAE (%) ^d	RRSE (%) ^e
Fingerprinter + 3D descriptors bundle					
Autocorrelation 3D (80)	0.442	0.375	0.510	79.329	88.773
CPSA (29)	0.531	0.358	0.480	75.587	83.593
Gravitational Index (9)	0.460	0.375	0.506	79.345	88.165
Length Over Breadth (2)	0.490	0.364	0.496	76.867	86.312
Moment of Inertia (7)	0.473	0.369	0.502	77.993	87.499
Petitjean Shape Index (3)	0.511	0.363	0.487	76.677	84.834
RDF (210)	0.396	0.406	0.523	85.921	91.132
WHIM (91)	0.413	0.391	0.518	82.657	90.129
All 3D set (431)	0.400	0.399	0.520	84.395	90.578
EState FP + 3D descriptors bundle					
Autocorrelation 3D (80)	0.478	0.413	0.498	87.411	86.790
CPSA (29)	0.390	0.395	0.522	83.539	90.288
Gravitational Index (9)	0.431	0.392	0.518	82.875	90.137
Length Over Breadth (2)	0.478	0.382	0.499	80.679	86.829
Moment of Inertia (7)	0.462	0.395	0.503	83.480	87.540
Petitjean Shape Index (3)	0.497	0.374	0.491	79.130	85.585
RDF (210)	0.281	0.437	0.552	92.322	96.077
WHIM (91)	0.231	0.451	0.562	95.388	97.862
All 3D set (431)	0.337	0.430	0.536	90.994	90.329
1D&2D descriptors + 3D descriptors bundle					
Autocorrelation 3D (80)	0.512	0.382	0.488	80.818	84.970
CPSA (29)	0.525	0.383	0.484	80.965	84.254
Gravitational Index (9)	0.510	0.384	0.488	81.226	84.943
Length Over Breadth (2)	0.489	0.389	0.495	82.262	86.241
Moment of Inertia (7)	0.473	0.392	0.500	82.796	87.127
Petitjean Shape Index (3)	0.501	0.378	0.491	79.867	85.456
RDF (210)	0.467	0.396	0.503	83.708	87.533
WHIM (91)	0.468	0.399	0.501	84.247	87.311
All 3D set (431)	0.437	0.399	0.511	84.448	88.951

a) R = Pearson's correlation coefficient. b) MAE = mean absolute error. c) RMSE = root mean squared error. d) RAE = relative absolute error. e) RRSE = relative root squared error.

This exercise allowed to define which bundles yielded the best predictive performance. By analysing **Table 11** it is seen that implementing three-dimensional information in the models can effectively improve predictive ability. The Correlation Coefficient was increased the most for the Fingerprinter set

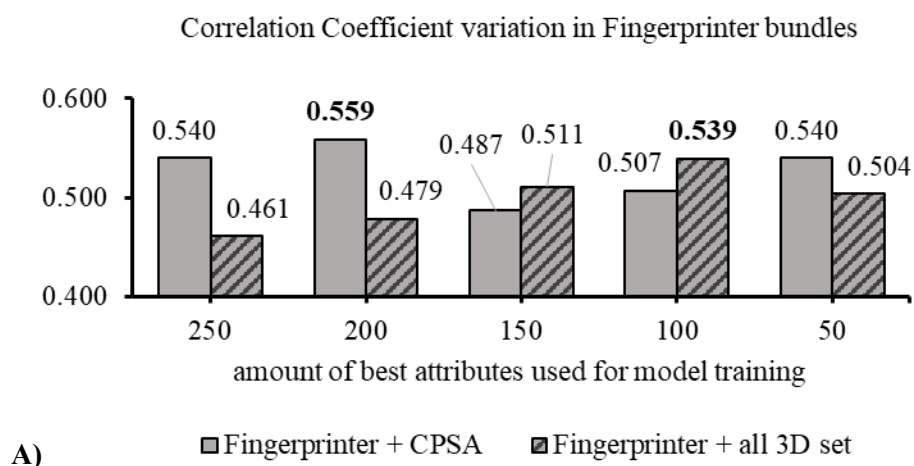
when using the CPSA descriptor in the bundle, for the EState set using the PSI descriptors in the bundle and for the 1D&2D set with the CPSA as well.

The error metrics MAE and RMSE were also comparably enhanced and the relative errors RAE and RRSE, despite being still high values, are all <100% which proved that using 3D descriptors to enrich the learning capabilities of the model is a good strategy.

It is noteworthy that, whilst helpful for complementing the FPs and 1D&2D descriptors at better denoting the chemical information required for training, choosing fitting 3D descriptors instead of using the entirety of the set available is an important task. As it is seen in **Figure 27**, the Correlation Coefficient for models trained with bundles that contained the whole 3D set is actually slightly lower (worst performance) than models with no three-dimensional data at all. This suggests that adding unfiltered, unselected 3D data can contribute to a model that ends up being too complex, and so, it is preferable to opt for a single 3D descriptor that complements nicely the input data. Nevertheless, it would be interesting to explore if by further optimising the models it would be possible to improve the performance of bundles with the entire 3D set. This was taken into consideration for the next step.

3.1.3.4 Descriptor attribute selection

Individual attribute importance was calculated using the R program, as done during Model 1 training, for each of the defined bundles: Fingerprinter + CPSA; Fingerprinter + all 3D set; EState FP + PSI; EState FP + all 3D set; 1D&2D + CPSA; 1D&2D + all 3D set, using all attributes and then sets of the 250; 200; 150; 100 and 50 best features. Results are indicated below in **Table 12** and **Figure 28**.



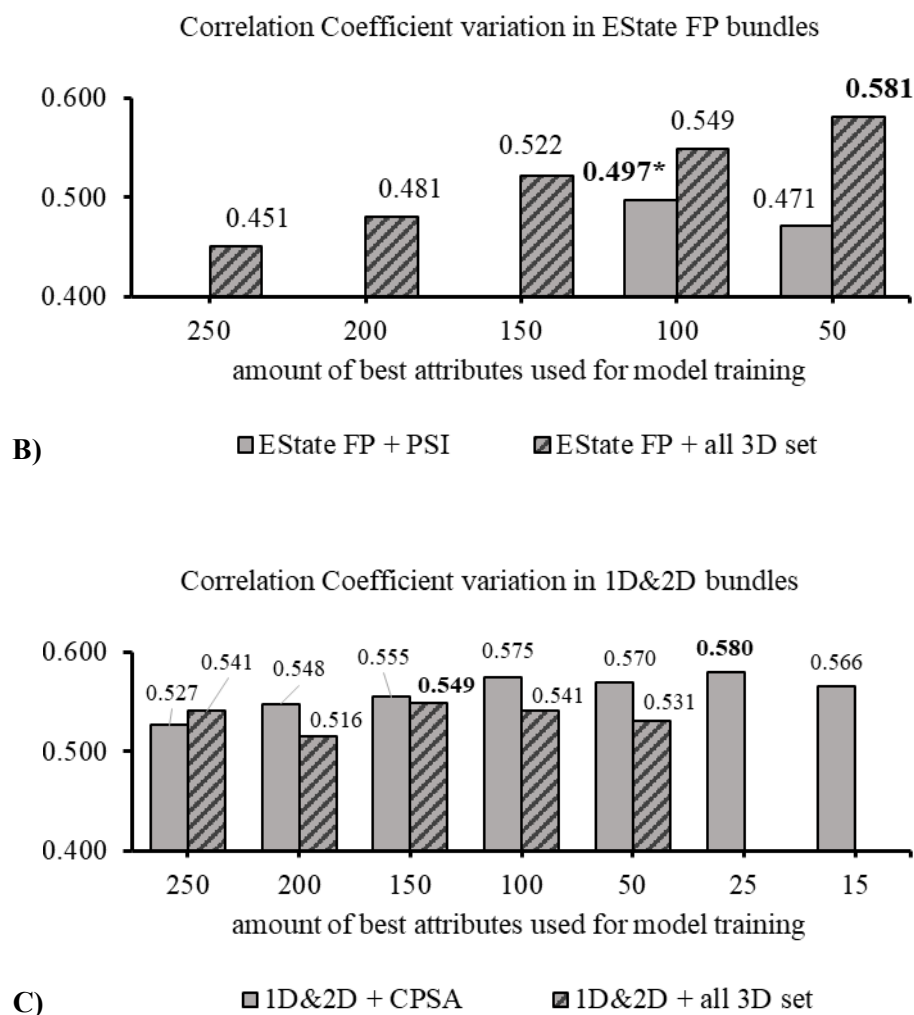


Figure 28. Bar chart for the variation in the Correlation Coefficient values for trial models built with the different defined FP/descriptor bundles. The graphs represent the R values obtained for the models described in Table 12 below. **A)** Model built with the Fingerprinter + CPSA bundle (grey) and Fingerprinter + all 3D set bundle (striped grey). **B)** Model built with the EState FP + PSI bundle (grey) and EState FP + all 3D set bundle (striped grey). The EState + PSI model marked with “*” was built with only 82 attributes, as this is the total number of features available for a combination of EState Fingerprints (79 features) and PSI descriptors (3 features), which also justifies why there are no models built with more attributes for this particular bundle. **C)** Model built with the 1D&2D + CPSA bundle (grey) and 1D&2D + all 3D set bundle (striped grey). The 1D&2D + CPSA models were trained further than the bundle using the entire 3D set with the 25 and 15 most importance weighted attributes. Bold values belong to the best-performing models.

Table 12. Analysis of trial models to assess how the predictive performance is affected by the number of attributes. Trial models were trained in WEKA for the defined bundles of Fingerprinter + CPSA; Fingerprinter + all 3D set; EState FP + PSI; EState FP + all 3D set; 1D&2D + CPSA and 1D&2D + all 3D set, built with all attributes, and then gradually downsizing to the 250, 200, 150, 100 and 50 best attributes as weighted by importance in the R software. Bold lines indicate the best-performing model for each set according to train evaluation metrics.

FP/Descriptor bundle (no. of attributes)	R ^a	MAE ^b	RMSE ^c	RAE (%) ^d	RRSE (%) ^e
Fingerprinter + 3D descriptors bundle					
Fingerprinter + CPSA (50)	0.540	0.331	0.483	70.024	84.092
Fingerprinter + CPSA (100)	0.507	0.352	0.495	74.442	86.116

Fingerprinter + CPSA (150)	0.487	0.365	0.499	77.220	86.945
Fingerprinter + CPSA (200)	0.559	0.334	0.470	70.535	81.837
Fingerprinter + CPSA (250)	0.540	0.342	0.477	72.258	83.158
Fingerprinter + all 3D set (50)	0.504	0.364	0.490	76.879	85.367
Fingerprinter + all 3D set (100)	0.539	0.356	0.477	75.294	83.037
Fingerprinter + all 3D set (150)	0.511	0.365	0.488	77.200	84.921
Fingerprinter + all 3D set (200)	0.479	0.381	0.497	80.451	86.518
Fingerprinter + all 3D set (250)	0.461	0.379	0.503	80.173	87.544
EState FP + 3D descriptors bundle					
EState FP + PSI (50)	0.471	0.377	0.502	79.722	87.352
EState FP + PSI (all, 82)	0.497	0.374	0.491	79.130	85.585
EState FP + all 3D set (50)	0.581	0.365	0.463	77.150	80.582
EState FP + all 3D set (100)	0.549	0.367	0.474	77.689	82.591
EState FP + all 3D set (150)	0.522	0.376	0.485	79.535	84.384
EState FP + all 3D set (200)	0.481	0.389	0.497	82.265	86.527
EState FP + all 3D set (250)	0.451	0.393	0.506	82.981	88.107
1D&2D descriptors + 3D descriptors bundle					
1D&2D + CPSA (15)	0.566	0.350	0.470	73.996	81.768
1D&2D + CPSA (25)	0.580	0.340	0.461	71.864	80.326
1D&2D + CPSA (50)	0.570	0.349	0.466	73.773	81.201
1D&2D + CPSA (100)	0.575	0.352	0.466	74.432	81.099
1D&2D + CPSA (150)	0.555	0.360	0.471	76.145	81.997
1D&2D + CPSA (200)	0.548	0.364	0.474	76.997	82.548
1D&2D + CPSA (250)	0.527	0.365	0.482	77.168	83.860
1D&2D + all 3D set (50)	0.531	0.364	0.480	77.029	83.557
1D&2D + all 3D set (100)	0.541	0.366	0.476	77.476	82.922
1D&2D + all 3D set (150)	0.549	0.355	0.473	75.079	82.374
1D&2D + all 3D set (200)	0.516	0.372	0.485	78.663	84.487
1D&2D + all 3D set (250)	0.541	0.369	0.477	78.051	83.123

a) R = Pearson's correlation coefficient. b) MAE = mean absolute error. c) RMSE = root mean squared error. d) RAE = relative absolute error. e) RRSE = relative root squared error.

These outcomes indicate that the best performance was obtained with the models produced using the Fingerprinter + CPSA bundle with the 200 best features ($R = 0.559$); the Fingerprinter + all 3D set bundle with the 100 best ($R = 0.539$); the EState FP + PSI bundle with all (82) features ($R = 0.497$); the EState FP + all 3D bundle with the 50 best ($R = 0.581$); the 1D&2D + CPSA bundle with the 25 best ($R = 0.580$) and the 1D&2D + all 3D bundle with the 150 best ($R = 0.549$).

When it comes to models built with the entire 3D descriptor set, it is possible to see that feature selection did in fact improve prediction metrics, with the Correlation Coefficient rising for most cases when compared to the initial, unselect models. Also, a small but consistent reduction in error metrics can be observed throughout this process. This concluded the preliminary optimisation round for Model 2.

During the final QSAR model building utilising the three different ML techniques (RF, SVM and NN), additional tweaking of the algorithms' parameters was performed, as it is discussed in the next section.

3.1.3.5 Finalised models built with RF, SVM and NN algorithms (and test set evaluation)

Final models were built using each algorithm as described before. For Random Forests, the algorithm's *mtry* and *no. of iterations* were adjusted during training to optimise results, with every other parameter kept with the default values. The final QSAR models and respective training set evaluation is presented in **Table 13**, with the best performance being achieved with the EState FP + all 3D set bundle, at an $R = 0.629$ (**Figure 29**). It is interesting to see that, despite initially producing worst results, using the whole 3D descriptor data can originate an accurate model once the proper fine-tuning is attained.

Table 13. Training set analysis of the Random Forest models. Each line in the table indicates the final models evaluation metrics generated during training in WEKA for each one of the FP/descriptor bundles attempted. Bold values indicate the best-performing model obtained.

Descriptor bundles	<i>no. of iterations</i> *	<i>mtry</i> **	R^a	MAE ^b	RMSE ^c	RAE ^d (%)	RRSE ^e (%)
Fingerprinter + CPSA	500	log ₂	0.576	0.334	0.463	70.535	80.622
Fingerprinter + all 3D set	100	√	0.545	0.363	0.475	76.771	82.745
EState FP + PSI	100	1/3	0.508	0.369	0.489	78.102	85.209
EState FP + all 3D set	100	1/3	0.629	0.350	0.454	72.999	77.075
1D&2D + CPSA	100	1/3	0.592	0.332	0.454	70.216	79.079
1D&2D + all 3D set	100	√	0.568	0.346	0.461	73.180	80.198

* The *no. of iterations* is the number of trees that were used for model training and it was varied from 100 to 500 trees.

** The *mtry* was set either as a third (1/3), the square root (√) or the log₂ of the total attribute number for each descriptor set.

For both these parameters, only the best results for each model using a different FP/descriptor bundle is indicated.

a) R = Pearson's correlation coefficient. b) MAE = mean absolute error. c) RMSE = root mean squared error. d) RAE = relative absolute error. e) RRSE = relative root squared error.

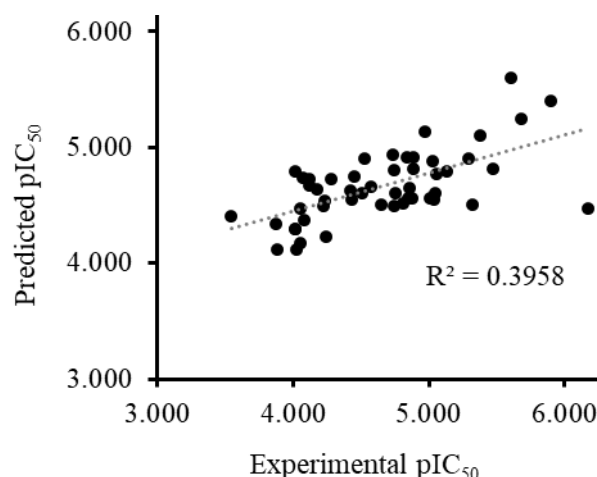


Figure 29. Scatter plot of the final RF model trained with the EState FP + all 3D set bundle. The graph plots the predicted pIC₅₀ for each instance as a function of their experimentally-measured values. The squared Pearson's Correlation Coefficient (R^2) is indicated.

Support Vector Machine models were also trained as before, optimising hyperparameters such as the *cost* and the kernel function. Finalised SVM models evaluated with the training set are indicated **Table 14**, with the 1D&2D descriptor + CPSA bundle yielding the best performance with an $R = 0.629$ (**Figure 30**), which was done with a radial basis kernel function and a *cost* of 1:10 negative-positive balance.

Table 14. Training set analysis of the Support Vector Machine models. Each line in the table indicates the final models evaluation metrics generated during training in WEKA for each one of the FP/descriptor bundles attempted. Bold values indicate the best-performing model obtained.

Descriptor bundle	Kernel function [*]	cost ^{**}	R ^a	MAE ^b	RMSE ^c	RAE ^d (%)	RRSE ^e (%)
Fingerprinter + CPSA	Linear	1	0.640	0.479	0.627	90.367	90.104
Fingerprinter + all 3D set	Radial basis	10	0.660	0.321	0.429	67.972	74.736
EState FP + PSI	Polinomial	1000	0.628	0.329	0.446	69.520	77.750
EState FP + all 3D set	Radial basis	1	0.555	0.347	0.473	73.434	82.359
1D&2D + CPSA	Radial basis	10	0.665	0.313	0.425	66.100	74.058
1D&2D + all 3D set	Polinomial	50	0.585	0.340	0.465	71.741	81.033

* The kernel function used by the SVM algorithm was set for each model, varying from linear, polynomial or radial basis.

** The *cost* parameter was tweaked for every model in order to find the class balance that provided the best results.

For all these parameters, only the best results for each model using a different FP/descriptor bundle are indicated in the table.

a) R = Pearson's correlation coefficient. b) MAE = mean absolute error. c) RMSE = root mean squared error. d) RAE = relative absolute error. e) RRSE = relative root squared error.

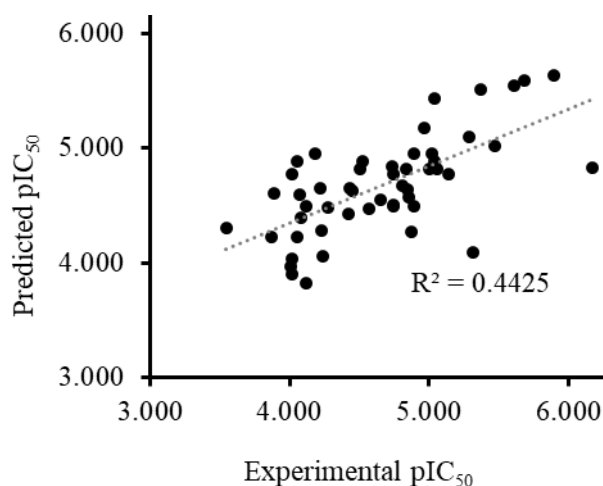


Figure 30. Scatter plot of the final SVM model trained with the 1D&2D descriptor + CPSA bundle. The graph plots the predicted pIC_{50} for each instance as a function of their experimentally-measured values. The squared Pearson's Correlation Coefficient (R^2) is indicated.

The final neural networks were built in Keras (**Table 15**) and the *no. of hidden layers* in the network; the *no. of neurons* and the *optimizer* algorithm were adjusted to provide accurate models, the best having been built with the 1D&2D descriptor + all 3D descriptor set bundle (**Figure 31**). This network made use of four layers (1 input + 3 hidden layers) with 150 neurons per layer and the Adamax *optimizer* to obtain a final $R^2 = 0.936$ ($R = 0.967$).

Table 15. Training set analysis of the Neural Network models. Each line in the table indicates the final models evaluation metrics generated during training in Keras for each one of the FP/descriptor bundles attempted. Bold values indicate the best-performing model obtained.

Descriptor bundle	<i>layers</i> [*]	<i>neurons</i> ^{**}	<i>optimizer</i> ^{***}	R^2 ^a	MAE ^b	RMSE ^c
Fingerprinter + CPSA	3	200	Adamax	0.929	0.103	0.159
Fingerprinter + all 3D set	5	100	Adam	0.793	0.183	0.267
EState FP + PSI	4	85	Adam	0.887	0.136	0.204
EState FP + all 3D set	4	50	Adam	0.667	0.246	0.356
1D&2D + CPSA	4	25	Adam	0.807	0.146	0.251
1D&2D + all 3D set	4	150	Adamax	0.936	0.096	0.156

* The *layers* parameter indicates the number of neuron layers that is settled to comprise the network.

** The *neurons* parameter denotes the number of neurons utilised for each layer in the network.

*** The *optimizer* algorithm was tried out using the options available in Keras, namely Adam, Adadelta, Adamax, Adagrad and SGD.

For all these parameters, only the best results for each model using a different FP/descriptor bundle are indicated in the table.

a) R^2 = squared Pearson's correlation coefficient. b) MAE = mean absolute error. c) RMSE = root mean squared error.

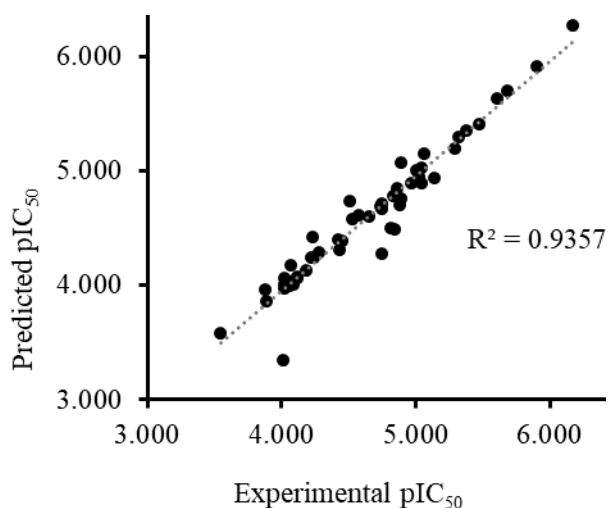


Figure 31. Scatter plot of the final NN model trained with the 1D&2D descriptor + all 3D set bundle. The graph plots the predicted pIC_{50} for each instance as a function of their experimentally-measured values. The squared Pearson's Correlation Coefficient (R^2) is indicated.

With the finalised models, external evaluation employing the defined test set was carried out. Due to the extremely limited total dataset available for this model (56 compounds), there were only five instances isolated to the test set and, moreover, one of these molecules was “compound VIII” (positive template D) which is known to interact with the protein but is thought not to wield any inhibitory effect, as no IC_{50} value could be registered. This meant that the test set encompassed a total of four mere molecules with a defined IC_{50} value. This is not a recommended situation and, once more, it is a tell-tale sign of this approach's major limitation – the amount of chemical data available that has been assayed towards PMM2. Yet, the test set, which is illustrated next in **Figure 32**, was submitted to the models to obtain a pIC_{50} prevision.

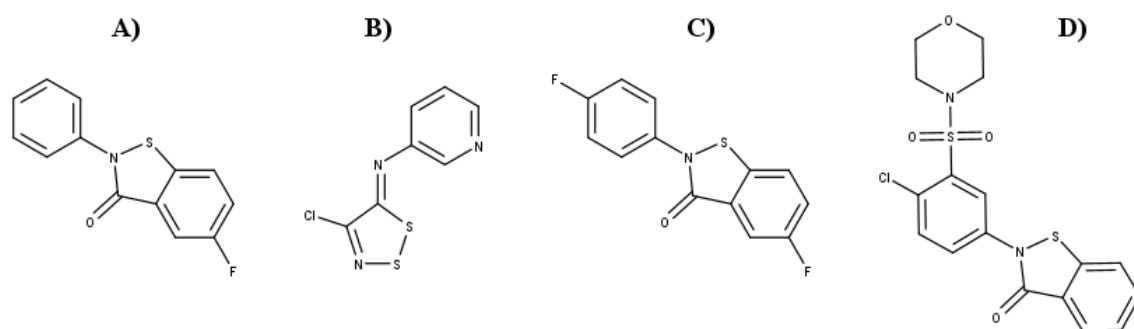


Figure 32. Testing set instances. **A)** 5-fluoro-2-phenyl-1,2-benzothiazol-3-one. **B)** 4-chloro-N-pyridin-3-ylthiazol-5-imine. **C)** 5-fluoro-2-(4-fluorophenyl)-1,2-benzothiazol-3-one. **D)** 2-(4-chloro-3-morpholin-4-ylsulfonylphenyl)-1,2-benzothiazol-3-one.

Table 16. Final test set evaluation of the QSAR regression models trained with different ML algorithms. The performance and prediction capability of the finalised models trained using Random Forest, Support Vector Machine, and Neural Networks were evaluated with the external test set. Each line in the table indicates one of the four test set instances (*vide* **Figure 32**) with its experimental pIC₅₀ value; the predicted pIC₅₀ value obtained with each best-performing model from the three ML algorithms (RF, SVM and NN); the quadratic error and the absolute error.

Test set compound	Exp. pIC ₅₀	Predicted pIC ₅₀			Quadratic Error ^a			Absolute Error ^b		
		RF	SVM	NN	RF	SVM	NN	RF	SVM	NN
A)	4.054	4.61	4.142	4.172	0.309	0.008	0.014	0.556	0.088	0.118
B)	4.182	4.327	3.691	4.349	0.021	0.241	0.028	0.145	0.491	0.167
C)	4.061	4.834	4.856	4.882	0.598	0.632	0.674	0.773	0.795	0.821
D)	4.257	4.522	4.945	4.608	0.070	0.473	0.123	0.265	0.688	0.351

a) The quadratic error is given by the squared difference of the experimental and predicted pIC₅₀: (exp. pIC₅₀ – pred. pIC₅₀)²

b) The absolute error is the modulus of the difference of the experimental and predicted pIC₅₀: |exp. pIC₅₀ – pred. pIC₅₀|

A theoretical pIC₅₀ value was generated in the RF, SVM and NN models for the four compounds and, since the test sample size was insufficient to correctly calculate a Pearson's correlation coefficient and the error metrics discussed, the quadratic error and the absolute error, alongside direct comparison between experimental and predicted values, were utilised to assess prevision quality (**Table 16**).

Whilst there is not enough data to observe a clear pattern in the performance of any model, analysis revealed some deviations from the experimental values: In the best predictions attained, compound A), which had an experimental IC₅₀ of 88.3 µM (pIC₅₀ = 4.054), got a score of 4.142 (pred. IC₅₀ = 72.1 µM); compound B), with an experimental IC₅₀ of 65.8 µM (pIC₅₀ = 4.182), was predicted a pIC₅₀ of 4.327 (pred. IC₅₀ = 47.1 µM); compound C), which reported an experimental IC₅₀ of 86.9 µM (pIC₅₀ = 4.061), was predicted a score of 4.834 (pred. IC₅₀ = 14.7 µM); and compound D), which had an experimental IC₅₀ of 55.3 µM (pIC₅₀ = 4.257), got a predicted pIC₅₀ of 4.522 (pred. IC₅₀ = 30.1 µM).

Despite most predictions being of the same order of magnitude as the experimentally-reported values, a general lack of precision is observed in the results. Test compound C) seems to be especially miss-predicted, yielding the highest error.

A pIC₅₀ estimation was also obtained for compound VIII: The RF, SVM and NN models predicted it to be 4.721 (IC₅₀ = 19.1 µM), 3.236 (IC₅₀ = 580.5 µM) and 2.924 (IC₅₀ = 1191.2 µM ≈ 1.19 mM), respectively. It does not seem to exist a consensus on the predicted IC₅₀ across models.

Given that compound VIII was not observed in literature to yield any sort of inhibition upon PMM2 activity and that the maximum concentration that was assayed was at 2.5 mM⁸⁷, it would be safe to assume that the expected prediction for this compound should be placed around a higher IC₅₀ value, as the more concentration is required to reach the half maximal inhibitory effect point, the least potent the inhibitor is. Following this logic, the 1.19 mM IC₅₀ calculated by the RF would be the most probable,

although, if it were to be accurate, the previous *in vitro* experimentation should have been able to detect it, as it is contained within the assay's concentration range.

All in all, Model 2 seems to be limited by the same factors as Model 1. While the training stage obtained promising results, even with Correlation Coefficients as high as $R = 0.9673$ for the neural network model, the external evaluation does not perform so well. It can be suggested that the model is again overfitted, as it predicts relatively well for the training set but fails to establish the structure-property relationships that would allow it to accurately generalise for external sets.

In the next section, a virtual drug screen was conducted in hopes of identifying potential drugs with a chaperoning effect on PMM2. Even with the models being subpar to what would be the ideal conditions of performance, their application in the drug screen was still carried out to explore this approach and to further study the models as a learning tool in the scope of this Master's project.

3.1.4 Virtual screening of approved drugs

Firstly, to conduct the virtual screen of approved drugs towards hPMM2, a screening dataset had to be assembled. This was done by performing a thorough search in ZINC database, setting up the *Bioactive and Drug*, plus the *for sale* filters in the search. These were employed purely for the practicality of this strategy: Whilst a wider and potentially more diverse universe of compounds could be screened for hPMM2, restricting the screening space to already approved drugs converts this task into a repurposing effort as discussed earlier. Additionally, limiting the ZINC outcomes to purchasable compounds is an efficient way of assuring that it is possible to acquire the identified hits, in case that such drugs yield interesting results with prospects of proceeding to *in vitro* biological assays.

During curation of the screen dataset, molecules with a molecular weight over 500 Da were discarded. This action was based on the Lipinski rule of five, which offers a set of practical guidelines widely used in drug development to evaluate the drug-likeness of a potential lead. According to one of its rules, a molecular weight of 500 Da is generally an adequate cut-off value, since larger molecules show tendency to have a lower bioavailability and therefore limited absorption²³⁶.

The total dataset amounted to 3306 unique molecules that were submitted to Model 1 to obtain a class prediction regarding the interaction effect of each instance in PMM2.

Unfortunately, from the entire screen set, there were only two compounds that were predicted as "positive" instances: Two drugs designated as ipsapirone and ebselen, shown in **Figure 33**. Both molecules exhibit a BIT scaffold, with ipsapirone containing a sulfone moiety at the S1 and an n-butyl substituent at N2 connected to a piperazin-1-ylpyrimidine group. Ipsapirone belongs to the azapirone family of selective 5-HT_{1A} receptor partial agonists, which are reported to have an antidepressant effect, being investigated for its application in generalized anxiety disorders²³⁷. In ebselen, the S1 is replaced by a selenium atom and the N2 contains a phenyl substituent. Ebselen is an organoselenium compound with

anti-inflammatory and anti-oxidant properties, being studied for its applications in the treatment of brain ischemia and strokes, amongst others²³⁸. Ebselen had been included in the dataset used for model construction, since it had been assayed in the HTS study described by Dahl *et al*²⁰⁸. Luckily, it was used for the external test set evaluation and, therefore, the model was not trained with this compound. However, as an IC₅₀ prediction had been reported in the literature (IC₅₀ for hPMM2 = 0.672 μ M), ebselen was used to train the QSAR regression Model 2 and, as so, it was not further explored. The fact that only two compounds were identified as “positives”, and that both contain the BIT scaffold, can be acknowledged as another indicator of the model’s poor predictive ability and bias, as discussed earlier.

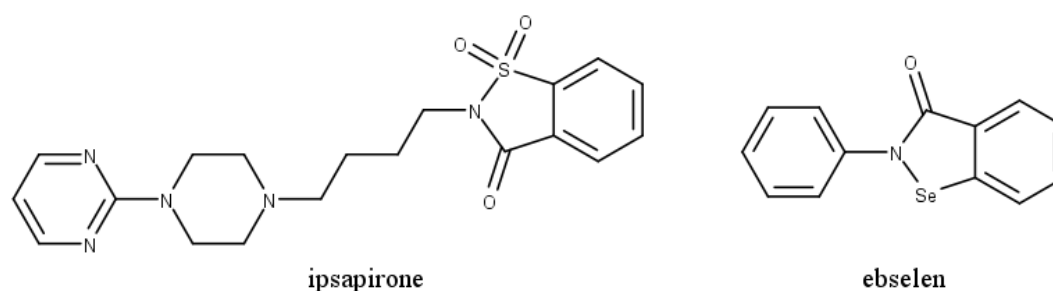


Figure 33. Ipsapirone (1,1-dioxo-2-[4-(4-pyrimidin-2-ylpiperazin-1-yl)butyl]-1,2-benzothiazol-3-one) and Ebselen (2-phenyl-1,2-benzoselenazol-3-one).

In parallel to Model 1 prediction calculations, the drug dataset underwent a filtering process by applying the established applicability domain to the set. For a drug to be deemed as falling within the QSAR models’ domain of application and continue to the regression Model 2, it must pass a majority of template thresholds, *i.e.*, by calculating the two- and three- dimensional Tanimoto coefficient for a certain drug towards each template, its similarity score must be equal or greater than the threshold that was established in **Section 3.1.2**.

Tanimoto calculations went accordingly and, at the end, 707 unique compounds passed the majority (at least more than half of templates) of templates for both 2D and 3D TCs, with 40 molecules being able to pass all templates.

The 707 drug passes were submitted to the regression QSAR Model 2, including ipsapirone, and an estimation to the pIC₅₀ value for each ligand was predicted.

3.1.5 Molecular Docking studies

To elaborate on the potential ligand-receptor interactions established between hPMM2 and selected drugs, a molecular docking examination was conducted in AutoDock Vina. This strategy was based on performing flexible docking to obtain an affinity score for each drug candidate and a binding pose at hPMM2. The compounds with the most promising results were intended to be acquired to carry on with *in vitro* biological experimentation, examining their effects on a PMM2-CDG disease model.

An important consideration had to be made at this stage: Over seven hundred molecules passed the applicability domain of the model and a pIC_{50} estimation was calculated for every single one. Performing a docking analysis for such a vast number of compounds would result in a very time-consuming and demanding endeavour in terms of computational processing power and would deem the posterior analysis of results much more complicated and exhaustive. A decision was made to downsize the number of ligands to be docked by applying the following selection criteria:

- The seven template compounds used to define the models' applicability domain were submitted to docking;
- Ipsapione and ebselen were admitted as well;
- The top 25 compounds for which the QSAR Model 2 predicted an IC_{50} value above 10 mM ($pIC_{50} < 2$) were selected. These molecules were chosen for exploration as, whilst high IC_{50} values are normally associated with a lack of interaction with the target, docking may reveal other non-inhibitory interaction mechanisms, which would be particularly interesting for pharmaco-chaperoning strategies that typically tend to be limited by the balance between the structural stabilisation of the protein target and the inhibition effect they exert;
- The top 50 compounds for which Model 2 predicted an IC_{50} value below 10 μ M ($pIC_{50} > 5$) were also selected, since lower IC_{50} values are related to potent inhibitors, meaning strong ligand-target interactions;
- At last, it was admitted for docking the eight compounds for which Model 2 predicted IC_{50} values contained within the range of 1.0 to 1.2 mM ($2.9 > pIC_{50} > 3.0$), which are the neighbouring values for the IC_{50} prediction for lead-like compound VIII (≈ 1.19 mM).

Docking was conducted as described in **Section 2.1.4**. Regarding the search space for docking at the protein surface, a user-defined search grid box is applied to set the boundaries inside of which the algorithm will generate the binding poses. Determining a suitable search space is a crucial step in the process, as an overly narrow docking box may limit the number of possible conformations and an excessively large box could lead to calculation of too many irrelevant binding poses, adding noise.

Whilst the dimensions of the grid box are heavily dependent on the active site at the target protein, literature examples suggest that an optimal docking accuracy can be obtained in AutoDock Vina for a search space around 2.9 times the radius of gyration of the explored ligand²³⁹.

As the present docking approach was done in batch for the about 90 ligands examined, the grid box was set in the centre of PMM2's active site at the interface between the cap and core domains and its dimensions were so that the search space fully involved the entire interfacial region.

Since the AutoDock Vina software uses a non-deterministic search algorithm to generate the binding pose, *i.e.*, an algorithm that performs random choices at certain steps during pose calculation and, therefore, does not necessarily yield the same output in every trial, docking was conducted by performing at least five experimental replicates, each generating the ten lowest energy ligand conformers for each compound submitted.

A binding pose and a docking score representing the free energy of binding ΔG , was calculated for each drug compound and, in order to understand what sort of results were the most interesting to be pursued, the experimental works described by Briso-Montiano and co-workers⁴⁷ were revisited: This study reported the complete crystalline structural model of hPMM2 and, although unsuccessful, attempts to crystallize PMM2 in complex with compound VIII were made. Despite this setback, isothermal titration calorimetry (ITC) experiments allowed to measure the interaction between compound VIII and PMM2, as well as for the enzyme's Glc 1,6-BP activator.

Quite briefly, the ITC technique is based on the thermodynamic principle that an interaction between two molecules, such as in the binding of a ligand to a protein, results in either heat generation (exothermic reaction) or absorption (endothermic reaction). By measuring the change in temperature along the reaction, it is possible to calculate thermodynamic parameters related to the molecular binding interaction, such as the binding constant K_D , the enthalpy and entropy of the system, ΔH and ΔS , as well as the reaction stoichiometry n ²⁴⁰.

Experimental ITC data (**Appendix, Figure 45**) quantified the binding of the Glc 1,6-BP activator with a $K_D \approx 1 \mu\text{M}$ in a stoichiometry of one molecule/monomer, indicative of a high-affinity interaction, just as enzyme activators generally are. Furthermore, compound VIII yielded a binding constant, $K_D \approx 9 \mu\text{M}$, again proving the ability of this compound to interact with the enzyme.

Remembering that the end goal of this CADD workflow was to identify potential leads for pharmacochaperoning, it was defined that potential hit molecules identified should present a free binding energy close to the one reported for compound VIII, whose interaction with PMM2 is well documented and experimentally validated. Also, to prevent a strong competitive inhibition with the enzyme's own sugar activator, its K_D value was set as an upper limit to the docking scores. From the K_D constants determined by ITC for compound VIII and Glc 1,6-BP to hPMM2 it is possible to calculate the standard free binding energy, ΔG° , by applying the **Equation 15** below:

$$\Delta G^\circ = RT \times \ln K_D$$

15

Where ΔG° is provided in kcal/mol, R is the perfect gas constant, equal to $1.987 \times 10^{-3} \text{ kcal} \cdot \text{K}^{-1} \cdot \text{mol}^{-1}$; T is the temperature, defined as 298.15 K and $\ln K_D$ is the natural logarithm of the dissociation constant in molarity²⁴¹. The calculated ΔG° for compound VIII and the Glc 1,6-BP activator towards hPMM2 were -6.86 and -7.80 kcal/mol, respectively. Keeping these values in mind, around 30 hit compounds were picked based on their docking scores.

A first analysis of results was carried out by examining the docking scores obtained for the template compound set and screened drug molecules. As shown in **Table 17** for the templates set, the calculated ΔG for compound VIII, template D), was -6.90 kcal/mol, which is extremely close the experimental -6.86 kcal/mol result. This appears to indicate that the docking approach is accurately predicting the affinity of the compound towards PMM2 and should therefore provide a trustworthy depiction of the

protein-ligand interaction established by this template at the enzyme's active site. Other positive templates reported free energy values in the same range as compound VIII, which seems to agree with the biological activities that are reported for these molecules. Interestingly, the affinity score for negative template compounds F) and G) is noticeably higher than the score obtained for the positive templates. Given that these compounds were seen in the dataset literature to be unable to interact with the enzyme, the calculation of a higher ΔG is a sensible observation, as these higher free energies of binding are associated with a decrease in the binding potency of the ligand²⁴².

Docking scores presented low standard deviation values between experimental replicates for each template, indicating that pose calculation and subsequent binding ranking was generally very precise.

Table 17. Docking scores obtained for the template compound set.

Template compound	Docking score* (kcal/mol)	σ^{**}
A)	-6.15	0.26
B)	-6.90	0.15
C)	-7.60	0.12
D)	-6.90	0.09
E)	-6.60	0.20
F)	-5.85	0.45
G)	-5.40	0.05

* The docking score indicates the free energy of binding ΔG for the protein-ligand interaction, expressed in kcal/mol, for the lowest energy conformer of the ligand. Values represent the median of at least five experimental replicates.

** σ indicates the standard deviation value verified between the experimental replicates for each ligand docked.

For the examination of the screen set results (**Table 18**), drug hits were grouped according to their chemical scaffolds. About 30 ligands achieved docking scores within the range of interest. Enantiomers of the same molecular species were verified amongst the hits, which makes sense given that a wide variety of drugs can be commercialized as either a pure enantiomer or a racemic mixture²⁴³. For simplicity's sake of the analysis, every drug containing enantiomers were identified, analysed and only one was included in the final examination.

As shown in the table below, there was a great diversity amongst the drug scaffolds observed. Amongst the chemical space of the hit ligands, the most represented classes comprised quinolone derivatives; phenylpropanoic acid derivatives; camptothecin derivatives; several nitrogen-containing moieties, which are common functional groups in drug molecules; and vitamin derivatives.

Drugs derived from vitamins tend to hold great therapeutic potential as they benefit from low toxicity values and the capacity to bind a myriad of protein receptors, targeting different enzymes and pathways. In a recent study, molecular docking experiments were conducted with selected vitamins and their derivatives to predict the binding to a potential SARS-CoV-2 main protein target²⁴⁴.

Quinolones (*vide* **Figure 34A**) are a class of antibiotics widely used to treat a range of bacterial infections that act by binding to prokaryotic topoisomerases and DNA gyrase exerting an inhibitory effect and corrupting enzyme function into producing double-stranded breaks in the bacterial chromosome²⁴⁵.

Phenylpropanoic acid (**Figure 34B**) englobe a class of nonsteroidal anti-inflammatory drugs, NSAIDs, that are highly bound to plasma proteins and possess the ability to inhibit synthesis of prostaglandins and thromboxane, being therefore primarily used for their analgesic and anti-inflammatory actions to reduce inflammation and pain symptoms²⁴⁶.

Camptothecin derivatives (**Figure 34C**) are a group of natural alkaloids extracted from *Camptotheca acuminata* trees and used as chemotherapy agents. They bind to and inhibit DNA topoisomerase to obstruct DNA replication in the rapidly dividing cancerous cells, yielding the desired anti-tumoral effect²⁴⁷.

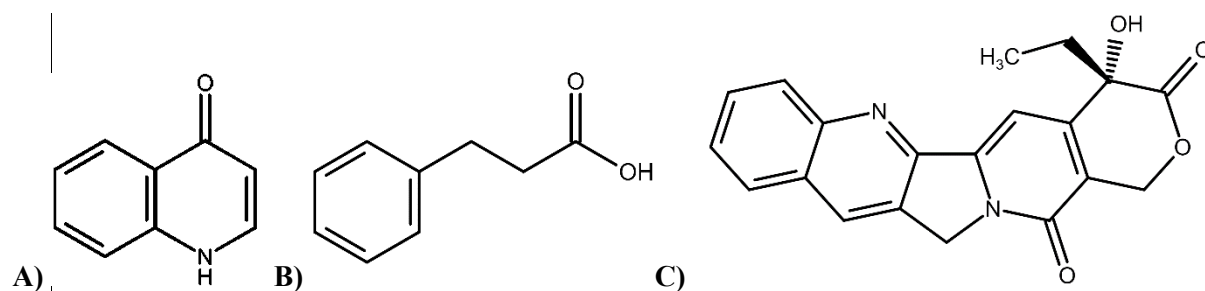


Figure 34. A) Quinolone, B) phenylpropanoic acid and C) camptothecin scaffolds.

Table 18. Scaffold grouping of potential hit compounds identified with molecular docking and respective docking scores.

Scaffold identification	Traditional name	Therapeutic applications	Docking score* (kcal/mol)	σ^{**}
Quinolone derivatives	Levofloxacin	Bacterial antibiotic ²⁴⁸	-7.50	0.22
Quinolone derivatives	Delafloxacin	Bacterial antibiotic for skin infections ²⁴⁹	-6.70	0.30
Quinolone derivatives	Ozenoxacin	Bacterial antibiotic for skin infections ²⁵⁰	-6.90	0.06
Quinolone derivatives	Trovafoxacin	Antibacterial, withdrawn from market due to hepatotoxicity ²⁵¹	-7.20	0.05
Quinolone derivatives	Grepafloxacin	Antibacterial, withdrawn from market due to cardiotoxicity ²⁵²	-6.75	0.19
Phenylpropanoic acid derivatives	Carprofen	Anti-inflammatory used for dogs, no reported human usage ²⁵³	-6.70	0.09
Phenylpropanoic acid derivatives	Flurbiprofen	Anti-inflammatory ²⁵⁴	-6.70	0.16
Phenylpropanoic acid derivatives	Pranoprofen	Anti-inflammatory ²⁵⁵	-7.00	0.10
Camptothecin derivatives	Exatecan	Chemo-therapy agent	-7.80	0.16
Camptothecin derivatives	Topotecan	Chemo-therapy agent ²⁵⁶	-7.05	0.18
N-benzoyl-L-glutamic acid derivate	(<i>R</i>)-pralatrexate	Anti-cancer drug (lymphoma treatment) ²⁵⁷	-6.90	0.19
Vitamin B ₉ derivative	Levoleucovorin 5-formyl	a.k.a. folinic acid, used as chemo-therapy agent ²⁵⁸	-7.15	0.30
Vitamin B ₂	Flavin mononucleotide	Various applications ²⁵⁹	-7.80	0.49
Vitamin B ₂	Riboflavin	Various applications ²⁵⁹	-6.90	0.24
Pyrazol-containing	Axitinib	Anti-cancer drug (TKI inhibitors) ²⁶⁰	-7.00	0.23
Pyrazol-containing	Lorlatinib	Anti-cancer drug (TKI inhibitors) ²⁶¹	-7.20	0.13
Piperazine-containing	Ranolazine	For angina, chest pain ²⁶²	-6.45	0.32
Piperazine-containing	Ipsapirone	Anti-depressive ²³⁷	-6.35	0.24
Pyrrolopyrimidine-containing	Vibegron	Anti-diuretic ²⁶³	-7.05	0.46

Oxymorphone derivative	Methylnaltrexone	For opioid-induced constipation ²⁶⁴	-6.60	0.06
Tetracycline-class antibiotics	Doxycycline	Bacterial antibiotic ²⁶⁵	-7.30	0.00
Cephalosporin derivatives	O-formylcefamandole	Bacterial antibiotic, discontinued as drug	-7.50	0.44
Retinoid	Tazarotenic acid	Bacterial antibiotic for skin infections ²⁶⁶	-6.35	0.19
Chromene derivative	Amlexanox	Bacterial antibiotic, discontinued in the USA by the FDA ²⁶⁷	-6.50	0.22

* The docking score indicates the free energy of binding ΔG for the protein-ligand interaction, expressed in kcal/mol, for the lowest energy conformer of the ligand. Values represent the median of at least five experimental replicates.

** σ indicates the standard deviation value verified between the experimental replicates for each ligand docked.

The binding poses were analysed for all docked ligands within the range of interest and the non-covalent interactions formed between each ligand and the protein, such as H-bonds, hydrophobic interactions and saline bridges, were explored using molecular visualisation softwares such as PyMOL or Chimera, and LigPlot+ to generate schematic diagrams representing these interactions. It would be impossible to illustrate in this thesis the complete structural analysis that was performed for all the docked ligands and so, as a way of exemplifying the general procedure employed during analysis, the case study of compound VIII is represented here.

Firstly, the docked ligand was visualised at the protein's surface as **Figure 35** shows. In the present work, the docking software was set to generate the ten lowest energy conformers for the ligand, meaning that ten independent structures were obtained. Every conformer should be observed and choosing which one is the most suitable to further explore the protein-ligand interactions must involve not only paying attention to the calculated ΔG for the interaction, as, in theory, the lower the free energy of binding is, the higher the complex's affinity is, but also minding the positioning of the ligand in the protein.

After selecting the conformer of interest, describing in detail the surrounding chemical environment at which the ligand is placed and the interactions that it establishes is essential to grasp some notion of the underlying binding mechanism. This exercise can be done by generating a diagram that represents the several non-covalent interactions at hand, as it was done in **Figure 36**. A list describing the interactions with intervening atoms and bond distances was also generated.

This process culminates on a comprehensive description of all the interactions observed between protein and ligand. For instance, in the case of compound VIII, it was seen that the ligand binds near PMM2's $\beta 6$ - $\beta 7$ loop and $\beta 2$ - $\alpha 2$ loop, establishing hydrophobic interactions with the side chain carboxyl of Glu129; the backbone of Asn216; the amide oxygens of Gly214, Gly215 and Ile174; the carboxyl sidechain of Asp 188; ring interactions with Phe183; the C α of Gly175; the OG of Ser173 side chain and the side chain methyl CG2 of Thr171.

By proceeding in this manner for the docking analysis, it was possible to characterise the binding environment and the complex interactions that each ligand established with the protein during docking.

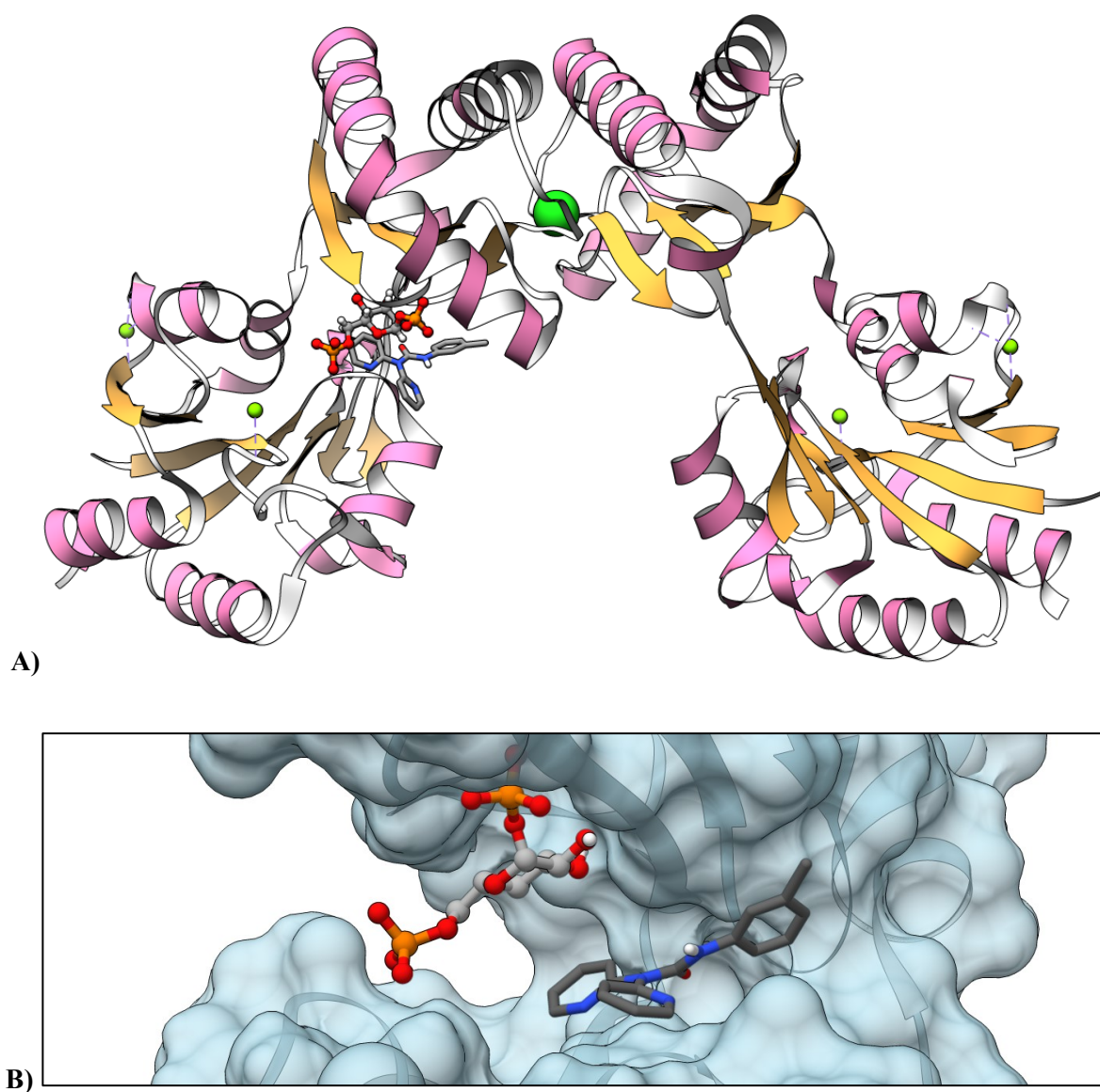


Figure 35. Structural representation of docked ligand 1-(3-chlorophenyl)-3,3-bis(pyridine-2-yl)urea in hPMM2. A) Ribbon representation of hPMM2 dimer showing secondary structure elements α -helices (in pink), β -sheets (in yellow) and turns (in white), with the enzyme's activator Glc 1,6-BP in ball-and-stick representation and ion co-factors Mg^{2+} and Cl^- at their respective coordination sites, as green spheres (PDB entry: 7O1B). The ligand compound VIII, illustrated in stick representation, is docked at one of the monomer's cap and core domains interface, near the active site. B) Detail of the ligand's binding site at the protein surface. Images generated using UCSF Chimera version 1.14.

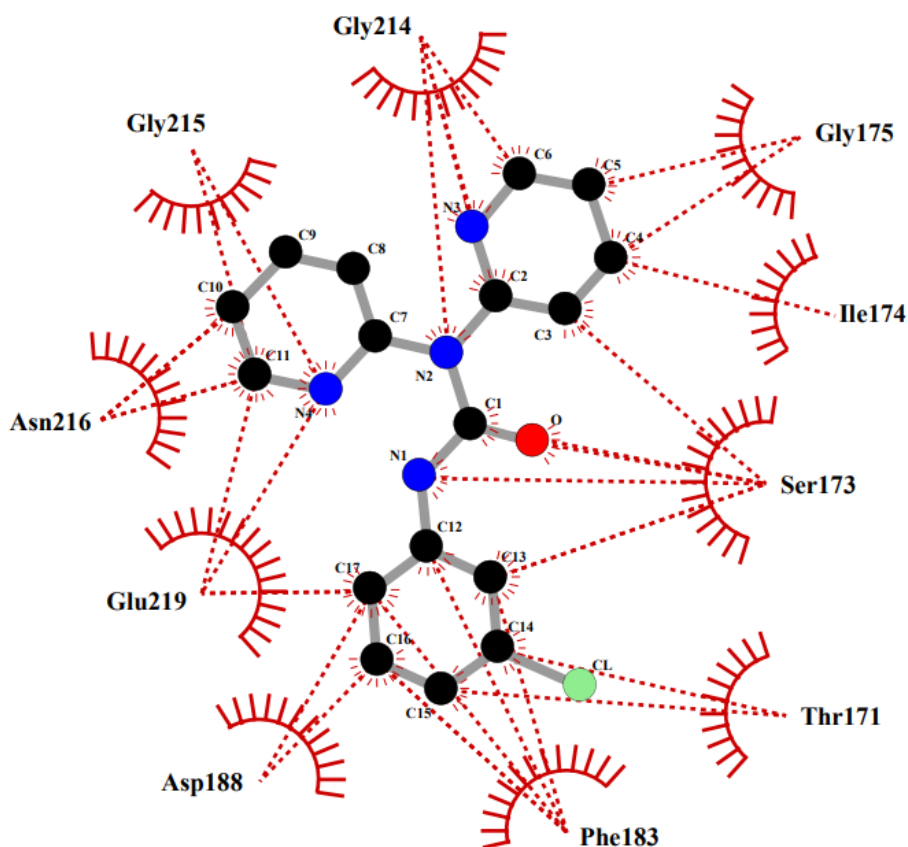


Figure 36. Schematic diagram representing the protein-ligand interaction established between hPMM2 and docked ligand 1-(3-chlorophenyl)-3,3-bis(pyridine-2-yl)urea. The ligand is depicted surrounded by protein residues and hydrophobic interactions are denoted in red trace, indicating the atoms at the ligand to which each interaction is established. Image generated using LigPlot+ version 2.2.5.

To infer on what key hPMM2 amino acids are most probable to participate in the binding of potential drug candidates, all non-covalent interactions that were estimated by docking and analysed in LigPlot+ for the positive templates and hit ligands from the drug screen set, encompassing H-bonding, saline bridges and hydrophobic interactions, were compiled and the relative occurrence of each residue in an interaction was calculated. In other words, the ratio between the number of times that each residue was found establishing an interaction with a ligand and the total number of interactions recorded was measured, with the results being presented below in **Figure 37**.

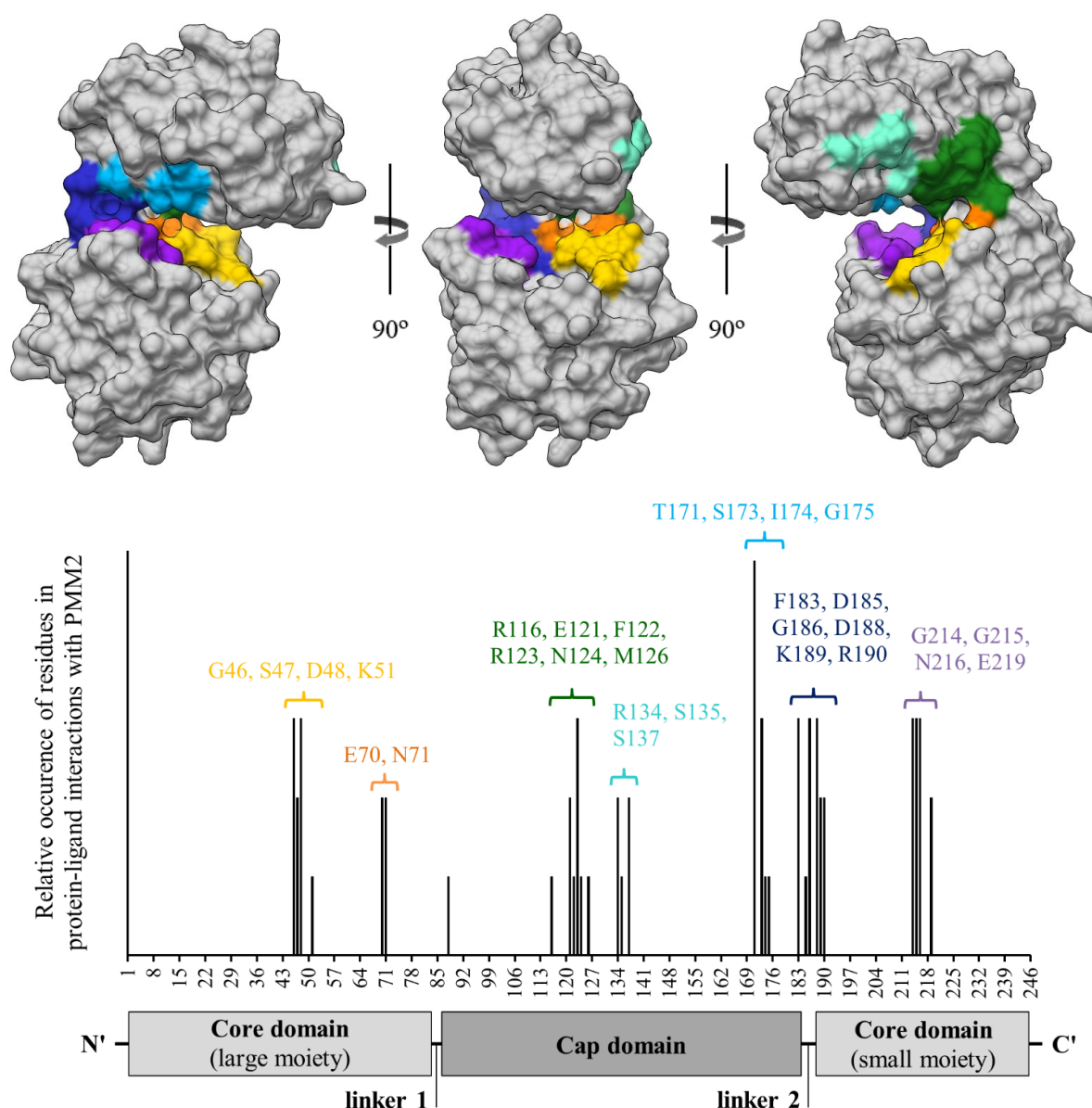


Figure 37. Relative occurrence of PMM2 residues in the protein-ligand interactions described for all docked ligands. At the bottom, the banner schematizes PMM2's polypeptide chain, divided into its two domains: the core domain (large moiety at aa 1-83 and small moiety at aa 189-246) and the cap domain (aa 87-184), connected by linker 1 (aa 84-86) and linker 2 (aa 185-188). For each residue, the bar chart describes the occurrence of that amino acid in interactions observed for docking protein-ligand complexes relative to the total number of analysed interactions. At the top, surface representation of hPMM2 monomer with interface residues coloured according to the region grouping described before. Image generated in UCSF Chimera version 1.14.

By analysing these results, it was possible to conclude that the examined ligands were all bound to the protein at the active site, located in the great pocket that the interface between the cap and core domains forms. Residues scattered across this region were seen to be relevant donors for non-covalent interactions, spanning from the flexible linker 2 region up to β -strands 6 and 7 close to the dimerization site. These observations seem to propose that the screened ligands follow a binding mechanism like that of

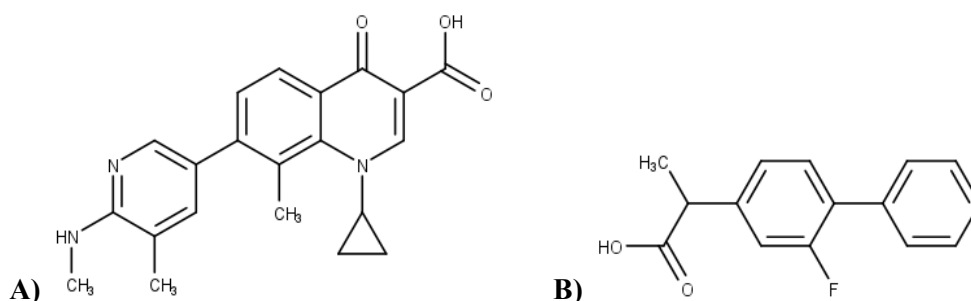
a first-generation pharmaco-chaperone, which typically tend to establish an interaction at the protein's active site and may exhibit an inhibitory effect, as they compete for the activators and substrates required for catalysis. Furthermore, the presence of a considerable number of interactions near the dimerization interface may loosely suggest that the binding of lead compounds could enhance the affinity of the functional holodimer, which would be an especially interesting effect for therapy of pathogenic PMM2 mutations affecting residues responsible for the dimerization interactions.

All in all, this docking workflow proved to be an invaluable tool to identify promising ligands for experimental validation from the virtual drug screen and allowed to further elucidate on the possible binding mechanisms and interaction dynamics that are thought to occur between PMM2 and the leads.

3.2 Biological component

In the “wet” lab component of this thesis project, an experimental validation was carried out with the objective of studying the pharmaco-chaperoning potential of the identified hit drugs for repurposing in a PMM2-CDG disease model consisting of patient-derived fibroblast cell lines.

To do so, four of the lead compounds that were analysed in the virtual screen and docking approaches were acquired (*vide* **Figure 38**). These comprised a quinolone derivative (A) known as ozenoxacin (CAS Registry no. 245765-41-7), the phenylpropanoic acid derivative (B) flurbiprofen (CAS Registry no. 5104-49-4), vitamin B₂ (C) (CAS Registry no. 83-88-5) and triazole compound BPTT (D) (CAS Registry no. 126651-85-2). Note that BPTT was one of the negative templates originated from the initial compound dataset used during QSAR model building to define the model's applicability domain. Including a negative control compound in this experimental study to explore the therapeutic potential of lead compounds was an essential task to ensure that the negative class predictions generated by the QSAR models were accurate, *i.e.*, that the negative template does not exert a therapeutic effect on the cellular disease model under examination.



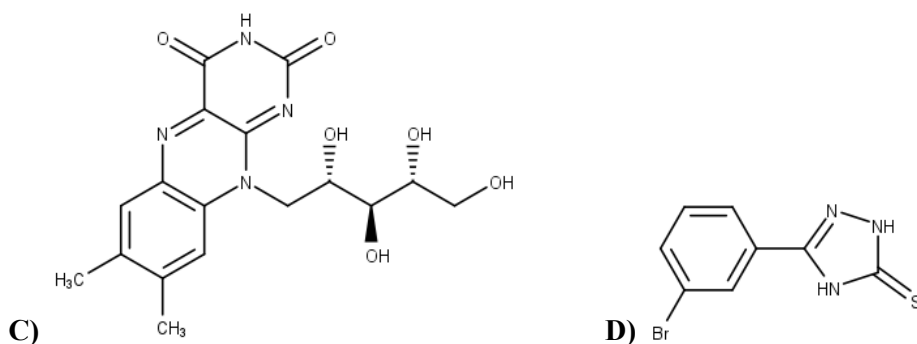


Figure 38. Compounds acquired for experimental validation with biological studies. **A)** Ozenoxacin: 1-cyclopropyl-8-methyl-7-[5-methyl-6-(methylamino)pyridin-3-yl]-4-oxoquinoline-3-carboxylic acid. **B)** Flurbiprofen: 2-(3-fluoro-4-phenylphenyl)propanoic acid. **C)** Riboflavin: 7,8-dimethyl-10-[(2S,3S,4R)-2,3,4,5-tetrahydroxypentyl]benzo[g]pteridine-2,4-dione. **D)** BPTT: 5-(3-Bromo-phenyl)-2,4-dihydro-[1,2,4]triazole-3-thione.

The biological evaluation component was conducted using as study model the patient-derived primary fibroblast cell line GM27386, and the wild-type control line GM00969. Whilst the appropriateness of fibroblast models for drug research has already been discussed in **Section 1.4.1**, some practical remarks on the selection of such models should be made:

From the small collection of PMM2-CDG patient-derived fibroblast lines available at the laboratory, the GM27386 line, which contains the heterozygous mutated variants c.422G>A and c.415G>A, translating into the two single-residue substitutions p.Arg141His and p.Glu139Lys, was selected. As mentioned before, the R141H mutation is the most represented across patients and, therefore, its study is of special relevance. In its turn, Glu139 is located at the beginning of helix $\alpha 5$ in the cap domain of PMM2 (*vide* **Figure 5**) and the E139K substitution replaces the glutamate's carboxylate negative charge by the positive charge of side chain amine of lysine, which might disturb charge capping at the helix 5. Despite this effect being yet to be explored, the fact that this mutation can be verified in a living patient alongside the deleterious R141 substitution seems to suggest that it constitutes a milder disease phenotype.

Moreover, when studying a pathogenic cell line, a healthy control must always be kept in parallel, so that a term of comparison may be drawn during experimentation. The differences between the control and the disease lines should be the minimal essentials, in this case, the underlying mutations in the *PMM2* gene. Choosing a control line derived from a female donor with a similar age as the patient from which the PMM2-CDG was collected, is a good practice to ensure comparability.

Another noteworthy aspect is that, to guarantee that the results obtained by manipulating the fibroblast lines were only originated due to the direct influence of the variables under analysis and no other external contamination factors, every cell batch that was utilised in the works here described were negatively tested towards *Mycoplasma spp.* contamination and grown in strict sepsis conditions, as described before.

3.2.1 Cell viability assays

To evaluate the effect of the four leads, it was necessary to prepare solutions for all of them. DMSO is typically used as solvent for compounds that report poor water solubility, as it was the case for ozenoxacin, flurbiprofen and BPTT. Since DMSO is widely known to have a cytotoxic effect²⁶⁸, an initial step for experimentation was to characterise the viability of both cell models towards exposure to different concentrations of DMSO in the DMEM culture medium. This exercise was performed using resazurin assays.

Resazurin assays provide a standard measurement for cellular viability. By employing a non-destructive technique that is based on a viable cell's capacity to reduce resazurin, a non-fluorescent blue azo-dye that has an absorption max at 600 nm, to the pinkish resorufin, which is highly fluorescent and absorbs at a maximum λ of 570 nm. This reduction reaction is carried out by mitochondrial reductase enzymes in an uncompromised, metabolically active cell and it can be easily monitored by UV-Vis spectroscopy or even spectrofluorometric methods. The difference in the absorbance contributions from resazurin and resorufin, $Abs_{570\text{ nm}} - Abs_{600\text{ nm}}$, (or the resorufin fluorescence intensity output) is proportional to the number of viable cells and therefore provides a simple method to examine the cytotoxic effects of several compounds, in various concentration ranges, in several cell types, including fibroblasts²⁶⁹.

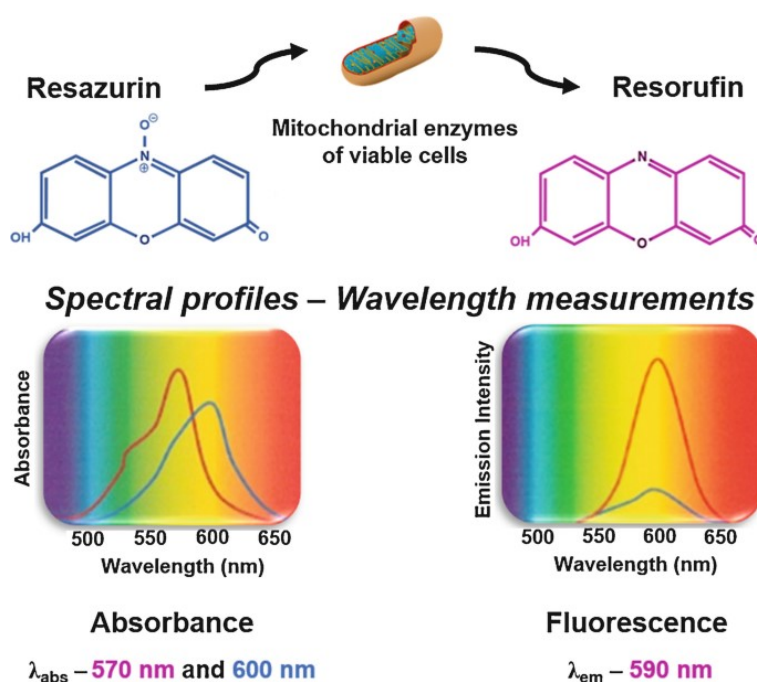


Figure 39. The underlying basis of methodology used in resazurin assays. Adapted from²⁷⁰.

Firstly, a working cell density for the resazurin assays was successfully defined for both GM00969 and GM27386 lines in a microplate assay by experimenting with different densities, namely 6.0×10^4 ,

3.0×10^4 and 1.5×10^4 cells/well (data not shown), having been set the intermediate option, 3.0×10^4 cells/well, as the most suitable under the assay and equipment's conditions.

Then, the differences in metabolisation rate between the control and the PMM2-deficient line were explored, to infer if the assay comparability between lines would be affected. As the plot in **Figure 40** shows, both cell lines describe a similar behaviour and a statistical analysis using a two-way ANOVA test to see if viability was significantly different between lines at each time point was made. This yielded no significant differences at $p < 0.05$ between the lines.

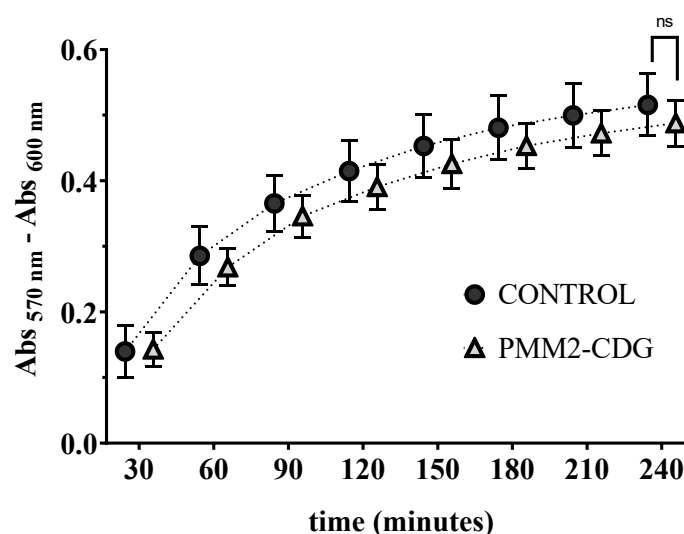


Figure 40. Scatter plot of resazurin metabolisation for control and PMM2-CDG fibroblast lines. The metabolic activity measured by the conversion rate of resazurin into resorufin was calculated by subtraction of background resazurin absorbance at 600 nm to the resorufin absorption maximum at 570 nm at each time point for the GM27386 and GM00969 lines. Represented values are the mean $\pm$ SD. Statistical significance tests performed in GraphPad Prism indicate that the differences observed between both lines at each time point are not significant to a $p < 0.05$ ($n \geq 3$).

Afterwards, the evaluation of the cytotoxic effect of increasing concentrations of DMSO at the growing fibroblasts DMEM medium was successfully conducted with a 24 hour-exposure resazurin assay to define an adequate DMSO percentage to be used in preparing lead compounds solutions without causing a negative decrease in cell viability. **Figure 41** portrays the results that were obtained for different concentrations. Note that 10% and 20% DMSO were used as negative controls, at which there is a considerable loss on viability reported. By analysing the results and further performing a statistical analysis similar to what was done before, it was seen that no considerable differences were registered for DMSO concentrations ranging from 0%; 0.1%; 0.5%; 1% until 2%, with the control 10% and 20% indicating, as expected, a significant decrease in viability.

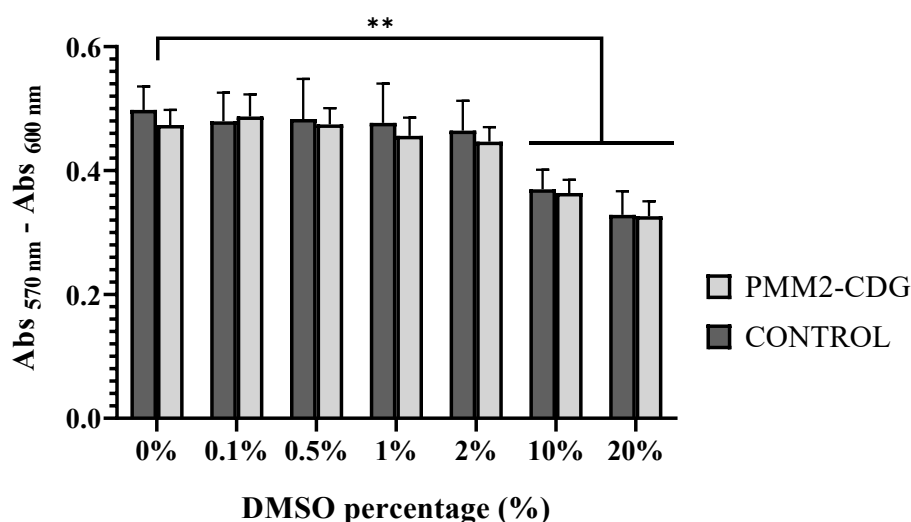


Figure 41. Bar chart of cell viability for control and PMM2-CDG fibroblast lines exposed to different concentrations of DMSO. The metabolic activity at four hours past resazurin addition was measured by the conversion rate of resazurin into resorufin for the GM27386 (lighter grey) and GM00969 (darker grey) lines exposed for 24 hours to increasing concentrations of DMSO in culture medium. Values are the Abs_{550 nm} – Abs_{600 nm} difference (mean $\pm$ SD). A statistically significant decrease in cell viability at ** $p < 0.05$ ($n \geq 3$) was observed for the higher 10% and 20% DMSO control solutions relatively to the remainder concentrations for both cell lines.

Considering the results obtained, a final DMSO percentage of 0.5% was cautiously deemed as appropriate to be used for the future therapeutic studies. The next step intended to again assess cell viability through a 48-hour exposure resazurin assay with the four lead compounds (**Figure 42**). This was done to understand how different concentrations of each hit drug affected the fibroblasts and to assure that any potential therapeutic effect that the drugs may exert would not be counteracted by a toxic effect resultant from an excessive concentration of the compounds at the cell.

A set of six test concentrations, placed around the values that are typically reported as being amenable for drug applications in *in vitro* experimentation with cell systems, was selected based on the former literature that explored PMM2 assays^{87,209}, ranging from 1 to 200 μ M.

Unfortunately, there was only time to perform a single attempt of this assay and so, no great correlations can be taken from the data, as the assay's variability and technical details still need to be perfected. Yet, it seems that a very subtle decrease in cell viability may take place for the higher concentrations, although this a very loose interpretation of this preliminary data.

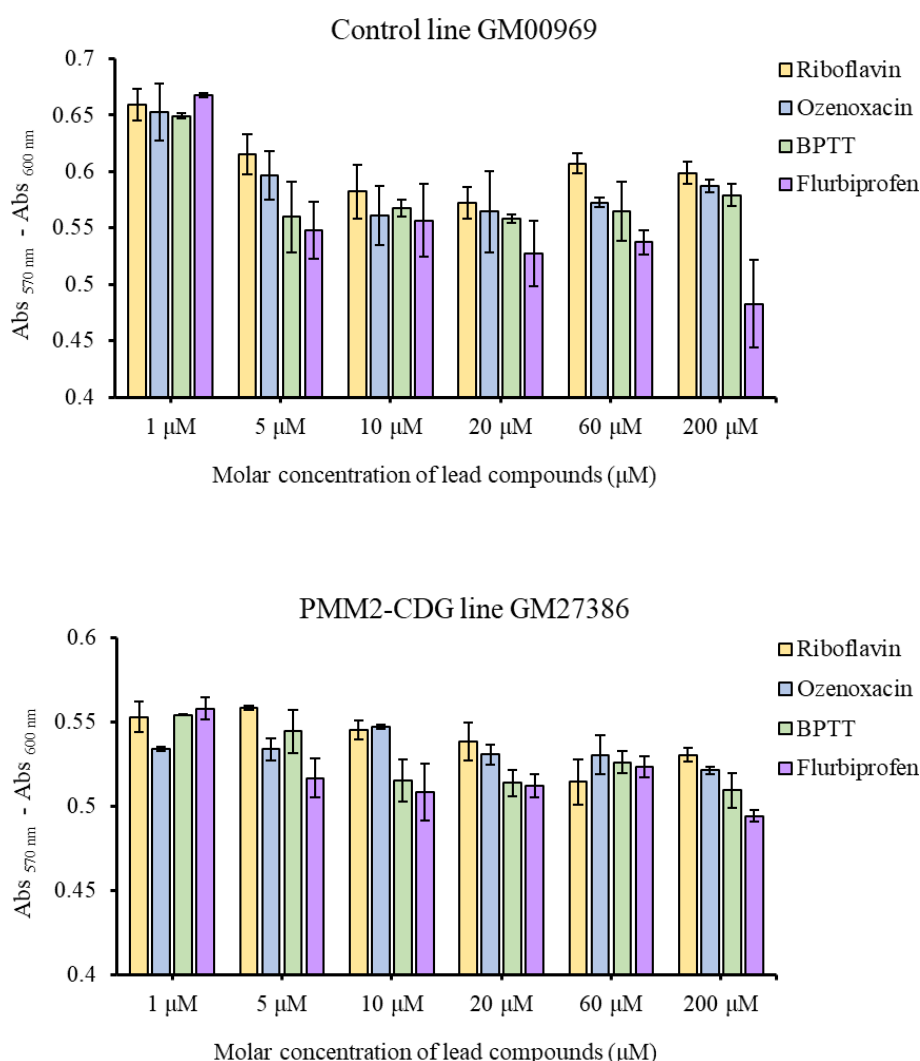


Figure 42. Bar charts of cell viability for control and PMM2-CDG fibroblast lines exposed to different concentrations of the four lead drugs. The metabolic activity at four hours past resazurin addition was measured by the conversion rate of resazurin into resorufin for the GM00969 (at the top) and GM27386 lines (at the bottom) exposed during 48 hours to a set of six different concentrations (1, 5, 10, 20, 60 and 200 μM) of riboflavin in 0.1 M NaOH aqueous solution (in yellow) and ozenoxacin (blue), BPTT (green) and flurbiprofen (purple) in 0.5% DMSO. Values are the $Abs_{570\text{ nm}} - Abs_{600\text{ nm}}$ difference (mean $\pm$ SD).

As detailed earlier at **Section 1.5**, this project aimed to study the therapeutic action of drug compounds by observing their effect on PMM2 and ICAM-1 expression/turnover patterns, which were analysed through western blotting methodologies, and on deficient PMM2 residual activity, measured by enzymatic assays.

3.2.2 Evaluation of PMM2 enzyme activity in fibroblast extracts

The kinetic assay described in **Section 2.2.4** for determining PMM2 activity was conducted accordingly, and the results are depicted next in **Figure 43**.

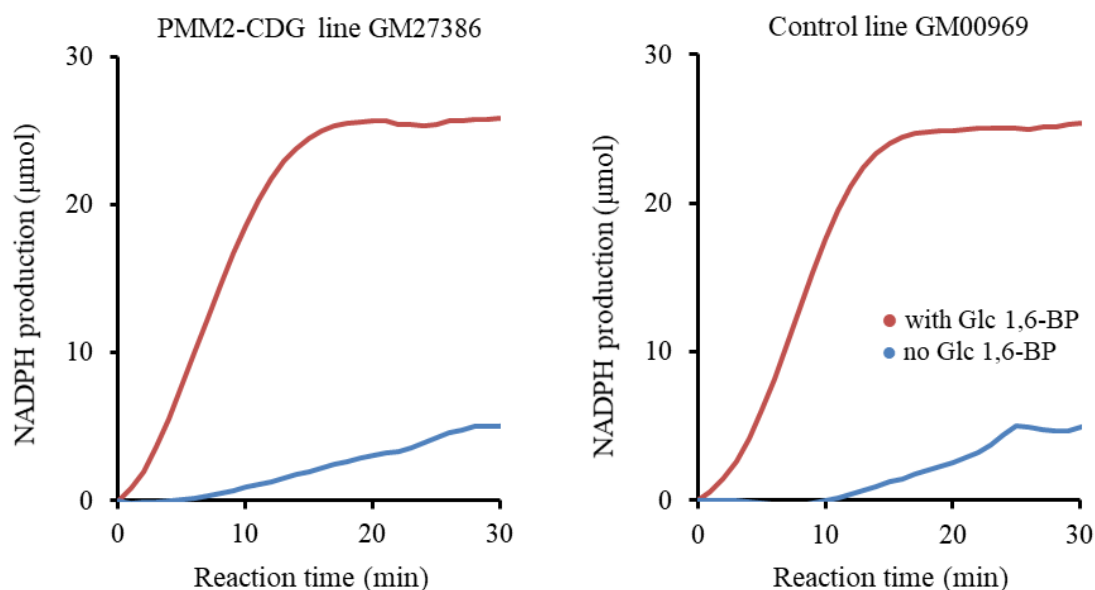


Figure 43. NADPH production as a function of reactional time. NADPH productions in soluble cell extracts from the control line GM00969 (right) and PMM2-CDG line GM27386 (left) were measured by UV-Vis spectroscopy at 340 nm (NADPH absorption max.) for 30 minutes at 37°C. The normal assay reaction mix contained all reagents (red lines), whilst the blank reaction mix did not contain the Glc 1,6-BP activator (blue lines). NADPH production is expressed as the number of micromoles of NADPH produced, subtracted by the initial $Abs_{340\text{ nm}}$ at the extract used. Values are the mean of three technical replicates.

Curve analysis revealed, as it was feared, that this experimental setup is not proper to study specific phosphomannomutase catalytic activity in fibroblast models. Differences between the curves for the control and the patient-derived cell extracts are virtually non-existent, indicating that the $NADP^+ \rightarrow NADPH$ conversion rate that is being measured is a result from the activity of other related enzymes, most probably enzymes with a phosphoglucomutase (PGM) activity, which is measured using the same assay as reported in literature²⁷¹. Some considerations on why this was assay unable to measure the intended outcomes related to previous observations that PMM and PGM activities in fibroblasts are generally very low and therefore difficult to detect, reason why this assay should be performed with purified PMM2 protein samples. Also, a range of enzymes related to PMMs utilise bisphosphate sugar as endogenous activators, including PGM²⁷¹, and so, the reaction is not specific for PMM2. Furthermore, whilst PMM2 can catalyse both activated forms of glucose and mannose, PGM is unable to utilise phosphorylated mannose as substrate, making the assay pathway that employed mannose as the starting reagent and requiring all the ancillary enzymes (PMI, PGI and G6PDH) (*vide* **Figure 15**) much more fitting. Additionally, an important experimental remark that should always be reflected upon when conducting colorimetric assays in a microplate setup continuously during a large timespan is that the baseline drift within the spectrophotometric measurement can affect the results. Baseline drift is a type of long-term noise that describes a change in the background baseline position along the assay as a consequence of

the effect that the cellular content present in the extracts has on the experience over time and eventual variations in temperature, amongst other factors. Since this drift can induce errors in quantitative analysis, it is necessary to ensure that they are accounted for. In this case, a practical manner of addressing the drift correction could be performing a sweeping spectrum for each data point and discounting the contributions from the absorbance at a region with a wavelength λ around 500-550 nm, where no peaks and major contributions originating from the NADPH reduction nor cellular contaminants are supposed to occur.

3.2.3 Western blotting with fibroblast extracts

Western blotting experiments were conducted to assess the differences in protein stability and turnover patterns in the two PMM2 and ICAM-1 biomarkers for both cellular models under examination. Experiments were performed successfully and the final membrane signal detection by ECL is depicted in **Figure 44** below.

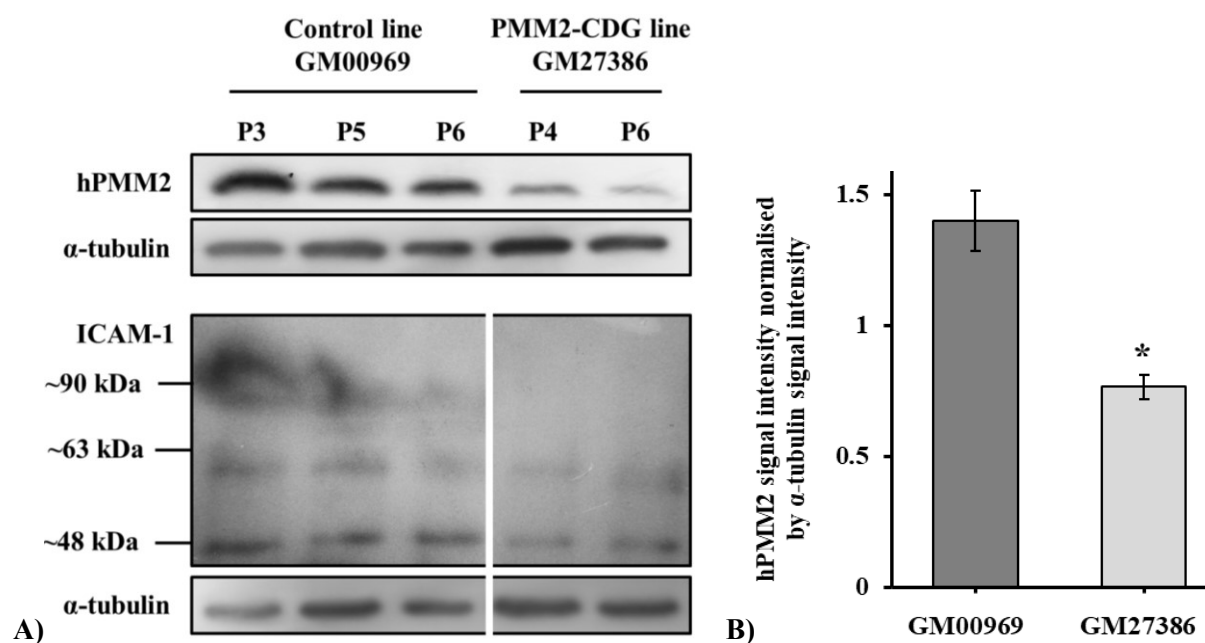


Figure 44. hPMM2 and ICAM-1 western blotting analysis. **A)** Western blots were performed using both the control and PMM2-CDG fibroblast lines at different sub-culturing stages (passages P3, P5 and P6 for the wild-type GM00969 line and P4 and P6 for the patient-derived GM27386 line). The figure indicates the final signal detection obtained by ECL for hPMM2 (at the top) and ICAM-1 (at the bottom). The relative positioning of some of the molecular weight standards from the *Prestained Protein Ladder - Broad molecular weight (10-245 kDa)* (Abcam) are indicated for ICAM-1 blotting. α -tubulin was used as a loading control for both experiments. **B)** Bar chart of the associated quantification for hPMM2 signal intensity normalised by the loading control for the wild-type GM00969 line, in darker grey, and PMM2-CDG GM27386 line, in lighter grey (mean $\pm$ SD). A significant decrease at * $p < 0.05$ ($n \geq 3$) was seen relative to control basal signal. P values were calculated by using a t-test in GraphPad Prism.

By analysing the PMM2 blot it is possible to clearly identify, as expected, the presence of weaker protein bands for the soluble extracts derived from the patient GM27386 line, comparatively to the ones originated from the control GM00969. This evidence is a confirmation that PMM2 is being prematurely degraded in the PMM2-CDG line due to the destabilising effect that the pathogenic amino acid substitutions Arg141His and Glu139Lys have on the protein's structure, increasing its turnover rate and resulting in a more faded signal. Statistical analysis further corroborates these significant differences. Moreover, the relative age of the cell lines seems to not exhibit an influence on signal intensity.

ICAM-1 western blotting was also completed with success. As this glycoprotein is present in different forms at the cell, different signals were observed. A band at around 90 kDa denotes the functional membrane-bound ICAM-1. These bands were not clear and well-defined, but rather stretched smears, which can perhaps be explained by the fact that ICAM-1 contains eight *N*-glycosylation sites, originating a range of possible glycoforms of different molecular weights. Alternative isoforms encompassing different numbers of domains may also contribute for this observation, as showcased in **Figure 13**. No bands at 90 kDa were observed for extracts produced with the PMM2-CDG model cells, suggesting that hypoglycosylation plays a key role on altering the turnover and possibly compartmentalisation of the functional membrane ICAM-1. Bands of smaller molecular weights were verified as well, namely at around 63 kDa and 48 kDa, according to the molecular ladder that was used, maintaining a tendency to decrease in signal strength for the patient fibroblast samples, when compared to the wild-type control. These bands may be cytosolic monomeric forms of ICAM-1. Unfortunately, no statistical analysis was performed as there were not sufficient independent replicates ($n < 3$) for ICAM-1 blotting from both the PMM2-CDG and control soluble extracts for comparison.

Possible differences in sample preparation methods for ICAM-1 blotting were also assessed by preparing different soluble fibroblast extracts from cell cultures of each line at the same sub-culturing passage. In parallel, fibroblasts were detached from the TC flask on which they were growing using trypsin-EDTA solution in one case, and in the other, cell scrapping using lysis solution was directed performed in the culture TC flask, as mentioned in **Section 2.2.5.2**. After performing the blotting procedure, no signal was observed for the membrane-bound ~90 kDa band of ICAM-1 in the trypsin treated sample (data not shown), in contrast to the signal obtained for the scrapped sample. This experimental remark seems to indicate that detaching adherent cells with trypsin may have a negative effect on transmembrane ICAM-1, which could be damaged, as it is more exposed to enzymatic cleavage at the cell membrane.

All things considered, western blotting proved to be a suitable technique to evaluate the presence of these biomarker proteins in rich soluble cell extracts obtained from patient-derived fibroblast lines. Their applicability may be further developed and they comprise an important tool to possibly assess the therapeutic effects that pharmaco-chaperones have on PMM2 stability and function.

CONCLUSION & FUTURE WORKS

This project had as its main goals the development of a computational approach to explore pharmacological chaperone-based therapies for PMM2-CDG, implementing a comprehensive workflow that included predictive QSAR machine learning models, a virtual drug screen and docking experimentation, followed by further biological examination of identified lead-like compounds to assess their therapeutic potential on patient-derived fibroblast lines. Overall, this dissertation accounts for several novel insights concerning the experimental work that was conducted:

It was seen that the considerable scarcity of data regarding exploratory screening bioassays for hPMM2, which comprises a common hurdle in rare diseases, was an overriding limitation for the training of the QSAR models in WEKA and Keras, built to predict the ability of a compound to interact with PMM2 and a theoretical IC_{50} value for such protein-ligand interactions.

Both the QSAR classification (Model 1) and the regression (Model 2) approaches yielded poor performances, with overfitting of data and the substantial imbalance in “positive” and “negative” compound classes being the main issues identified. The models’ generalisation capacity is subpar to what was initially wanted, as an insufficient amount of positive input instances incapacitated machine learning algorithms. The fact that the dataset’s reduced chemical space was slightly biased towards the BIT scaffold was yet another constrain.

Despite the relative unreliability in the QSAR model predictions, they were utilised as tools in the virtual screen of a large approved drug dataset collected from the ZINC databases, with the scope of searching for and select potential hit compounds for pharmacological chaperoning.

The identification of template compounds whose interaction with PMM2 had been described proved to be essential in order to set the model’s applicability domain. From the 707 compounds that passed at this domain, about 90 ligands proceeded to docking studies in AutoDock Vina.

The molecular docking stage was overall successful, with the binding affinity scores calculated for template compound VIII being similar to the experimentally-verified value for this molecule’s K_d . This observation provided some reassurance on the accuracy of the docking method.

From the total number of analysed ligands, around 30 molecules were perceived to achieve free energy of binding, ΔG , values that were contained within the range that was defined as relevant for the intended use. The generated binding poses for each ligand candidate and its protein-ligand interactions with the PMM2 receptor were fully examined in a complete step making use of different bioinformatics tools such as Chimera, PyMOL and LigPlot+, and the final results provided important information on what residues at the protein's surface are expected to play a role in the binding mechanisms for the identified hits.

All in all, the virtual screening procedure resulted in a final list of around 30 lead-like compounds for drug repositioning towards PMM2 chaperoning, efficiently grouped according to their chemical scaffolds. Four of these lead compounds were acquired for the biological experimentation stage, namely riboflavin (vitamin B2), the phenylpropanoic acid derivative flurbiprofen, a quinolone drug known as ozenoxacin and BPTT.

Patient-derived fibroblast cell line GM27386 (Coriell) was selected as an *in vitro* PMM2-CDG disease model and its characterisation indicated that this line could be easily manipulated in the laboratory setting and compared with wild-type control lines.

Relevant disease biomarkers for pharmaceutical studies in PMM2-CDG were identified amongst ICAM-1 and the own PMM2 enzyme. Both proteins were analysed in soluble fibroblast extracts by western blotting, which proved to be a suitable technique to carry out their detection in such complex samples. Analysis of PMM2 blotting can be conducted to explore protein stability and turnover patterns, as pathogenic mutations in the PMM2 gene were directly associated with a loss of signal intensity, which correlates to the fact that a misfolded protein is much more susceptible to proteasome degradation and therefore will denote an altered turnover rate.

ICAM-1 glycoprotein can also be considered as a direct evaluation of deficiencies at the cellular *N*-glycosylation mechanisms, being negatively affected in PMM2-CDG cell lines.

It is possible that, in future endeavours, these biomarkers can be effectively used to analyse the therapeutic effect of drug supplementation in growing fibroblast lines.

Moreover, the kinetic assays described to evaluate PMM2 enzyme activity in cell extracts deemed not to be suitable for the disease model setup that was utilised and a complete reformulation of the assay would be necessary. This includes re-thinking the use of pure protein samples over protein-rich, complex cell extracts and measuring the specific phosphomannomutase activity by employing the more costly enzymatic pathway.

In the wake of the results obtained in this dissertation, the continuation of upcoming investigation must embrace several lines of action into further developing the study of cheminformatics applied to PMM2-CDG therapy development and the therapeutic effect of the lead compounds that were identified for pharmaco-chaperoning efforts.

If, in the future, more high-throughput screening assays for PMM2 are to be made available, this novel data should be used to iteratively reframe and improve the predictive QSAR models that were built.

More studies are required to understand the effect of the four selected compounds in the fibroblast disease model, including defining the most suitable range of therapeutic concentrations amenable to be used in the studied cell systems. Redrafting the PMM2 enzymatic activity assay into a functional procedure would as well represent a valuable addition to the biologic toolset available to explore PMM2-CDG therapies.

Considering the inclusion of binding assays in the experimental workflow to measure protein-ligand interactions could also constitute a great advantage for this drug development approach. Defining the underlying thermodynamic aspects for the complex established between the PMM2 target and lead drugs can be regarded as a must-have practice to extensively evaluate the on-target affinity of such compounds.

BIBLIOGRAPHY

1. HART, G. Glycosylation. *Curr Opin Cell Biol* **4**, 1017–1023 (1992).
2. Reily, C., Stewart, T. J., Renfrow, M. B. & Novak, J. Glycosylation in health and disease. *Nat Rev Nephrol* **15**, 346–366 (2019).
3. *Essentials of Glycobiology*. (Cold Spring Laboratory Press, 2022).
4. *Carbohydrate Analysis by Modern Liquid Phase Separation Techniques*. (Elsevier, 2021). doi:10.1016/C2019-0-01776-1.
5. Schjoldager, K. T., Narimatsu, Y., Joshi, H. J. & Clausen, H. Global view of human protein glycosylation pathways and functions. *Nat Rev Mol Cell Biol* **21**, 729–749 (2020).
6. Jefferis, R. Human IgG Glycosylation in Inflammation and Inflammatory Disease. in *Comprehensive Glycoscience* 373–392 (Elsevier, 2007). doi:10.1016/B978-0-444-51967-2/00099-4.
7. Freeze, H. H., Chong, J. X., Bamshad, M. J. & Ng, B. G. Solving Glycosylation Disorders: Fundamental Approaches Reveal Complicated Pathways. *The American Journal of Human Genetics* **94**, 161–175 (2014).
8. Ohtsubo, K. & Marth, J. D. Glycosylation in Cellular Mechanisms of Health and Disease. *Cell* **126**, 855–867 (2006).
9. Shivatare, S. S., Shivatare, V. S. & Wong, C.-H. Glycoconjugates: Synthesis, Functional Studies, and Therapeutic Developments. *Chem Rev* **122**, 15603–15671 (2022).
10. Marth, J. D. & Grewal, P. K. Mammalian glycosylation in immunity. *Nat Rev Immunol* **8**, 874–887 (2008).
11. Hirata, T. & Kizuka, Y. N-Glycosylation. in 3–24 (2021). doi:10.1007/978-3-030-70115-4_1.
12. Jaeken, J. & Matthijs, G. Congenital Disorders of Glycosylation: A Rapidly Expanding Disease Family. *Annu Rev Genomics Hum Genet* **8**, 261–278 (2007).
13. Ng, B. G. & Freeze, H. H. Perspectives on Glycosylation and Its Congenital Disorders. *Trends in Genetics* **34**, 466–476 (2018).
14. Mousa, J., Veres, L., Mohamed, A., De Graef, D. & Morava, E. Acetazolamide treatment in late onset CDG type 1 due to biallelic pathogenic DHDDS variants. *Mol Genet Metab Rep* **32**, 100901 (2022).
15. Ondruskova, N., Cechova, A., Hansikova, H., Honzik, T. & Jaeken, J. Congenital disorders of glycosylation: Still “hot” in 2020. *Biochimica et Biophysica Acta - General Subjects* vol. 1865 Preprint at <https://doi.org/10.1016/j.bbagen.2020.129751> (2021).
16. Jaeken, J. Congenital disorders of glycosylation. in 1737–1743 (2013). doi:10.1016/B978-0-444-59565-2.00044-7.
17. Freeze, H. H., Eklund, E. A., Ng, B. G. & Patterson, M. C. Neurological Aspects of Human Glycosylation Disorders. *Annu Rev Neurosci* **38**, 105–125 (2015).
18. Park, J. H. *et al.* N-glycome analysis detects dysglycosylation missed by conventional methods in SLC39A8 deficiency. *J Inher Metab Dis* **43**, 1370–1381 (2020).
19. Bruneel, A. & Fenaille, F. Integrating mass spectrometry-based plasma (or serum) protein N-glycan profiling into the clinical practice? *Ann Transl Med* **7**, S225–S225 (2019).
20. Abu Bakar, N. *et al.* Synergistic use of glycomics and single-molecule molecular inversion probes for identification of congenital disorders of glycosylation type-1. *J Inher Metab Dis* **45**, 769–781 (2022).

21. Verheijen, J., Tahata, S., Kozicz, T., Witters, P. & Morava, E. Therapeutic approaches in Congenital Disorders of Glycosylation (CDG) involving N-linked glycosylation: an update. *Genetics in Medicine* **22**, 268–279 (2020).
22. Banderali, G., Salvatici, E., Rovelli, V. & Jaeken, J. PMM2-CDG and nephrotic syndrome: A case report. *Clin Case Rep* **10**, (2022).
23. Pajusalu, S. *et al.* The Estimated Prevalence of N-Linked Congenital Disorders of Glycosylation Across Various Populations Based on Allele Frequencies in General Population Databases. *Front Genet* **12**, (2021).
24. *The Proteomics Protocols Handbook*. (Humana Press, 2005). doi:10.1385/1592598900.
25. Andreotti, G., Pedone, E., Giordano, A. & Cubellis, M. V. Biochemical phenotype of a common disease-causing mutation and a possible therapeutic approach for the phosphomannomutase 2-associated disorder of glycosylation. *Mol Genet Genomic Med* **1**, 32–44 (2013).
26. Pirard, M. *et al.* Kinetic properties and tissular distribution of mammalian phosphomannomutase isozymes. *Biochem J* vol. 339 (1999).
27. Citro, V. *et al.* A mutant of phosphomannomutase1 retains full enzymatic activity, but is not activated by IMP: Possible implications for the disease PMM2-CDG. *PLoS One* **12**, e0189629 (2017).
28. Thiel, C., Lübke, T., Matthijs, G., von Figura, K. & Körner, C. Targeted Disruption of the Mouse Phosphomannomutase 2 Gene Causes Early Embryonic Lethality. *Mol Cell Biol* **26**, 5615–5620 (2006).
29. Cromphout, K. *et al.* The Normal Phenotype of Pmm1-Deficient Mice Suggests that Pmm1 Is Not Essential for Normal Mouse Development. *Mol Cell Biol* **26**, 5621–5635 (2006).
30. Matthijs, G., Schollen, E., Van Schaftingen, E., Cassiman, J.-J. & Jaeken, J. Lack of Homozygotes for the Most Frequent Disease Allele in Carbohydrate-Deficient Glycoprotein Syndrome Type 1A. *The American Journal of Human Genetics* **62**, 542–550 (1998).
31. Matthijs, G. *et al.* Mutations in PMM2, a phosphomannomutase gene on chromosome 16p13 in carbohydrate-deficient glycoprotein type I syndrome (Jaeken syndrome). *Nat Genet* **16**, 88–92 (1997).
32. Bjursell, C. *et al.* Detailed mapping of the phosphomannomutase 2 (PMM2) gene and mutation detection enable improved analysis for Scandinavian CDG type I families. *European Journal of Human Genetics* **6**, 603–611 (1998).
33. Matthijs, G., Schollen, E., Heykants, L. & Grünwald, S. Phosphomannomutase Deficiency: The Molecular Basis of the Classical Jaeken Syndrome (CDGS Type Ia). *Mol Genet Metab* **68**, 220–226 (1999).
34. Matthijs, G., Schollen, E., Heykants, L. & Grü, S. MINIREVIEW Phosphomannomutase Deficiency: The Molecular Basis of the Classical Jaeken Syndrome (CDGS Type Ia). *Mol Genet Metab* **68**, 220–226 (1999).
35. Schollen, E. *et al.* Comparative analysis of the phosphomannomutase genes PMM1, PMM2 and PMM2 ψ : the sequence variation in the processed pseudogene is a reflection of the mutations found in the functional gene. *Human Molecular Genetics* vol. 7 (1998).
36. Veiga-da-Cunha, M., Vleugels, W., Maliekal, P., Matthijs, G. & Van Schaftingen, E. Mammalian Phosphomannomutase PMM1 Is the Brain IMP-sensitive Glucose-1,6-bisphosphatase. *Journal of Biological Chemistry* **283**, 33988–33993 (2008).
37. Freeze, H. H. Towards a therapy for phosphomannomutase 2 deficiency, the defect in CDG-Ia patients. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1792**, 835–840 (2009).
38. Quental, R., Moleirinho, A., Azevedo, L. & Amorim, A. Evolutionary History and Functional Diversification of Phosphomannomutase Genes. *J Mol Evol* **71**, 119–127 (2010).
39. Cromphout, K. *et al.* The Normal Phenotype of Pmm1-Deficient Mice Suggests that Pmm1 Is Not Essential for Normal Mouse Development. *Mol Cell Biol* **26**, 5621–5635 (2006).
40. Stiers, K. M., Muenks, A. G. & Beamer, L. J. Biology, Mechanism, and Structure of Enzymes in the α -d -Phosphohexomutase Superfamily. in 265–304 (2017). doi:10.1016/bs.apcsb.2017.04.005.
41. Bateman, A. *et al.* UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* **51**, D523–D531 (2023).
42. Kim, Y. *et al.* Structure- and function-based characterization of a new phosphoglycolate phosphatase from *Thermoplasma acidophilum*. *J Biol Chem* **279**, 517–26 (2004).
43. Koonin, E. V. & Tatusov, R. L. Computer Analysis of Bacterial Haloacid Dehalogenases Defines a Large Superfamily of Hydrolases with Diverse Specificity. *J Mol Biol* **244**, 125–132 (1994).

44. Kedzierski, L. *et al.* Structure of *Leishmania mexicana* Phosphomannomutase Highlights Similarities with Human Isoforms. *J Mol Biol* **363**, 215–227 (2006).
45. Collet, J. F., Stroobant, V., Pirard, M., Delpierre, G. & Van Schaftingen, E. A new class of phosphotransferases phosphorylated on an aspartate residue in an amino-terminal DXDX(T/V) motif. *Journal of Biological Chemistry* **273**, 14107–14112 (1998).
46. Morais, M. C. *et al.* The Crystal Structure of *Bacillus cereus* Phosphonoacetaldehyde Hydrolase: Insight into Catalysis of Phosphorus Bond Cleavage and Catalytic Diversification within the HAD Enzyme Superfamily. *Biochemistry* **39**, 10385–10396 (2000).
47. Briso-Montiano, A. *et al.* Insight on molecular pathogenesis and pharmacochaperoning potential in phosphomannomutase 2 deficiency, provided by novel human phosphomannomutase 2 structures. *J Inherit Metab Dis* **45**, 318–333 (2022).
48. Silvaggi, N. R. *et al.* The X-ray crystal structures of human α -phosphomannomutase 1 reveal the structural basis of congenital disorder of glycosylation type 1a. *Journal of Biological Chemistry* **281**, 14918–14926 (2006).
49. Andreotti, G. *et al.* Conformational response to ligand binding in Phosphomannomutase2: Insights into inborn glycosylation disorder. *Journal of Biological Chemistry* **289**, 34900–34910 (2014).
50. Lahiri, S. D., Zhang, G., Dunaway-Mariano, D. & Allen, K. N. The pentacovalent phosphorus intermediate of a phosphoryl transfer reaction. *Science* (1979) **299**, 2067–2071 (2003).
51. Allen, K. N. & Dunaway-Mariano, D. Catalytic scaffolds for phosphoryl group transfer. *Current Opinion in Structural Biology* vol. 41 172–179 Preprint at <https://doi.org/10.1016/j.sbi.2016.07.017> (2016).
52. Webster, C. E. High-Energy Intermediate or Stable Transition State Analogue: Theoretical Perspective of the Active Site and Mechanism of β -Phosphoglucomutase. *J Am Chem Soc* **126**, 6840–6841 (2004).
53. Allen, K. N. & Dunaway-Mariano, D. Phosphoryl group transfer: evolution of a catalytic scaffold. *Trends Biochem Sci* **29**, 495–503 (2004).
54. Frasse, P. M. *et al.* Enzymatic and structural characterization of HAD5, an essential phosphomannomutase of malaria-causing parasites. *Journal of Biological Chemistry* **298**, 101550 (2022).
55. Lee, Y., Villar, M. T., Artigues, A. & Beamer, L. J. Promotion of Enzyme Flexibility by Dephosphorylation and Coupling to the Catalytic Mechanism of a Phosphohexomutase. *Journal of Biological Chemistry* **289**, 4674–4682 (2014).
56. Altassan, R. *et al.* International clinical guidelines for the management of phosphomannomutase 2-congenital disorders of glycosylation: Diagnosis, treatment and follow up. *J Inherit Metab Dis* **42**, 5–28 (2019).
57. Giurgea, I., Michel, A., Le Merrer, M., Seta, N. & de Lonlay, P. Underdiagnosis of mild congenital disorders of glycosylation type Ia. *Pediatr Neurol* **32**, 121–123 (2005).
58. Stenson, P. D. *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* **133**, 1–9 (2014).
59. Yuste-Checa, P. *et al.* The Effects of PMM2-CDG-Causing Mutations on the Folding, Activity, and Stability of the PMM2 Protein. *Hum Mutat* **36**, 851–860 (2015).
60. Westphal, V. *et al.* A frequent mild mutation in ALG6 may exacerbate the clinical severity of patients with congenital disorder of glycosylation Ia (CDG-Ia) caused by phosphomannomutase deficiency. *Hum Mol Genet* **11**, (2002).
61. Citro, V. *et al.* The analysis of variants in the general population reveals that PMM2 is extremely tolerant to missense mutations and that diagnosis of PMM2-CDG can benefit from the identification of modifiers. *Int J Mol Sci* **19**, (2018).
62. Biddanda, A., Rice, D. P. & Novembre, J. A variant-centric perspective on geographic patterns of human allele frequency variation. *Elife* **9**, (2020).
63. Quelhas, D. *et al.* Assessing the effects of PMM2 variants on protein stability. *Mol Genet Metab* **134**, 344–352 (2021).
64. Vega, A. I. *et al.* Expression analysis revealing destabilizing mutations in phosphomannomutase 2 deficiency (PMM2-CDG). *J Inherit Metab Dis* **34**, 929–939 (2011).
65. Troskie, R., Faulkner, G. J. & Cheetham, S. W. Processed pseudogenes: A substrate for evolutionary innovation. *BioEssays* **43**, 2100186 (2021).
66. Pink, R. C. *et al.* Pseudogenes: Pseudo-functional or key regulators in health and disease? *RNA* **17**, 792–798 (2011).

67. Jones, M. A. *et al.* Molecular diagnostic testing for congenital disorders of glycosylation (CDG): Detection rate for single gene testing and next generation sequencing panel testing. *Mol Genet Metab* **110**, 78–85 (2013).
68. Ligezka, A. N. *et al.* Sorbitol Is a Severity Biomarker for PMM2-CDG with Therapeutic Implications. *Ann Neurol* **90**, 887–900 (2021).
69. Vega, A. I. *et al.* Functional analysis of three splicing mutations identified in the PMM2 gene: Toward a new therapy for congenital disorder of glycosylation type Ia. *Hum Mutat* **30**, 795–803 (2009).
70. Pushpakom, S. *et al.* Drug repurposing: Progress, challenges and recommendations. *Nat Rev Drug Discov* **18**, (2018).
71. Bain, P. G., O'Brien, M. D., Keevil, S. F. & Porter, D. A. Familial periodic cerebellar ataxia: A problem of cerebellar intracellular pH homeostasis. *Ann Neurol* **31**, 147–154 (1992).
72. Izquierdo-Serra, M. *et al.* Stroke-Like Episodes and Cerebellar Syndrome in Phosphomannomutase Deficiency (PMM2-CDG): Evidence for Hypoglycosylation-Driven Channelopathy. *Int. J. Mol. Sci* **19**, 619 (2018).
73. Martínez-Monseny, A. F. *et al.* AZATAX: Acetazolamide Safety and Efficacy in Cerebellar Syndrome in PMM2 Congenital Disorder of Glycosylation (PMM2-CDG) and the CDG Spanish Consortium. *ANN NEUROL* **85**, 740–751 (2019).
74. Cascão, R., Fonseca, J. E. & Moita, L. F. Celastrol: A Spectrum of Treatment Opportunities in Chronic Diseases. *Front Med (Lausanne)* **4**, (2017).
75. Vilas, A. *et al.* Proteostasis regulators as potential rescuers of PMM2 activity. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1866**, 165777 (2020).
76. Mu, T.-W. *et al.* Proteostasis Regulators and Pharmacologic Chaperones Synergize to Correct Protein Misfolding Diseases. *Cell* **134**, 769–781 (2008).
77. Kawai, T. *et al.* Effects of epalrestat, an aldose reductase inhibitor, on diabetic peripheral neuropathy in patients with type 2 diabetes, in relation to suppression of Nε-carboxymethyl lysine. *J Diabetes Complications* **24**, 424–432 (2010).
78. Iyer, S. *et al.* Repurposing the aldose reductase inhibitor and diabetic neuropathy drug epalrestat for the congenital disorder of glycosylation PMM2-CDG. *DMM Disease Models and Mechanisms* **12**, (2019).
79. Liguori, L. *et al.* Pharmacological chaperones: A therapeutic approach for diseases caused by destabilizing missense mutations. *International Journal of Molecular Sciences* vol. 21 Preprint at <https://doi.org/10.3390/ijms21020489> (2020).
80. Convertino, M., Das, J. & Dokholyan, N. V. Pharmacological Chaperones: Design and Development of New Therapeutic Strategies for the Treatment of Conformational Diseases. *ACS Chemical Biology* vol. 11 1471–1489 Preprint at <https://doi.org/10.1021/acscchembio.6b00195> (2016).
81. Leidenheimer, N. J. & Ryder, K. G. Pharmacological chaperoning: A primer on mechanism and pharmacology. *Pharmacological Research* vol. 83 10–19 Preprint at <https://doi.org/10.1016/j.phrs.2014.01.005> (2014).
82. Tran, M. L., Génisson, Y., Ballereau, S. & Dehoux, C. Second-generation pharmacological chaperones: Beyond inhibitors. *Molecules* vol. 25 Preprint at <https://doi.org/10.3390/molecules25143145> (2020).
83. Wang, Y. J., Di, X. J. & Mu, T. W. Using pharmacological chaperones to restore proteostasis. *Pharmacological Research* vol. 83 3–9 Preprint at <https://doi.org/10.1016/j.phrs.2014.04.002> (2014).
84. Fan, J.-Q. A counterintuitive approach to treat enzyme deficiencies: use of enzyme inhibitors for restoring mutant enzyme activity. *bchm* **389**, 1–11 (2008).
85. Jorge-Finnigan, A. *et al.* Pharmacological chaperones as a potential therapeutic option in methylmalonic aciduria cblB type. *Hum Mol Genet* **22**, 3680–3689 (2013).
86. Muntau, A. C., Leandro, J., Staudigl, M., Mayer, F. & Gersting, S. W. Innovative strategies to treat protein misfolding in inborn errors of metabolism: pharmacological chaperones and proteostasis regulators. *J Inherit Metab Dis* **37**, 505–523 (2014).
87. Yuste-Checa, P. *et al.* Pharmacological Chaperoning: A Potential Treatment for PMM2-CDG. *Hum Mutat* **38**, 160–168 (2017).
88. Monticelli, M., Liguori, L., Allocca, M., Andreotti, G. & Cubellis, M. V. β-Glucose-1,6-Bisphosphate stabilizes pathological phosphomannomutase2 mutants in vitro and represents a lead compound to develop pharmacological chaperones for the most common disorder of glycosylation, PMM2-CDG. *Int J Mol Sci* **20**, (2019).

89. DiMasi, J. A., Feldman, L., Seckler, A. & Wilson, A. Trends in Risks Associated With New Drug Development: Success Rates for Investigational Drugs. *Clinical Pharmacology and Therapeutics* **87**, 272–277 (2010).
90. Sernadela, P. *et al.* Linked Registries: Connecting Rare Diseases Patient Registries through a Semantic Web Layer. *Hindawi BioMed Research* **2017**, 13 (2017).
91. Richter, T. *et al.* Rare Disease Terminology and Definitions—A Systematic Global Review: Report of the ISPOR Rare Disease Special Interest Group. *Value in Health* **18**, 906–914 (2015).
92. Brasil, S. *et al.* Artificial intelligence (AI) in rare diseases: Is the future brighter? *Genes (Basel)* **10**, (2019).
93. Roessler, H. I., Knoers, N. V. A. M., van Haelst, M. M. & van Haaften, G. Drug Repurposing for Rare Diseases. *Trends Pharmacol Sci* **42**, 255–267 (2021).
94. Xue, H., Li, J., Xie, H. & Wang, Y. Review of Drug Repositioning Approaches and Resources. *Int J Biol Sci* **14**, 1232–1244 (2018).
95. Oprea, T. I. *et al.* Drug repurposing from an academic perspective. *Drug Discov Today Ther Strateg* **8**, 61–69 (2011).
96. Schaefer, J., Lehne, M., Schepers, J., Prasser, F. & Thun, S. The use of machine learning in rare diseases: A scoping review. *Orphanet J Rare Dis* **15**, (2020).
97. Elbadawi, M., Gaisford, S. & Basit, A. W. Advanced machine-learning techniques in drug discovery. *Drug Discov Today* **26**, 769–777 (2021).
98. Patel, V. & Shah, M. A comprehensive study on artificial intelligence and machine learning in drug discovery and drug development. *Intelligent Medicine* (2021) doi:10.1016/j.imed.2021.10.001.
99. Delavan, B. *et al.* Computational drug repositioning for rare diseases in the era of precision medicine. *Drug Discovery Today* vol. 23 Preprint at <https://doi.org/10.1016/j.drudis.2017.10.009> (2018).
100. Engel, T. Basic overview of chemoinformatics. *Journal of Chemical Information and Modeling* vol. 46 2267–2277 Preprint at <https://doi.org/10.1021/ci600234z> (2006).
101. Polanski, J. Chemoinformatics. in *Comprehensive Chemometrics* 635–676 (Elsevier, 2020). doi:10.1016/B978-0-12-409547-2.14327-6.
102. Augustine, E. F., Adams, H. R. & Mink, J. W. Clinical Trials in Rare Disease: Challenges and Opportunities. *J Child Neurol* **28**, 1142–1150 (2013).
103. Challa, A. P. *et al.* Human and Machine Intelligence Together Drive Drug Repurposing in Rare Diseases. *Front Genet* **12**, (2021).
104. *Bioinformatics and Drug Discovery*. vol. 1939 (Springer New York, 2019).
105. *Chemoinformatics*. (Wiley, 2003). doi:10.1002/3527601643.
106. Lu, Z. PubMed and beyond: a survey of web tools for searching biomedical literature. *Database* **2011**, baq036–baq036 (2011).
107. *Small Molecule Drug Discovery*. (Elsevier, 2020). doi:10.1016/C2018-0-03933-X.
108. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* **28**, 31–36 (1988).
109. Daylight. <https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>.
110. Heller, S. R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J Cheminform* **7**, 23 (2015).
111. Dalby, A. *et al.* Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J Chem Inf Comput Sci* **32**, 244–255 (1992).
112. Todeschini, R. & Consonni, V. *Handbook of Molecular Descriptors*. (Wiley, 2000). doi:10.1002/9783527613106.
113. Senese, C. L., Duca, J., Pan, D., Hopfinger, A. J. & Tseng, Y. J. 4D-Fingerprints, Universal QSAR and QSPR Descriptors. *J Chem Inf Comput Sci* **44**, 1526–1539 (2004).
114. Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J Chem Inf Comput Sci* **42**, 1273–1280 (2002).
115. PubChem Substructure Fingerprints. ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_Fingerprints.pdf.
116. Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* **32**, 1466–1474 (2011).

117. Verma, R. P. & Hansch, C. Camptothecins: A SAR/QSAR Study. *Chem Rev* **109**, 213–235 (2009).
118. Shahlaei, M. Descriptor selection methods in quantitative structure-activity relationship studies: A review study. *Chemical Reviews* vol. 113 8093–8103 Preprint at <https://doi.org/10.1021/cr3004339> (2013).
119. Leelananda, S. P. & Lindert, S. Computational methods in drug discovery. *Beilstein Journal of Organic Chemistry* **12**, 2694–2718 (2016).
120. Pereira, F., Latino, D. & Gaudêncio, S. A Chemoinformatics Approach to the Discovery of Lead-Like Molecules from Marine and Microbial Sources En Route to Antitumor and Antibiotic Drugs. *Mar Drugs* **12**, 757–778 (2014).
121. Breiman, L. Random Forests. *Mach Learn* **45**, 5–32 (2001).
122. Weaver, S. & Gleeson, M. P. The importance of the domain of applicability in QSAR modeling. *J Mol Graph Model* **26**, 1315–1326 (2008).
123. Efron, B. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *J Am Stat Assoc* **78**, 316 (1983).
124. Vihinen, M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics* **13**, S2 (2012).
125. Matveieva, M. & Polishchuk, P. Benchmarks for interpretation of QSAR models. *J Cheminform* **13**, 41 (2021).
126. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* **405**, 442–451 (1975).
127. Roy, K., Kar, S. & Das, R. N. *A Primer on QSAR/QSPR Modeling*. (Springer International Publishing, 2015). doi:10.1007/978-3-319-17281-1.
128. Begam, B. F. & Kumar, J. S. Computer Assisted QSAR/QSPR Approaches – A Review. *Indian J Sci Technol* **9**, (2016).
129. Gramatica, P. & Sangion, A. A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology. *J Chem Inf Model* **56**, 1127–1131 (2016).
130. Kiralj, R. & Ferreira, M. M. C. Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. *J Braz Chem Soc* **20**, 770–787 (2009).
131. Mitchell, T. *Machine Learning*. (McGraw-Hill, 1997).
132. Bender, A. & Cortés-Ciriano, I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. *Drug Discov Today* **26**, 511–524 (2021).
133. Brasil, S. *et al.* Systematic Review: Drug Repositioning for Congenital Disorders of Glycosylation (CDG). *International Journal of Molecular Sciences* vol. 23 Preprint at <https://doi.org/10.3390/ijms23158725> (2022).
134. Cutler, A., Cutler, D. R. & Stevens, J. R. Random Forests. in *Ensemble Machine Learning* 157–175 (Springer US, 2012). doi:10.1007/978-1-4419-9326-7_5.
135. Lavecchia, A. Machine-learning approaches in drug discovery: Methods and applications. *Drug Discovery Today* vol. 20 318–331 Preprint at <https://doi.org/10.1016/j.drudis.2014.10.012> (2015).
136. Leach, A. R. & Gillet, V. J. *An Introduction To Chemoinformatics*. (Springer Netherlands, 2007). doi:10.1007/978-1-4020-6291-9.
137. Cortes, C. & Vapnik, V. Support-vector networks. *Mach Learn* **20**, 273–297 (1995).
138. Mammone, A., Turchi, M. & Cristianini, N. Support vector machines. *WIREs Computational Statistics* **1**, 283–289 (2009).
139. Mohan, V., Gibbs, A., Cummings, M., Jaeger, E. & DesJarlais, R. Docking: Successes and Challenges. *Curr Pharm Des* **11**, 323–333 (2005).
140. Meng, X.-Y., Zhang, H.-X., Mezei, M. & Cui, M. Molecular Docking: A powerful approach for structure-based drug discovery. *Curr Comput Aided Drug Des.* **7**, 146–157 (2011).
141. Stanzione, F., Giangreco, I. & Cole, J. C. Use of molecular docking computational tools in drug discovery. in 273–343 (2021). doi:10.1016/bs.pmch.2021.01.004.
142. Huey, R., Morris, G. M., Olson, A. J. & Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *J Comput Chem* **28**, 1145–1152 (2007).

143. Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J Chem Inf Model* **61**, 3891–3898 (2021).
144. Fischer, E. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft* **27**, 2985–2993 (1894).
145. Koshland, D. E. Correlation of Structure and Function in Enzyme Action. *Science* (1979) **142**, 1533–1541 (1963).
146. Frenkel, D. *Understanding Molecular Simulation From Algorithms to Applications*. (Academic Press, 2007).
147. Gilson, M. K. & Zhou, H.-X. Calculation of Protein-Ligand Binding Affinities. *Annu Rev Biophys Biomol Struct* **36**, 21–42 (2007).
148. Huey, R., Morris, G. M., Olson, A. J. & Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *J Comput Chem* **28**, 1145–1152 (2007).
149. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* **30**, 2785–2791 (2009).
150. Åqvist, J., Luzhkov, V. B. & Brandsdal, B. O. Ligand Binding Affinities from MD Simulations. *Acc Chem Res* **35**, 358–365 (2002).
151. Verkhivker, G. M. *et al.* Deciphering common failures in molecular docking of ligand-protein complexes. *J Comput Aided Mol Des* **14**, 731–51 (2000).
152. Feher, M., Deretey, E. & Roy, S. BHB: a simple knowledge-based scoring function to improve the efficiency of database screening. *J Chem Inf Comput Sci* **43**, 1316–27 (2003).
153. Nikolic, M., Sustersic, T. & Filipovic, N. In vitro Models and On-Chip Systems: Biomaterial Interaction Studies With Tissues Generated Using Lung Epithelial and Liver Metabolic Cell Lines. *Front Bioeng Biotechnol* **6**, (2018).
154. Plikus, M. V. *et al.* Fibroblasts: Origins, definitions, and functions in health and disease. *Cell* **184**, 3852–3872 (2021).
155. Fernandes, I. R. *et al.* Fibroblast sources: Where can we get them? *Cytotechnology* **68**, 223–228 (2016).
156. Murray, L. A., Knight, D. A. & Laurent, G. J. Fibroblasts. in *Asthma and COPD* 193–200 (Elsevier, 2009). doi:10.1016/B978-0-12-374001-4.00015-8.
157. Vangipuram, M., Ting, D., Kim, S., Diaz, R. & Schüle, B. Skin Punch Biopsy Explant Culture for Derivation of Primary Human Fibroblasts. *Journal of Visualized Experiments* (2013) doi:10.3791/3779.
158. Villegas, J. & McPhaul, M. Establishment and Culture of Human Skin Fibroblasts. *Curr Protoc Mol Biol* **71**, (2005).
159. Ravikanth, M., Manjunath, K., Ramachandran, C., Soujanya, P. & Saraswathi, T. Heterogeneity of fibroblasts. *Journal of Oral and Maxillofacial Pathology* **15**, 247 (2011).
160. Mesdom, P. *et al.* Human Dermal Fibroblast: A Promising Cellular Model to Study Biological Mechanisms of Major Depression and Antidepressant Drug Response. *Curr Neuropsychopharmacol* **18**, 301–318 (2020).
161. Brasil, S. *et al.* CDG therapies: From bench to bedside. *International Journal of Molecular Sciences* vol. 19 Preprint at <https://doi.org/10.3390/ijms19051304> (2018).
162. Bella, J. & Rossmann, M. G. ICAM-1 receptors and cold viruses. *Pharm Acta Helv* **74**, 291–297 (2000).
163. Lawson, C. & Wolf, S. ICAM-1 signaling in endothelial cells. *Pharmacological Reports* **61**, 22–32 (2009).
164. Ramos, T. N., Bullard, D. C. & Barnum, S. R. ICAM-1: Isoforms and Phenotypes. *The Journal of Immunology* **192**, 4469–4474 (2014).
165. Hubbard, A. K. & Rothlein, R. Intercellular adhesion molecule-1 (ICAM-1) expression and cell signaling cascades. *Free Radic Biol Med* **28**, 1379–1386 (2000).
166. Scott, D. W. & Patel, R. P. Endothelial heterogeneity and adhesion molecules N-glycosylation: Implications in leukocyte trafficking in inflammation. *Glycobiology* **23**, 622–633 (2013).
167. He, P., Srikrishna, G. & Freeze, H. H. N-glycosylation deficiency reduces ICAM-1 induction and impairs inflammatory response. *Glycobiology* **24**, 392–398 (2014).
168. Bruneel, A., Cholet, S., Tran, N. T., Mai, T. D. & Fenaille, F. CDG biochemical screening: Where do we stand? *Biochimica et Biophysica Acta (BBA) - General Subjects* **1864**, 129652 (2020).
169. Hnasko, T. S. & Hnasko, R. M. The Western Blot. in 87–96 (2015). doi:10.1007/978-1-4939-2742-5_9.
170. Kim, B. Western Blot Techniques. in 133–139 (2017). doi:10.1007/978-1-4939-6990-6_9.

171. Gwozdz, T. & Dorey, K. Western Blot. in *Basic Science Methods for Clinical Researchers* 99–117 (Elsevier, 2017). doi:10.1016/B978-0-12-803077-6.00006-0.
172. Meftahi, G. H., Bahari, Z., Zarei Mahmoudabadi, A., Iman, M. & Jangravi, Z. Applications of western blot technique: From bench to bedside. *Biochemistry and Molecular Biology Education* **49**, 509–517 (2021).
173. Jones, M. A. *et al.* DDOST Mutations Identified by Whole-Exome Sequencing Are Implicated in Congenital Disorders of Glycosylation. *The American Journal of Human Genetics* **90**, 363–368 (2012).
174. Schaftingen, E. Van & Jaeken, J. *Phosphomannomutase deficiency is a cause of carbohydrate-deficient glycoprotein syndrome type I. FEBS Letters* vol. 377 (1995).
175. De Koning, T. J. *et al.* A Novel Disorder of N-Glycosylation Due to Phosphomannose Isomerase Deficiency. *BIO-CHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS* vol. 245 (1998).
176. Rover, L. *et al.* Study of NADH Stability Using Ultraviolet–Visible Spectrophotometric Analysis and Factorial Design. *Anal Biochem* **260**, 50–55 (1998).
177. Held, P. Determination of NADH Concentrations with the Synergy™ 2 Multi-Detection Microplate Reader using Fluorescence or Absorbance. <https://www.biotek.com/resources/application-notes/determining-nadh-concentrations-with-synergy-2-multi-mode-microplate-reader-using-fluorescence-or-absorbance/> (2006).
178. Aykul, S. & Martinez-Hackert, E. Determination of half-maximal inhibitory concentration using biosensor-based protein interaction analysis. *Anal Biochem* **508**, 97–103 (2016).
179. Newton, P., Harrison, P. & Clulow, S. A Novel Method for Determination of the Affinity of Protein: Protein Interactions in Homogeneous Assays. *SLAS Discovery* **13**, 674–682 (2008).
180. Sevinc, A. Web of science: a unique method of cited reference searching. *J Natl Med Assoc* **96**, 980–3 (2004).
181. Kim, S. *et al.* PubChem 2023 update. *Nucleic Acids Res* **51**, D1373–D1380 (2023).
182. Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res* **45**, D945–D954 (2017).
183. Irwin, J. J. *et al.* ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J Chem Inf Model* **60**, 6065–6073 (2020).
184. Frank, E., Hall, M. & Witten, I. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques*. (Morgan Kaufmann Publishers, 2016).
185. R Foundation for Statistical Computing. R: A Language and Environment for Statistical Computing. . [http:// www.R-project.org](http://www.R-project.org) (2014).
186. Liaw, A. & Wiener, M. R News. *Classification and Regression by Randomforest* 18–22 <http://CRAN.R-project.org/doc/Rnews/> (2002).
187. Teixeira, A. L., Leal, J. P. & Falcao, A. O. Random forests for feature selection in QSPR Models - an application for predicting standard enthalpy of formation of hydrocarbons. *J Cheminform* **5**, 9 (2013).
188. Chollet, F. Keras. <https://keras.io> (2015).
189. Abadi, M. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv Forum* <https://arxiv.org/abs/1603.04467> (2016).
190. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminform* **7**, 20 (2015).
191. Gohlke, B. O. *et al.* 2D and 3D similarity landscape analysis identifies PARP as a novel off-target for the drug Vatalanib. *BMC Bioinformatics* **16**, (2015).
192. Landrum, G. RDKit: Open-Source Cheminformatics Software. https://github.com/rdkit/rdkit/releases/tag/Release_2016_09_4 (2016).
193. Zagidullin, B., Wang, Z., Guan, Y., Pitkänen, E. & Tang, J. Comparative analysis of molecular fingerprints in prediction of drug combination effects. *Brief Bioinform* **22**, (2021).
194. Halgren, T. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J Comput Chem* (1996).
195. Zhang, Q. *et al.* A QSPR approach for the fast estimation of DFT/NBO partial atomic charges. *Chemometrics and Intelligent Laboratory Systems* **134**, 158–163 (2014).

196. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* NA-NA (2009) doi:10.1002/jcc.21334.
197. Berman, H. M. The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
198. Carvalho, A. L., Santos-Silva, T., Romao, M. J., Cabrita, E. J. & Marcelo, F. Structural Elucidation of Macromolecules. in *Essential Techniques for Medical and Life Scientists: A Guide to Contemporary Methods and Current Applications with the Protocols* 30–91 (BENTHAM SCIENCE PUBLISHERS, 2018). doi:10.2174/9781681087092118010005.
199. O’Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *J Cheminform* **3**, 33 (2011).
200. Weiner, S. J. *et al.* A new force field for molecular mechanical simulation of nucleic acids and proteins. *J Am Chem Soc* **106**, 765–784 (1984).
201. Wade, R. C. & Goodford, P. J. Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 2. Ligand probe groups with the ability to form more than two hydrogen bonds. *J Med Chem* **36**, 148–156 (1993).
202. Mehler, E. L. & Solmajer, T. Electrostatic effects in proteins: comparison of dielectric and charge models. *‘Protein Engineering, Design and Selection’* **4**, 903–910 (1991).
203. Stouten, P. F. W., Frömmel, C., Nakamura, H. & Sander, C. An Effective Solvation Term Based on Atomic Occupancies for Use in Protein Simulations. *Mol Simul* **10**, 97–120 (1993).
204. Pettersen, E. F. *et al.* UCSF Chimera?A visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605–1612 (2004).
205. Wallace, A. C., Laskowski, R. A. & Thornton, J. M. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *‘Protein Engineering, Design and Selection’* **8**, 127–134 (1995).
206. Veskoukis, A. S., Margaritelis, N. V., Kyparos, A., Paschalis, V. & Nikolaidis, M. G. Spectrophotometric assays for measuring redox biomarkers in blood and tissues: the NADPH network. *Redox Report* **23**, 47–56 (2018).
207. Niesen, F. H., Berglund, H. & Vedadi, M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nat Protoc* **2**, 2212–2221 (2007).
208. Dahl, R. *et al.* Potent, Selective, and Orally Available Benzoisothiazolone Phosphomannose Isomerase Inhibitors as Probes for Congenital Disorder of Glycosylation Ia. *J Med Chem* **54**, 3661–3668 (2011).
209. Bravo, Y. *et al.* Design, synthesis and evaluation of benzoisothiazolones as selective inhibitors of PHOSPHO1. *Bioorg Med Chem Lett* **24**, 4308–4311 (2014).
210. Kalliokoski, T., Kramer, C., Vulpetti, A. & Gedeck, P. Comparability of Mixed IC50 Data – A Statistical Analysis. *PLoS One* **8**, e61007 (2013).
211. Willighagen, E. L. *et al.* The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J Cheminform* **9**, 33 (2017).
212. Hall, L. H. & Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J Chem Inf Comput Sci* **35**, 1039–1045 (1995).
213. Hsu, H.-H., Hsu, Y.-C., Chang, L.-J. & Yang, J.-M. An integrated approach with new strategies for QSAR models and lead optimization. *BMC Genomics* **18**, 104 (2017).
214. Lewis, R. A. A General Method for Exploiting QSAR Models in Lead Optimization. *J Med Chem* **48**, 1638–1648 (2005).
215. Scornet, E., Biau, G. & Vert, J.-P. Consistency of random forests. *The Annals of Statistics* **43**, (2015).
216. Probst, P. & Boulesteix, A. To Tune or Not to Tune the Number of Trees in Random Forest. *JMLR* **18**, 1–18 (2018).
217. Oshiro, T. M., Perez, P. S. & Baranauskas, J. A. How Many Trees in a Random Forest? in 154–168 (2012). doi:10.1007/978-3-642-31537-4_13.
218. Yerushalmy, J. Statistical Problems in Assessing Methods of Medical Diagnosis, with Special Reference to X-Ray Techniques. *Public Health Reports (1896-1970)* **62**, 1432 (1947).
219. Bénard, C., Da Veiga, S. & Scornet, E. Mean decrease accuracy for random forests: inconsistency, and a practical solution via the Sobol-MDA. *Biometrika* **109**, 881–900 (2022).

220. Chen, Y., Rosenkranz, C., Hirte, S. & Kirchmair, J. Ring systems in natural products: structural diversity, physico-chemical properties, and coverage by synthetic compounds. *Nat Prod Rep* **39**, 1544–1556 (2022).
221. Aldeghi, M., Malhotra, S., Selwood, D. L. & Chan, A. W. E. Two- and Three-dimensional Rings in Drugs. *Chem Biol Drug Des* **83**, 450–461 (2014).
222. Taylor, R. D., MacCoss, M. & Lawson, A. D. G. Rings in Drugs. *J Med Chem* **57**, 5845–5859 (2014).
223. Lowe, D. The Most Common Heterocycles in Drugs. *Science* <https://www.science.org/content/blog-post/most-common-heterocycles-drugs> (2014).
224. Scott, K. A. & Njardarson, J. T. Analysis of US FDA-Approved Drugs Containing Sulfur Atoms. *Top Curr Chem* **376**, 5 (2018).
225. Burden, F. R. Molecular identification number for substructure searches. *J Chem Inf Comput Sci* **29**, 225–227 (1989).
226. Doshi, S. Various Optimization Algorithms For Training Neural Network. <https://towardsdatascience.com/optimizers-for-training-neural-network-59450d71caf6> (2019).
227. Amari, S., Murata, N., Muller, K.-R., Finke, M. & Yang, H. H. Asymptotic statistical theory of overtraining and cross-validation. *IEEE Trans Neural Netw* **8**, 985–996 (1997).
228. Bilbao, I. & Bilbao, J. Overfitting problem and the over-training in the era of data: Particularly for Artificial Neural Networks. in *2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS)* 173–177 (IEEE, 2017). doi:10.1109/INTELCIS.2017.8260032.
229. Zakharov, A. V., Peach, M. L., Sitzmann, M. & Nicklaus, M. C. QSAR Modeling of Imbalanced High-Throughput Screening Data in PubChem. *J Chem Inf Model* **54**, 705–712 (2014).
230. Swinney, D. C. Molecular Mechanism of Action (MMoA) in Drug Discovery. in 301–317 (2011). doi:10.1016/B978-0-12-386009-5.00009-6.
231. Hendriks, B. S. Functional pathway pharmacology: chemical tools, pathway knowledge and mechanistic model-based interpretation of experimental data. *Curr Opin Chem Biol* **14**, 489–497 (2010).
232. Stanton, D. T. & Jurs, P. C. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies. *Anal Chem* **62**, 2323–2329 (1990).
233. Katritzky, A. R., Mu, L., Lobanov, V. S. & Karelson, M. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *J Phys Chem* **100**, 10400–10407 (1996).
234. Petitjean, M. Applications of the radius-diameter diagram to the classification of topological and geometrical shapes of chemical compounds. *J Chem Inf Comput Sci* **32**, 331–337 (1992).
235. Gasteiger, J., Schuur, J., Selzer, P., Steinhauer, L. & Steinhauer, V. Finding the 3D structure of a molecule in its IR spectrum. *Fresenius J Anal Chem* **359**, 50–55 (1997).
236. Pollastri, M. P. Overview on the Rule of Five. *Curr Protoc Pharmacol* **49**, (2010).
237. Chessick, C. A. *et al.* Azapirones for generalized anxiety disorder. *Cochrane Database of Systematic Reviews* **2015**, (2006).
238. Parnham, M. J. & Sies, H. The early research and development of ebselen. *Biochem Pharmacol* **86**, 1248–1253 (2013).
239. Feinstein, W. P. & Brylinski, M. Calculating an optimal box size for ligand docking and virtual screening against experimental and predicted binding pockets. *J Cheminform* **7**, 18 (2015).
240. Srivastava, V. K. & Yadav, R. Isothermal titration calorimetry. in *Data Processing Handbook for Complex Biological Data Sources* 125–137 (Elsevier, 2019). doi:10.1016/B978-0-12-816548-5.00009-5.
241. Borea, P. A., Varani, K., Gessi, S., Gilli, P. & Dalpiaz, A. Receptor binding thermodynamics as a tool for linking drug efficacy and affinity. *Il Farmaco* **53**, 249–254 (1998).
242. Wan, S., Bhati, A. P., Zasada, S. J. & Coveney, P. V. Rapid, accurate, precise and reproducible ligand–protein binding free energy prediction. *Interface Focus* **10**, 20200007 (2020).
243. McConathy, J. & Owens, M. J. Stereochemistry in Drug Action. *Prim Care Companion CNS Disord* **5**, (2003).
244. Belhassan, A. *et al.* In silico detection of potential inhibitors from vitamins and their derivatives compounds against SARS-CoV-2 main protease by using molecular docking, molecular dynamic simulation and ADMET profiling. *J Mol Struct* **1258**, 132652 (2022).

245. Aldred, K. J., Kerns, R. J. & Osheroﬀ, N. Mechanism of Quinolone Action and Resistance. *Biochemistry* **53**, 1565–1574 (2014).
246. Day, R. O., Graham, G. G. & Williams, K. Propionic Acid Derivative Drugs (Profens). in *Compendium of Inflammatory Diseases* 1106–1110 (Springer Basel, 2016). doi:10.1007/978-3-7643-8550-7_58.
247. Martino, E. *et al.* The long story of camptothecin: From traditional medicine to drugs. *Bioorg Med Chem Lett* **27**, 701–707 (2017).
248. Wimer, S. M., Schoonover, L. & Garrison, M. W. Levofloxacin: a therapeutic review. *Clin Ther* **20**, 1049–1070 (1998).
249. Candel, F. J. & Peñuelas, M. Delafloxacin: design, development and potential place in therapy. *Drug Des Devel Ther* **Volume11**, 881–891 (2017).
250. Vila, J. *et al.* Ozenoxacin: a review of preclinical and clinical efficacy. *Expert Rev Anti Infect Ther* **17**, 159–168 (2019).
251. Haria, M. & Lamb, H. M. Trovafloxacin. *Drugs* **54**, 435–445 (1997).
252. Wagstaff, A. J. & Balfour, J. A. Grepafloxacin. *Drugs* **53**, 817–824 (1997).
253. Fox, S. & Johnston, S. Use of carprofen for the treatment of pain and inflammation in dogs. *J Am Vet Med Assoc* **210**, 1493–1498 (1997).
254. Davies, N. M. Clinical Pharmacokinetics of Flurbiprofen and its Enantiomers. *Clin Pharmacokinet* **28**, 100–114 (1995).
255. Abrego, G. *et al.* Biopharmaceutical profile of pranoprofen-loaded PLGA nanoparticles containing hydrogels for ocular administration. *European Journal of Pharmaceutics and Biopharmaceutics* **95**, 261–270 (2015).
256. Herben, V. M. M., ten Bokkel Huinink, W. W. & Beijnen, J. H. Clinical Pharmacokinetics of Topotecan. *Clin Pharmacokinet* **31**, 85–102 (1996).
257. Marchi, E., Mangone, M., Zullo, K. & O'Connor, O. A. Pralatrexate Pharmacology and Clinical Development. *Clinical Cancer Research* **19**, 6657–6661 (2013).
258. Danenberg, P. V. *et al.* Folates as adjuvants to anticancer agents: Chemical rationale and mechanism of action. *Crit Rev Oncol Hematol* **106**, 118–131 (2016).
259. Olfat, N., Ashoori, M. & Saedisomeolia, A. Riboflavin is an antioxidant: a review update. *British Journal of Nutrition* **128**, 1887–1895 (2022).
260. Chen, Y. *et al.* Clinical Pharmacology of Axitinib. *Clin Pharmacokinet* **52**, 713–725 (2013).
261. Syed, Y. Y. Lorlatinib: First Global Approval. *Drugs* **79**, 93–98 (2019).
262. Rayner-Hartley, E. & Sedlak, T. Ranolazine: A Contemporary Review. *J Am Heart Assoc* **5**, (2016).
263. Kennelly, M. J. *et al.* Efficacy of Vibegron and Mirabegron for Overactive Bladder: A Systematic Literature Review and Indirect Treatment Comparison. *Adv Ther* **38**, 5452–5464 (2021).
264. Garnock-Jones, K. P. & McKeage, K. Methylnaltrexone. *Drugs* **70**, 919–928 (2010).
265. Riond, J. & Riviere, J. Pharmacology and toxicology of doxycycline. *Vet Hum Toxicol* **30**, (1988).
266. Jain, M. S. & Barhate, S. D. Tazarotene, commonly marketed as Tazorac, Avage, and Zorac is member of the acetylenic class of Retinoids: A Review. *Asian Journal of Pharmaceutical Research* **9**, 206 (2019).
267. Bell, J. Amlexanox for the Treatment of Recurrent Aphthous Ulcers. *Clin Drug Investig* **25**, 555–566 (2005).
268. Awan, M. *et al.* Dimethyl sulfoxide: a central player since the dawn of cryobiology, is efficacy balanced by toxicity? *Regenerative Med* **15**, 1463–1491 (2020).
269. Präbst, K., Engelhardt, H., Ringgeler, S. & Hübner, H. Basic Colorimetric Proliferation Assays: MTT, WST, and Resazurin. in 1–17 (2017). doi:10.1007/978-1-4939-6960-9_1.
270. Pereira, M. I. A. *et al.* Resazurin-Based Assay to Evaluate Cell Viability After Quantum Dot Interaction. in 213–221 (2020). doi:10.1007/978-1-0716-0463-2_12.
271. Vignogna, R. C. *et al.* Evolutionary rescue of phosphomannomutase deficiency in yeast models of human disease. *Elife* **11**, (2022).

APPENDIX

A.1 Considerations

Unless indicated otherwise, all chemical structures presented in this thesis were generated with MarvinSketch version 21.3 (Chemaxon Ltd., Budapest, Hungary).

An Excel Appendix for this thesis can be found on-line at <https://docs.google.com/spreadsheets/d/1DOpm9Ji3Q43rTxEPNvt3-iFvI7uSyzdk/edit?usp=sharing&oid=103504592920775921468&rtpof=true&sd=true>

A.2 Supplementary Figures and Tables

Table 19. Chemaxom Standardizer SMILES standardization list with defined action commands displayed in hierarchical order.

Add explicit hydrogens
Clear isotopes
Mesomerise
Neutralise
Remove solvents
Remove fragments
Strip salts
Tautomerise
Transform aromatic <i>N</i> -oxide
Transform nitro
Transform nitroso
Transform sulfoxide
Transform tertiary <i>N</i> -oxide
Aromatize

Table 20. Recipes for various solutions used throughout the biological experiments.

PBS 10× (makes 1L)	• 17.8 g Na₂HPO₄•2H₂O (final concentration 100 mM) • 2.4 g KH₂PO₄ (final concentration 18 mM) • 80 g NaCl (final concentration 1.37 M) • 2 g KCl (final concentration 27 mM) • Adjust to final volume with dH₂O
PMM2 activity assay buffer 5× (makes 250 mL)	• 5.96 g HEPES (final concentration 100 mM) • 0.595 g MgCl₂ (final concentration 25 mM) • Dissolve in 200 mL d H₂O • Adjust pH to 7.5 with HCl 37% • Adjust to final volume with dH₂O
Separating gel 10% acrylamide (per gel, 1.5 mm-thick)	• 3.1 mL dH₂O • 0.9 mL glycerol • 3.33 mL 30% acrylamide (37.5:1) <i>Rotiphorese® Gel 30</i> (Carl Roth) • 2.5 mL 1.5 M Tris pH 8.8 • 100 µL 10% SDS • 30 µL TEMED • 100 µL 10% APS
Stacking gel 4% acrylamide (per gel, 1.5 mm-thick)	• 1.8 mL dH₂O • 402 µL 30% acrylamide (37.5:1) <i>Rotiphorese® Gel 30</i> (Carl Roth) • 0.75 mL 0.5 M Tris pH 6.8 • 30 µL 10% SDS • 3 µL TEMED • 30 µL 10% APS
Running buffer 10× (makes 1 L)	• 144.4 g glycine (final concentration 1.92 M) • 30.3 g Tris base (final concentration 0.25 M) • 10 g SDS • Dissolve in 800 mL dH₂O • Adjust to final volume with dH₂O
Transfer buffer 10× (makes 1 L)	• 144.4 g glycine (final concentration 1.92 M) • 30.3 g Tris base (final concentration 0.25 M) • Dissolve in 500 mL dH₂O • Adjust to final volume with dH₂O
TBS 10× (makes 1 L)	• 24 g Tris base (final concentration 0.2 M) • 88 g NaCl (final concentration 1.5 M) • Dissolve in 900 mL dH₂O • Adjust pH to 7.6 with HCl 37% • Adjust to final volume with dH₂O
TBS-T 0.1% (v/v) (makes 1L)	• 100 mL TBS 10× • 1 mL Tween® 20 detergent • Dissolve in 900 mL dH₂O

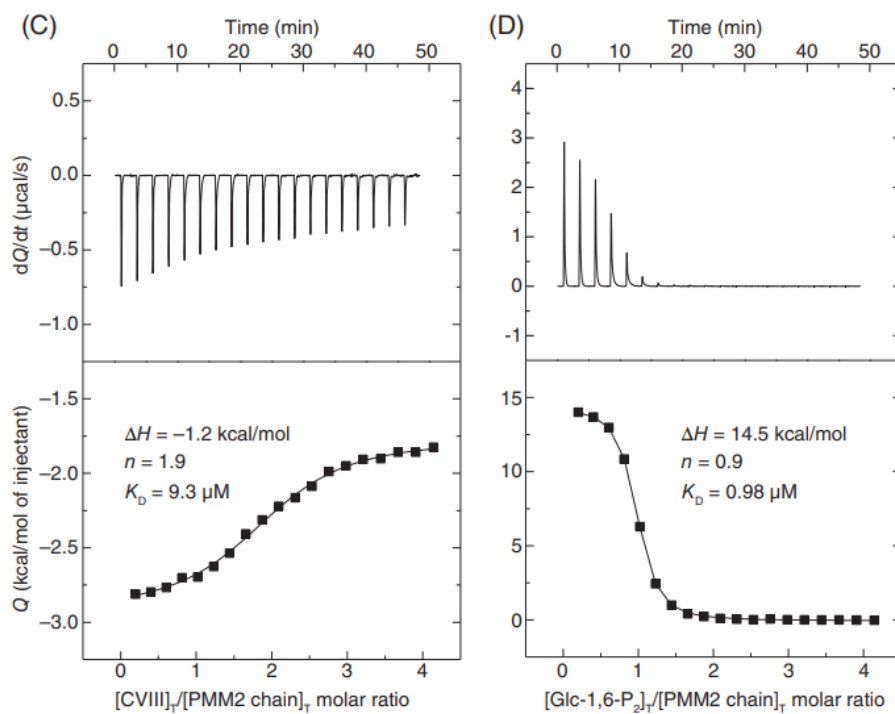


Figure 46. Calorimetric titrations measured for the binding of compound VIII and the Glc 1,6-BP activator to hPMM2.

The upper plots illustrate the thermograms (thermal variation plotted as a function of time) as calculated by isothermal titration calorimetry (ITC) for compound VIII (**C**) and the biphosphate sugar co-factor (**D**). The lower plots show the binding isotherm (normalized heat/injection as a function of the molar ligand/protein ratio). The variation in the enthalpy, ΔH , the number of estimated binding sites for the ligand at the target, n , and their dissociation constants, K_D , are denoted as well. The figure was adapted from⁴⁷.





DESIGN A
DISORD

M. M.

2023