

Harnessing Computational Chemistry and Structural Biology Platforms for *in silico* Drug Discovery

Ben Stuart Cree

Thesis submitted for the degree of
Doctor in Philosophy



School of Natural & Environmental Sciences

Newcastle University

Newcastle upon Tyne

United Kingdom

September 2024

Contents

1	Introduction	1
2	Computer Aided Drug Discovery (CADD)	4
2.1	Theory of Drug Binding	5
2.1.1	Water	6
2.2	Cheminformatics	8
2.2.1	Descriptors	8
2.2.2	Rule of 5 (Ro5)	9
2.2.3	Absorption Distribution Metabolism Excretion Toxicology (ADMET)	10
2.2.4	Quantitative Estimate of Druglikeness (QED)	11
2.3	High Throughput Screening (HTS) and Virtual Libraries (VL)	11
2.3.1	Synthetic Access	12
2.4	Structure Based Drug Design (SBDD)	13
2.4.1	Fragment-based Screening	15
2.4.2	Fragment Growing	16
2.4.3	Fragment Linking	18
2.5	Ligand Based Drug Design (LBDD)	20
2.5.1	Representations of Molecules	21
2.5.2	Similarity	22
2.6	Experimental Assay Techniques	23
2.6.1	Fluorescence Polarization (FP) Assay	23
2.6.2	Surface Plasmon Resonance (SPR)	24
2.7	Summary	27
3	Theory	28
3.1	Computational Chemistry	28
3.2	Quantum Mechanics (QM)	29
3.3	Molecular Dynamics (MD)	33

3.3.1	Integrators	34
3.3.2	Thermostats	35
3.3.3	Force Fields (FF)	36
3.3.4	Conformer Generation	39
3.3.5	Machine Learning Potentials and ML/MM (Machine Learning/Molecular Mechanics)	41
3.3.6	Free Energy Calculations	47
3.4	Docking	52
3.4.1	gnina	53
3.5	Uniform Manifold Approximation (UMAP)	55
3.6	Active Learning (AL)	57
3.6.1	Gradient Boosted Machine (GBM)	57
3.6.2	Gaussian Process	57
4	FEgrow	59
4.1	Introduction	59
4.2	Results	63
4.2.1	Workflow Design	63
4.2.2	Input and Constrained Conformer Generation	63
4.2.3	Geometry Optimisation	64
4.2.4	Binding Pose Scoring	65
4.2.5	Molecular Property Filters	67
4.2.6	Analysis of Lennard-Jones and Electrostatic Scaling Factors.	67
4.3	Case Study I: Protein-Ligand Benchmarks	69
4.4	Discussion	73
4.5	Computational Methods	75
4.5.1	Free Energy Calculations	75
5	CACHE (Critical Assessment of Computational Hit finding Experiments)	77

5.1	CACHE (Critical Assessment of Computational Hit finding Experiments)	
	Challenge 2	78
5.2	Target Background	79
5.3	Round 1: Hit and Free Energy Calculations	82
	5.3.1 Design Strategy	82
	5.3.2 Compound Designs	86
	5.3.3 Docked Enamine Compounds	91
	5.3.4 Synthesis	93
	5.3.5 Round 1 Results	94
5.4	Round 2: Hit Expansion	96
	5.4.1 Round 2 Hit Data	96
	5.4.2 Free Energy Perturbation (FEP) Calculations	99
	5.4.3 Docking	100
5.5	Conclusion	103
6	Active Learning	104
6.1	Introduction	104
6.2	Methods	107
	6.2.1 Workflow Design	107
	6.2.2 Database Search	107
	6.2.3 Computational Details	108
6.3	Results	109
	6.3.1 Interfacing FEgrow with Active Learning Enables Efficient Search of Chemical Space	109
	6.3.2 Active Learning Driven Fragment Expansion Identifies Potential SARS-CoV-2 MPro Inhibitors	114
	6.3.3 Analysis of Hit Compounds	115
6.4	Discussion and Conclusions	119

7 Conclusion	121
7.1 Future of Fragment-based Drug Design	121
7.2 Future of FEgrow	123
S8 Appendix 1: FEgrow Supplementary Information	126
S8.1 Case Study II: SARS-CoV-2 Main Protease	126
S9 Appendix 2: Active Learning Supplementary Information	136
References	146

Abstract

The integration of computational methods in drug discovery has become essential, necessitating the development of open-source, modular, and reproducible workflows that are adaptable to an evolving field. In this context, the synergistic combination of molecular mechanics (MM) and machine learning (ML) offers avenues for expediting the identification and optimisation of potential drug candidates in the hit-to-lead stage. Automated free energy calculations for the prediction of binding free energies of congeneric series of ligands to a protein target are growing in popularity, but choosing transformations and building reliable initial binding poses for the ligands remains challenging.

In this thesis, an open-source python package, FEgrow, is presented. This package automates the required construction and evaluation of congeneric compound series within protein binding pockets, by employing hybrid ML/MM potential energy functions. FEgrow optimises suggested compounds’ bioactive conformers using physics-based methods and scores them using a convolutional neural network (CNN) scoring function, rapidly finding relevant areas of chemical space, as well as generating accurate 3D structures which can then be utilised in more rigorous calculations, such as free energy perturbation (FEP).

This workflow was applied to the CACHE#2 (Critical Assessment of Computational Hit-finding Experiments) Challenge, which serves as a validation exercise that aims to establish benchmarks in molecular design by providing high-quality experimental feedback on *in silico* design predictions. In this challenge, compounds designed via FEgrow yielded multiple low micromolar hits for the NSP13 helicase of SARS-CoV-2, demonstrating the efficacy of the workflow. To further streamline the FEgrow workflow, an updated Active Learning (AL) approach along with the utilisation of large on-demand library searches (such as Enamine REAL) was developed. This addition enhances the exploration of chemical space for *de novo* design and was validated through experimental assays, where three designed compounds showed weak activity in a fluorescence-based M^{pro} assay.

Acknowledgements

This thesis is dedicated to my family, friends and colleagues. I wish to express gratitude to those who have offered their support - no matter the kind, or frequency. I would like to thank Dr. Daniel Cole, my (ever fastidious) supervisor, whose patience and insights have not only facilitated this work, but shaped my scientific thinking. Thanks too to my co-supervisor Dr. Natalie Tatum for her time and support, especially in collaboration with Cancer Research UK.

I am grateful to everyone in the group for providing me with a wonderful environment in which to do science, in particular to Dr. Mateusz Bieniek, whose technical prowess has taught me a great deal and who has helped me through many challenges (and pints). Thanks to both to the CACHE organisation, for running the challenge which forms a large part of this thesis, and to the whole MosMed CDT for organising events, training, as well as project funding.

My parents have always had faith in me, and have been stalwart in their guidance, encouragement and assistance. They have allowed me to take my (educational) journey, and I wish to thank them for their love and kindness - they are an inspiration to me. Finally, I wish to thank my wife, Ruth, whom I love unendingly. She has supported me in more ways than she knows, and I am thankful to have her in my life.

Declaration of Authorship

Parts of this work have been published.

- **Chapter 4** - B. Cree, M. Bieniek, R. Pirie, J. Horton, N. Tatum, and D. Cole, *Commun. Chem.*, 2022, **5**, 136.
- **Chapter 6** - B. Cree, M. Bieniek, S. Amin, A. Kawamura, and D. Cole, *Dig. Discov.*, 2025, **4**.

List of Abbreviations

^{19}F	Fluorine-19 Nuclear Magnetic Resonance Spectroscopy
ACE	Atomic Cluster Expansion
ADMET	Absorption Distribution Metabolism Excretion Toxicology
AL	Active Learning
BAR	Bennett’s Acceptance Ratio
CACHE	Critical Assessment of Computational Hit Finding Experiments
CADD	Computer Aided Drug Design
CDK2	Cyclin Dependent Kinase
CNN	Convolutional Neural Network
CoVs	Coronaviruses
CSD	Cambridge Structural Database
DFT	Density Functional Theory
DG	Distance Geometry
ETKDG	Experimental-Torsion Distance Geometry with Basic Knowledge
FDA	Federal Drug Administration
FEP	Free Energy Perturbation
FF	Force Field
FLOPs	Floating-Point Operations
GBM	Gradient Boosted Machine
GP	Gaussian Process
GPCR	G-protein-coupled Receptors
HBA	Hydrogen Bond Acceptors

HBD Hydrogen Bond Donors

HPC High Performance Computing

HTS High Throughput Screening

kNN k-Nearest Neighbour

LBDD Ligand Based Drug Design

LLM Large Language Model

MBAR Multistate Bennett Acceptance Ratio

MC Monte Carlo

MD Molecular Dynamics

ML Machine Learning

MM Molecular Mechanics

MMGBSA Molecular Mechanics with Generalised Born and Surface Area Solvation

MOE Molecular Operating Environment

Mpro Main Protease

NMR Nuclear Magnetic Resonance

NNP Neural Network Potential

NSP Nonstructural Protein

OpenFF OpenForceField

PAINS Pan-Assay Interference Structure

PCA Principal Component Analysis

PDB Protein Data Bank

PES Potential Energy Surface

PK Pharmacokinetics

QED Quantitative Estimate of Druglikeness

QM Quantum Mechanics

QPU Quantum Processing Units

QSAR Quantitative Structure-Activity Relationship

RMSD Root Mean Square Displacement

RMSE Root Mean Square Error

Ro5 Rule of 5

SA Synthetic Accessibility

SAR Structure Activity Relationship

SARS-CoV-2 Mpro SARS-CoV-2 Main Protease

SB-CADD Structure-Based Computer-Aided Drug Design

SBDD Structure Based Drug Design

SMARTS Simplified Molecular Input Line Entry System Arbitrary Target Specification

SMILES Simplified Molecular Input Line Entry System

SPR Surface Plasmon Resonance

TI Thermodynamic Integration

UCB Upper Confidence Bound

UFF Universal Force Field

VL Virtual Library

XFELs X-ray Free Electron Lasers

List of Figures

1	Binding of a drug molecule to HIV-1 protease (PDB: 1EBZ), illustrating the interaction between an inhibitor and a receptor's (here an enzyme) active site. The drug molecule (ligand) forms interactions with the receptor through hydrogen bonds, hydrophobic interactions, and van der Waals forces, blocking the protease's active site and consequently its ability to process viral polyproteins required for HIV replication.	2
2	A non-physical scheme of an alchemical thermodynamic cycle used to obtain the relative binding free energies for two different, but homologous, ligands A and B (shown in orange and grey, respectively). The relative difference in binding free energy ($\Delta\Delta G_{bind,A\rightarrow B} = \Delta G_{bind,B} - \Delta G_{bind,A}$) can be computed as a difference between two alchemical transformations, $\Delta G_{bound} - \Delta G_{unbound}$, where ΔG_{bound} is the free energy difference for the transformation $A \rightarrow B$ in the receptor and $\Delta G_{unbound}$ is the free energy difference for the solvation for the transformation $A \rightarrow B$ in water.	7
3	Various methods of protein structure determination used in structure-based drug discovery. a) X-ray electron density, with fitted amino acid residues for a portion of myoglobin (PDB: 2NRL), b) NMR ensemble of human ubiquitin ¹ , c) Cryo-EM structure of a bacterial ribosome, with common landmarks/domains labelled.	15
4	An example of two R groups added to a furan ring core with FEgrow. a) carbon linker, N-Methylacetamide R group; b) phenyl linker, formamide R group. R groups shown in purple, with linkers shown in orange. There are in excess of 1 million possible linker/R group combinations, which can be added to replace either a methyl group or a hydrogen.	17
5	chromen-2-one. a) 1D SMILES representation, b) 2D skeletal structure, c) 3D conformer with explicit hydrogens	22
6	In creating an ECFP fingerprint, each non-hydrogen atom is assigned an identifier, and those identifiers are iteratively combined with identifiers of neighbouring atoms until a specified distance (diameter) is reached. Progressive iterations capture circular neighbourhoods of increasing size around each atom, which are then encoded into integer values via a hashing algorithm, which are collected into a list. A schematic of two iterations of this procedure is shown, with distances of 0 and 1 in light and dark grey, respectively.	23

7	A monochromatic laser is directed at a thin metal surface, exciting surface plasmons (oscillations of electron density) at a specific resonance angle. The reflected light is monitored by a detector, which identifies a distinct intensity dip corresponding to the resonance condition. As analytes, such as antigens, bind to immobilized receptors (e.g., antibodies) on the surface, the added mass alters the local refractive index. This change shifts the resonance angle, which is quantitatively recorded by the detector. The magnitude of this shift is directly proportional to the analyte concentration, allowing measurement of affinities for e.g. binding events.	25
8	The two main dimensions of accuracy/complexity in modelling quantum systems. Electron correlation capability is the ability for the level of theory to capture the coupling of electronic motion, by increasing the sophistication of the configuration interaction (CI). Methods such as coupled cluster single-double-triple (CCSDT) include configurations that contain single, double and triple excitations configurations. Basis sets are a set of mathematical functions which can be combined in a linear way to represent the total electronic wavefunction. The basis sets shown here are Slater-Type Orbitals (STO) basis function (equation 11), where a SZ basis set has only a single s-function (for second row elements it has two s functions and one p function), DZ has double the number of functions, TZ has triple, and so on. When both the full CI and basis set limit are reached, the exact solution to the Schrödinger Equation, up to the Born-Oppenheimer approximation, is reached.	32
9	Torsional-angle distributions from the CSD, ETKDG (only acyclic torsion patterns together with the fitted torsional-angle potentials for the first three SMARTS patterns that were fitted. ²	40
10	A single-layered neural network comprised of two input neurons (green), five hidden neurons (blue) and one output neuron (yellow). The weights and biases (eq 24) can be adjusted to minimise the error in the output of the network. For example predicting the dissociation energy of a diatomic molecule, where the input neurons represent the constituent atoms, and the output neuron represents a scalar energy value.	42
11	The competitive accuracy of a ML forcefield, MACE-OFF ³ relative to DFT at the CCSD(T) level of theory. Barrier height was calculated for the torsion angle between two aromatic rings in the biaryl torsion benchmark, ⁴ which contains 78 molecules.	47

12	Thermodynamic cycle used to calculate relative binding free energies ($\Delta\Delta G_{bind} = \Delta G_{bA}^{\circ} - \Delta G_{bB}^{\circ}$) between congeneric ligands. The horizontal legs correspond to the physical binding process, whereas vertical legs indicate the unphysical transformation of ligand A (blue) into ligand B (green) performed in bulk solvent (left) and in the protein binding site (right).	49
13	An example network of alchemical transformations used for calculation of relative binding free energies of 7 inhibitors for a helicase of SARS-CoV2 (section 5.4.1).	50
14	The FEGrow workflow. (left) The user specifies the receptor, ligand core, and a list of functional groups, along with their attachment points. (centre) RDKit ⁵ is used to attach the selected R-group(s) and enumerate the available conformers with a rigid core. (right) Possible bioactive conformers undergo structural optimisation using a hybrid ML/MM potential energy function. The binding affinity is predicted using a convolutional neural network scoring function ⁶ and molecular properties are assessed. Final structures are output for further free energy based binding affinity assessment.	62
15	Two different R groups (purple) and linkers (orange) grown from a common core (grey) in an example protein receptor. The core is restrained but the added groups are kept flexible to be optimised (optionally with a machine learning potential) to produce low energy conformers, which can be input into free energy calculations.	64
16	Overlay of protein–ligand benchmark dataset structures for the BACE(Hunt) target (PDB: 4JPC). Crystal structure in yellow and grown compound in grey. a) including water in the binding pocket as part of the receptor, b) using ANI for optimisation, c) using GAFF for optimisation, d) setting relative permittivity (ϵ) and the Lennard-Jones radii scaling factor to 1.0.	66
17	Effect of the LJ radii scaling factor and relative dielectric permittivity (ϵ) used during optimisation on the correlation between predicted and experimental binding free energies for the set of thrombin inhibitors (see Case Study I).	69

18	Overlay of protein-ligand benchmark dataset structures (crystal structures in yellow and grown compounds in grey). a) TYK2 (PDB: 4GIH ⁷), b) Thrombin (PDB: 2ZFF ⁸), c) P38 (PDB: 3FLY ⁹), d) PTP1B with force field optimisation (PDB: 2QBS ¹⁰), e) PTP1B using ML/MM optimisation, and f) BACE(Hunt) (PDB: 4JPC ¹¹). Root-mean-square distances (RMSD) between predicted and experimental coordinates of atoms in the built R-groups were calculated using RDKit ⁵	70
19	Absolute binding free energies of congeneric series of ligands taken from the protein-ligand benchmark set, using the gnina CNN affinity, compared with experiment. Protein targets from top left: BACE, BACE(Hunt), BACE(P2), CDK2, JNK1, MCL1, P38, PTP1B, Thrombin, TYK2.	71
20	A schematic flowchart of the four constituent phases of a CACHE challenge. This figure is licensed under CC BY 4.0. ¹²	78
21	Timeline of this CACHE challenge.	79
22	CoV genomic architecture and nonstructural proteins (NSPs). A. Schematic of the SARS-CoV-2 genome with the NSPs and the structural and accessory proteins. B. Schematic of SARS-CoV-2 nsp13-HEL. C. A model of probable NSPs 7–16 assembly in a multi-protein complex on viral template RNA, based on biochemical and structural studies. NSP13 helicase [HEL] shown in purple. This figure is licensed under CC BY 4.0. ¹³	80
23	Structure overview of NSP13 (PDB: 6ZSL) with domains labeled and coloured individually, the interface of the 2A and 1B domains (circled in orange) form the binding site for this challenge. Figure is licensed under CC BY 4.0. ¹⁴	81
24	All fragments crystallographic fragments used in the design process; a) 5RMM (blue), b) 5RLH (brown), c) 5RLZ (grey) overlayed, with interacting residues identified (orange).	82
25	Water map constructed from 52 Protein Data Bank (PDB) entries of NSP13. Structures from Newman <i>et al.</i> that had waters in the RNA binding site were collated and overlayed to give areas of the pocket that were frequently hydrated. ¹⁵	83
26	Crystallographic fragments: a) 5RLH (blue), with the trifluoro group occupying the R ₃ sub-pocket, b) 5RMM (grey), with the phenyl group occupying the R ₁ sub-pocket. Key residues shown in orange.	84
27	An exemplar 5-membered ring added to the carboxylate core from 5RLH with FEGrow showing three main vectors used for growth, similar to 5RMM.	84

28	Best scoring <i>de novo</i> design (yellow), with a predicted gnina affinity of 1.4 μ M, with crystal fragments 5RLH (blue) and 5RMM (grey), along with the water map.	85
29	Exemplar 5-membered rings added to the carboxylate fragment (of 5RLZ) with FEGrow: a) pyrazole, b) imidazoline	86
30	An example of a hydrophobic R group design off the R ₁ vector that scored favourably. Oxazole ring, predicted gnina affinity 95 μ M, amide linker, phenyl R group.	87
31	Furan ring. 360 μ M, R ₂ vector, ketone linker, furan R group (purple). The R ₂ vector was not well suited to R group addition due to the positioning of the core that was used, with the nearest polar residue on the other side of the pocket (orange) at a distance of 10.8 Å.	87
32	Furan ring. 28 μ M, R ₃ vector, no linker, benzopyran R-group. The lack of a linker group restricted the ability of the benzopyran to reach the groove.	88
33	Test set of linker atoms from a pyrazole ring with pyridazine R-group. a) Oxygen, b) Carbon, c) Nitrogen, d) No linker, e) Sulphur and f) Ketone.	89
34	An example of a <i>de novo</i> design utilising the C linker (orange), overlayed with 5RLH (blue). Predicted gnina affinity, 43 μ M	90
35	A large molecule grown with FEGrow, incorporating all three vectors with a predicted gnina affinity of 1.2 μ M.	91
36	a) Example docked enamine structure (green), b) 5RLH (blue), with a predicted gnina affinity of 17 μ M.	92
37	A top-ranked Enamine compound, with water map shown in red. gnina predicted affinity 8.7 μ M	92
38	Top docked Enamine compound. predicted gnina affinity 0.85 μ M	93
39	Compound designs of custom molecules to be synthesised in-house.	94
40	Original 9 μ M hit structure as predicted with gnina, and watermap overlayed.	95
41	Steady-state SPR affinity data for the round 1 hit compound, 9 μ M.	96
42	SPR steady-state affinity measurement for the hit from round 1, 124 μ M.	97
43	SPR steady-state affinity measurements for all round 2 hits: a) 49 μ M , b) 64 μ M , c) 87 μ M (orthogonally confirmed via ¹⁹ F NMR), d) 78 μ M	98
44	FEP transformation network for investigating follow-up compounds to the Round 1 hit.	99
45	Compound 100, predicted Δ G to be -0.8 kcal/mol relative to the hit compound.	100
46	Scatter plot of 2200 compounds submitted from all participants, coloured by number of molecules. Predicted pK determined by gnina.	101

47	Comparison of performance according to the two evaluation metrics. This work is 1438. Y axis represents the ability of a protocol to accurately predict active compounds for NSP13. X axis is the combined score for all hit compounds, for each team.	102
48	A) Example building and scoring of a SARS-CoV-2 inhibitor ¹⁶ using the interactive FEgrow workflow ¹⁷ . The fixed core (grey) is extended using a user-defined, flexible linker (pink) and R-group (yellow), and scored using gnina ⁶ . B) Compound libraries with substructures that match the rigid core can now be automatically grown and scored, treating the rest of the molecule as fully flexible. C) Proposed active learning cycle. Compounds are grown, built in the binding pocket and scored with FEgrow. The outputs are used to train a machine learning model, which is used to select the next batch of compounds. Optionally, the chemical space can be seeded using compounds available from on-demand chemical libraries.	106
49	a) The position of the ligand core (in the M ^{pro} active site) and definitions of binding pocket labels, the purple sphere is the hydrogen atom for replacement. b) Histogram of computed pK for the 47 K compound oracle dataset. c) UMAP of entire 47,000 oracle chemical space, coloured by computed pK (the activity limit of 4.5 was arbitrarily set). 2D structures of representative strong binders are included. d) A known, 4 μ M, uracil-based binder. ¹⁸	110
50	Recall and F1 score for diverse initial selection GBM (left) and GP (right) models, and greedy acquisition for identification of top 2 % scoring compounds for different cycle sizes. Error bars show standard errors over five runs.	111
51	Recall and F1 score for diverse initial selection using GP and UCB acquisition (repeating the same protocol three times with different β values) with cycle sizes of 200 (left) and 400 (right) for identification of top 2 % scoring compounds. Error bars show standard errors over five runs.	112
52	F1/recall for Experiment: Random initial molecule selection, GP regression model and greedy acquisition at 5 % as a function of different cycle sizes. .	112
53	Difference in selection for first (left) and final (right) active learning cycles, for a GP model with UCB acquisition function ($\beta = 10$), a cycle size of 200 and a diverse set of starting compounds showing a narrowing into areas predicted to be potent and avoiding unpromising areas.	113

54	Active learning drives improvements in predicted binding affinity. A GP model is used, with UCB acquisition function ($\beta = 0.1$), a cycle size of 200 and a diverse set of starting compounds. The solid horizontal line shows the average score for 377 compounds randomly selected from the REAL database that were built with FEgrow.	114
55	Top-scoring compounds from active learning FEgrow runs of the main protease of SARS-CoV2 (PDB: 5R83) using different scoring functions: a) gnina predicted pK (grey), b) protein-ligand interaction profile (blue), c) combined scoring function (pink) and d) Fragment 5RGI (red and teal) (H-bond donation by Gly143, Ser144, Cys145 and His163), and 5RF7 in green (hydrophobic and H-bond donation with Glu166).	116
56	IC ₅₀ determination of selected compounds with Mpro. Compounds 10 , 12 and 14 were tested at a top concentration of 1000 μ M. Nirmatrelvir was tested at a top concentration of 10 μ M as a positive control. Datapoints presented as mean \pm SD; pIC ₅₀ presented as mean \pm SEM; two biological repeats consisting of three technical replicates. 10 consists of one biological repeat with three technical replicates. Conditions: Mpro (0.2 μ M), 12-hour pre-incubation with compounds, 20 μ M fluorescent substrate, 50 mM Tris-HCl (pH 7.3), 1 mM EDTA and temperature 25°C.	118
57	Predicted bound structures docked via gnina, of compounds 12 (Z1470573089) and 14 (Z8969017446).	118
58	a) Experimental Moonshot compound (literature IC ₅₀ 17 μ M) ¹⁹ and most similar compound from this study, from active learning optimisation of predicted gnina pK ($\beta=10$), b) Experimental Moonshot compound (literature IC ₅₀ 54 μ M) ¹⁹ and most similar compound from this study, from active learning optimisation of predicted pK ($\beta=10$), c) Experimental Moonshot compound (literature IC ₅₀ 57 μ M) ¹⁹ and most similar compound from this study, from active learning optimisation of combination scoring function.	119
S1	a) Cyanophenyl-based M ^{pro} inhibitors. b) X-ray crystal structure of 4 in complex with the protease, with discussed binding pockets labelled. c,d) Uracil-based M ^{pro} inhibitors.	126
S2	Overlay of (a) 5 and PDBID: 7L11, (b) 26 and 7L14, (c) 14 and 7L12, (d) 21 and 7L13. Crystal structures are coloured in yellow, and modelled binding poses in grey. Root-mean-square distances (RMSD) between predicted and experimental coordinates of atoms in the built R-groups were calculated using RDKit ⁵	127

S3	Comparison between free energy calculations and experiment. Binding free energies of 13 analogs of the uracil-based M ^{pro} inhibitors, relative to compound 10 . The error bars indicate one standard error based on least square fitting ²⁰	128
S4	Network of alchemical transformations used for calculation of relative binding free energies of 13 analogs of the uracil-based M ^{pro} inhibitors. . . .	134
S5	Comparison between glna and experiment. Absolute binding free energies of 13 analogs of the uracil-based M ^{pro} inhibitors using the glna CNN affinity.	135
S1	Distribution of molecular weights (MW, Da) for the 47 K compound oracle dataset.	137
S2	Correlation between the predicted pK using a Gaussian process regression model and the oracle predictions. Overall RMSE between the predictions is 0.97 pK units. (Cycle size = 200, diverse initial selection of molecules, UCB acquisition function, $\beta = 10$).	137
S3	Histogram of the number of Enamine molecules added for each experiment, as a fraction of the total number of molecules built.	138
S4	Top 2% activity	139
S5	Top 5% activity	139
S6	F1/recall for Experiment: Random initial molecule selection, GBM regression model and greedy acquisition at 2 and 5 % as a function of different cycle sizes.	139
S7	Top 2% activity	140
S8	Top 5% activity	140
S9	F1/recall for Experiment: Diverse (MaxMin) initial molecule selection, GBM regression model and greedy acquisition at 2 and 5 % as a function of different cycle sizes.	140
S10	Top 2% activity	140
S11	F1/recall for Experiment: Random initial molecule selection, GP regression model and greedy acquisition at 2 % as a function of different cycle sizes.	140
S12	Top 2%	141
S13	Top 5%	141
S14	F1/recall for Experiment: Diverse (MaxMin) initial molecule selection, GP regression model and greedy acquisition at 2 and 5 % as a function of different cycle sizes.	141
S15	Active learning drives improvements in predicted CS scoring function. A GP model is used, with UCB acquisition function ($\beta = 0.1$), a cycle size of 200 and a diverse set of starting compounds.	141

S16	Active learning drives improvements in predicted PLIP scoring function. A GP model is used, with UCB acquisition function ($\beta = 0.1$), a cycle size of 200 and a diverse set of starting compounds.	142
S17	Active learning drives improvements in predicted binding affinity scoring function. A GP model is used, with UCB acquisition function ($\beta = 10$), a cycle size of 200 and a diverse set of starting compounds.	142
S18	2D structures of the Enamine compounds ordered, along with their compound number and Enamine IDs. Note that compound 17 is a control compound taken from a previous study ²¹	143
S19	Initial compound screening for inhibition of Mpro enzyme activity. Compounds were tested for inhibition of Mpro catalytic activity at concentrations of 1000 μ M, 500 μ M and 100 μ M. Compounds 17 and 21 were included as controls. Compounds 12 and 14 reduced the Mpro activity below the threshold (≤ 50 % Mpro activity) at 1000 μ M and were selected for subsequent IC50 analysis. 8 was not chosen for further analysis due to background auto-fluorescent activity. Data represented as mean \pm SD; 2 biological repeats consisting of 3 technical replicates. 10 consists of 1 biological repeat with 3 technical replicates. Conditions: Mpro (0.2 μ M) 12-hour pre-incubation with compounds, 20 μ M fluorescent substrate, 50 mM Tris-HCl (pH 7.3), 1 mM EDTA and temp: 25°C. Compounds 5 and 6 were excluded from the analysis due to poor solubility in assay conditions.	144

List of Tables

1	Effect of the LJ radii scaling factor and relative dielectric permittivity (ϵ) used during optimisation on the accuracy of predicted affinities and structures from FEgrow for the set of thrombin inhibitors (see Case Study I).	69
2	Forward and Backward SOMD Free Energy Differences (MBAR) in kcal/mol. Errors are MBAR estimates.	100
S3	PDB ID, number of R-groups grown, net ligand charge, and 2D common core structure for each target. Attachment vectors are labelled by “-R”. . .	131
S4	(Continued) PDB ID, number of R-groups grown, net ligand charge, and 2D common core structure for each target. Attachment vectors are labelled by “-R”.	132
S5	Root mean square error (RMSE) and correlation coefficient (R^2) between gnina CNN affinities (converted to free energies) and experimental binding free energy, calculated as $RT \times \ln(IC_{50})$	133
S6	Comparison between free energy calculations and experiment. Binding free energies of 13 analogs of the uracil-based M ^{pro} inhibitors, relative to compound 10	136
S7	Cycle closure errors for the network of M ^{pro} inhibitors (Figure S3). Errors are calculated from the raw free energy data from SOMD, averaged over duplicate runs and forward/backward transitions.	136

1 Introduction

Since prehistory, medicines have been deeply important for humanity, and charms, incantations, and plants have always been used to treat afflictions of all kinds. Our modern potions typically consist of compounds that treat or alleviate conditions or symptoms by binding to biological targets in the body. Targets in biological systems are typically macromolecules — such as proteins, nucleic acids, or polysaccharides — that interact with compounds to induce physiological changes. Binding to a given target can change its behaviour (for instance by blocking an active site), ideally in such a way that ameliorates any dysfunction that causes disease, saving lives and reducing suffering.

These potions do not grow on trees — though some do — and tend to require significant ingenuity to concoct. One way to create new medicines is by the *de novo* design of compounds, utilising an understanding of the molecular mechanism of drug action. That is, the physical fact that the strength of binding is determined by simple laws — although emergent complexity prevents equally simple methods from solving the problem of molecular design. To overcome this challenge, a confluence of simulation techniques, structural information gathered from the target through experiment, and activity data are used to build models which are then deployed to design a ligand for therapeutic use. A ligand is a molecule that has potent affinity and selectivity for the active or allosteric site of a given target (a secondary pocket that serves a regulatory function — often harder to successfully drug, but usually offering greater specificity.²²)

Potency and selectivity are attained by exploiting a ligand’s interaction with the receptor, shown in Figure 1 (usually a protein of interest in a disease/condition), which is in turn dictated by its size, shape and electron distribution — that is, the arrangement of its electrons, especially those of reactive functional groups that govern polarity, charge, and bonding upon complexation.²³ Given that estimates of the number of synthesisable molecules are on the order of 10^{24} , finding a small molecule that embodies the required characteristics is a highly nontrivial task.²⁴

Drug design has two main approaches: rational and phenotypic. Rational design uses specific information (structural and/or ligand based) about a target of interest to initially find weak inhibitors, or hits, typically in the millimolar range²⁵ possessing low molecular weights (~ 300 Da). These are subsequently optimised to exhibit stronger binding.

This optimisation process itself has phases. The initial focus is on exclusively decreasing the free energy of binding (increasing affinity), a drive towards a ‘lead’ like compound. This is a compound that demonstrates strong binding but lacks some of the desired physiochemical properties that would truly categorise it as a drug, which can be later engineered. Small molecular fragments are a natural choice for an initial screen, as they are small enough to be easily elaborated while still exhibiting appreciable binding

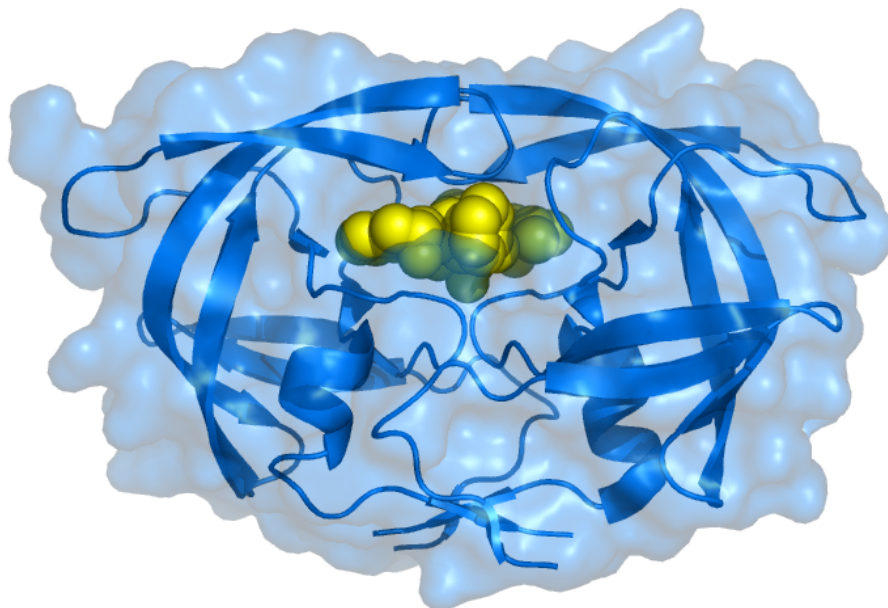


Figure 1: Binding of a drug molecule to HIV-1 protease (PDB: 1EBZ), illustrating the interaction between an inhibitor and a receptor's (here an enzyme) active site. The drug molecule (ligand) forms interactions with the receptor through hydrogen bonds, hydrophobic interactions, and van der Waals forces, blocking the protease's active site and consequently its ability to process viral polyproteins required for HIV replication.

affinities.²⁶ Diverse arrays of fragments can be tested against a particular target, allowing a structure based approach to identify 'hot spots' or areas with multiple key interactions, seen when multiple fragments bind in an active site. These interactions can then be used to elaborate small fragments into larger molecules with higher potency and selectivity. This can be achieved, for example, by linking bound fragments together or growing fragments towards one another. With this 'bottom-up' approach, the relationship between the structure and activity is well understood, especially when 3D structures have been solved via X-ray crystallography. However, translating X-ray structures of protein-ligand complexes and assays to real world efficacy (*in vivo*) remains a speculative task.

The second type of approach is one of screening compounds for their effect both *in vitro* and *in vivo* with neither a specific hypothesis nor any knowledge about the mechanisms of their action. This is referred to as a phenotypic approach, and this target agnostic tactic has proven successful and borne out the development of many first-in-class drugs.²⁷ Phenotypic screening is performed by testing a wide range of compounds against a biological system or cellular signalling pathway and, in the event that a hit is found, it can then be taken as a starting point for lead optimisation, although this is easily confounded by a lack of target information.²⁸ These sort of screens are antipodal to the previous 'bottom-up' strategy and are only concerned with overall efficacy without

attempting to rationalise what is observed. They are especially productive for target identification.²⁹ The rationale behind this sort of approach is an appreciation of the fact that no matter how carefully designed a drug may be, it is very likely that it will have unintended side effects elsewhere. This approach avoids optimising a weaker compound’s affinity in a target-centric fashion, which can potentially compromise target specificity or increase toxicity. This is a common occurrence since compounds at therapeutic levels are liable to interact with various different targets simultaneously, which is an issue that can be obviated with a phenotypic approach.³⁰

Successful creation of a drug can help improve innumerable lives, but it is a long and arduous process, typically taking over a decade and costing in excess of GBP 1 billion.³¹ Reducing the time needed and capital required in order to increase both the quality and quantity of drugs is a prudent focus of industry and academia alike. The methods of *in silico* drug discovery as applied to drug design and their integration into automated workflows are the focus and goal of this thesis.

Chapter 2 broadly introduces the topic of drug design, relevant computational concepts and toolkits, and how they are applied in the context of a drug discovery campaign.

In Chapter 3, the theoretical foundations of both computational chemistry and simulation are presented, along with a discussion of the methodologies implemented in the remainder of the thesis.

Chapter 4 covers the integration of computational chemistry and structural biology platforms, and the development of FEgrow, an open-source workflow for evaluating *de novo* compound design *in silico*. Here, its validation via a retrospective study predicting binding modes of inhibitors for SARS-CoV-2’s main protease (Mpro) is detailed.

Chapter 5 is concerned with the application of FEgrow, in collaboration with medicinal chemists to a real-world (if condensed) drug discovery campaign where inhibitors of a nonstructural protein of SARS-CoV-2 were designed. This work resulted in multiple low micromolar hits, experimentally validated via enzymatic assay, surface plasmon resonance (SPR), and Fluorine-19 nuclear magnetic resonance spectroscopy (¹⁹F NMR).

In Chapter 6, the extension of FEgrow to incorporate active learning-based R-group selection, parallelisation suitable for leverage on HPC clusters, and a redesign of the interface to aid ease of use is demonstrated. This is illustrated through the undertaking of a prospective inhibitor design for M^{pro} of Sars-CoV-2, carried out in collaboration with the Kawamura group, which resulted in weak micromolar hits as determined by inhibition assay.

The thesis is concluded in Chapter 7, where the limitations and potential future developments of FEgrow are discussed, and the overall status of open-source scientific software is explored.

2 Computer Aided Drug Discovery (CADD)

Computers lie at the heart of modern life, and drug discovery is no exception. Creating computational models of ligands and protein-ligand complexes has been a mainstay in the pharmacological industry for decades.³² This is due to the simple fact that these techniques increase the likelihood of creating a successful drug in addition to reducing the time required. Often, these models can be enacted more cheaply than (some) traditional experiments, while also providing certain types of data that are unattainable in a conventional lab across physical, temporal, and biological scales.³³ The stages of drug discovery to which computational chemistry techniques are applicable range from hit discovery using virtual libraries and machine learning for chemical space exploration,³⁴ to rigorous physics simulations of protein-ligand structures in a physiological environment that give free energies of binding for lead optimisation.³⁵ Simulations of this type are referred to as *in silico* experiments and can rival wet experiments for accuracy,³⁶ and in some cases can even refine traditional experimental results.³⁷

The origins of computational chemistry interrogating condensed matter systems can be traced to 1957, with the time-dependant simulation of a monoatomic gas as hard spheres.³⁸ Modern incarnations of these simulations are referred to as Molecular Dynamics (MD) and are composed of complex non-rigid molecules that utilise experimental or quantum data to create classical 3D models of biomolecular systems.

The subsequent ubiquity of computing power since the dawn of computational science has led to a proliferation of computational chemistry research, which has been exploited to achieve impressive feats like simulating entire cells,³⁹ or reaching microsecond time scales for large condensed matter systems.⁴⁰ The field has gone from a niche academic interest to an integral part of both academic and industry toolkits. MD has even been utilised in distributed computing projects involving the general public, with efforts like the COVID Moonshot project using the PCs of volunteers to run molecular dynamics simulations in order to design an inhibitor for COVID-19’s main protease.¹⁹ In addition to changes due to a raw increase in the computing power available, advancements in machine learning (ML) and artificial intelligence (AI) in recent years have caused a paradigm shift in modern computational science. These advances allow the training of models that can predict various properties that are traditionally expensive to compute, such as quantum mechanical energies, kinetics and emergent behaviours of macroscopic systems.⁴¹ Creating more accurate models, especially those that generalise outside the training data used to create them, is the next challenge for the field and is a primary focus that will hopefully enable the avoidance of calculations that are currently prohibitively expensive for use in domains they would otherwise be applicable to.⁴²

2.1 Theory of Drug Binding

A deeper understanding of molecules and the mechanisms behind their action in the body allows us to enact rational strategies to design therapeutic compounds that are extremely potent, selective in what they bind to and non-toxic. With that said, it is surprising how the best examples of drugs we have are irrationally self-created with no knowledge of any medicinal chemistry whatsoever. These are the drugs found in nature, so called ‘natural products’, which exhibit some of the most astounding properties,⁴³ with some still even being beyond the scope of modern synthetic chemistry.⁴⁴ These are the substances our ancestors extracted from the living world around them, many of which are still used today, or are the origin of the most ubiquitous medicines we have (aspirin is derived from salicylic acid contained within the bark of the willow tree, and penicillin was discovered when it was serendipitously observed that *Penicillium Rubens* had anti-bacterial properties)^{45,46} and natural products have afforded us unique drugs that would have almost certainly not been discovered otherwise.

A ligand binding to a target is not a single process. It is best understood by viewing ligand binding not as an association but as an exchange.⁴⁷ Everything exists in equilibrium, and ligand association is analogous to other equilibria like sugar cyclisation and keto-enol tautomerisation. For a general association reaction between a protein, P , and a ligand, L



where the equilibrium constant, known as the association constant K_A , is defined as

$$K_A = \frac{[PL]}{[P][L]} \quad (2)$$

the inverse of this constant, called the dissociation constant K_D , shares the same unit as concentration, and this characteristic makes K_D physically and chemically significant

$$K_D = \frac{1}{K_A} = \frac{[P][L]}{[PL]} \quad (3)$$

and the Gibbs free energy change of the reaction ΔG is given by

$$\Delta G = -RT \ln K_A = RT \ln K_D \quad (4)$$

where R is the molar gas constant ($\approx 8.314 \text{ J K}^{-1} \text{ mol}^{-1}$) and T is the temperature (typically 298.15 K or 25 °C). The position of this equilibrium is dictated by the free energy change associated with the process (the weighted probability a microstate will occur), so understanding factors that determine the free energy change like electrostatic

interactions, orbital interactions and entropy changes due to the ligand's conformational rigidity upon binding (or the reverse effect of a ligand on the conformation of a protein) dictate how much of a ligand is bound at any one time (at a particular concentration).

The most obvious change upon complexation is the formation of hydrogen bonds, which are responsible for the stabilisation of the 3D structure in all biological macromolecules.⁴⁸ One goal of rational drug design is to form as many of these polar interactions as possible, by pairing up donors to acceptors, increasing not only affinity but specificity. This is achieved by taking advantage of the fact these interactions are less likely to be formed in other binding sites due to the idiosyncratic geometrical restraints required.⁴⁹ Polar interactions are not the only contribution to binding affinity and in many cases are of secondary importance, with non-polar and hydrophobic effects known to dictate affinity, modulating it by orders of magnitude.⁵⁰

If equilibrium (under physiological conditions) favours the bound state, which is essentially a proxy of therapeutic activity, then less of the drug will be required to achieve the same effect, reducing the likelihood of toxicity and side effects. An example highlighting the complexities of rational drug design is the impact of optimisation on solubility. Even small structural modifications to a ligand, intended to improve its affinity for a specific target, can produce unexpected and adverse effects elsewhere. A salient example of this is a series of acyclic secondary amides, where the addition of a hydrophobic methyl group surprisingly increased solubility in the homogenous series. This phenomenon can be explained by the dependence of the solubility on the equilibrium between the solid and solvated phases. Although the methyl group reduced the solvation, it had an even greater destabilising effect on the solid state, ultimately leading to improved overall solubility.⁵¹ Cases like these serve as evidence of how altering structure, even slightly, can lead to surprising outcomes - and the pitfalls that await those who partake in drug discovery.

2.1.1 Water

Water is the solvent of life, and has an important part to play in every biochemical process. For a ligand to bind to a protein, it must first be desolvated as part of a thermodynamic cycle⁵² (a way of quantifying how changes in ligand structure or receptor environment influence binding by breaking the process into intermediate steps, each with a calculable free energy contribution, the sum of which equals zero, Figure 2) so that the ligand can interact with the receptor and the water molecules occupying the prospective binding can be displaced (even in the absence of any bound ligand, water fills voids in the protein that would otherwise be vacant.⁵³).

In protein-ligand complexes water's role is multifaceted, it can either be displaced upon binding or perform a structural role which contributes to the protein's 3D structure,

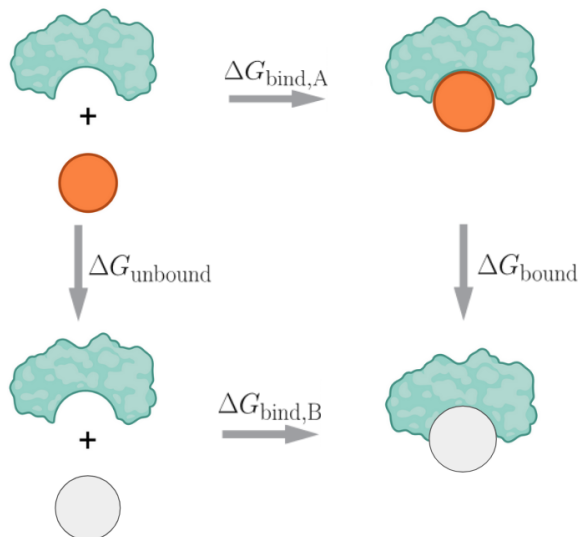


Figure 2: A non-physical scheme of an alchemical thermodynamic cycle used to obtain the relative binding free energies for two different, but homologous, ligands A and B (shown in orange and grey, respectively). The relative difference in binding free energy ($\Delta\Delta G_{bind,A\rightarrow B} = \Delta G_{bind,B} - \Delta G_{bind,A}$) can be computed as a difference between two alchemical transformations, $\Delta G_{bound} - \Delta G_{unbound}$, where ΔG_{bound} is the free energy difference for the transformation $A \rightarrow B$ in the receptor and $\Delta G_{unbound}$ is the free energy difference for the solvation for the transformation $A \rightarrow B$ in water.

almost indistinguishable from that of amino acids.⁵³ Water is the ultimate cause of an indirect force that favours binding, the hydrophobic effect. This effect is commonly understood as a disfavouring of polar moieties being mixed with non-polar for entropic reasons. This is the ‘iceberg formation’ model proposed in 1945,⁵⁴ and according to this model a subset of water molecules closest to the solute adopt an ‘ice-like’ structure, forming a pseudocrystalline cage around the hydrophobe. This model purportedly explains the changes in entropy, enthalpy, and specific heat capacity that characterise the transfer of nonpolar solutes into water and consequent increase in binding affinity. This effect is highly non-specific however, and will increase the affinity of a ligand for all receptors in the body, rendering it an ineffective strategy in general. Even without considering the effect such modifications would have on the PK/ADMET (pharmacokinetics/absorption, distribution, metabolism, excretion, toxicology) profile of the molecule. Modern interpretations of this effect ascribe the volume of exclusion of the solute as the principal cause of the entropy change, not the slight alteration of translational and re-orientational dynamics of the solvation shell in contact with hydrophobic groups.⁵⁵

Water can also serve as a mediator of polar bonds between a ligand and its receptor, allowing the formation of interactions that could not be made otherwise. Water’s amphoteric nature can also be exploited in this way, a simple rotation of a single water

molecule can create ‘nanobuffers’ in a binding pocket which sensitively control the pH of whatever they are interacting with.⁵⁶ Non-structural waters that are not functioning as a conduit for interactions are prone to be displaced by a ligand, and these labile waters can contribute significantly to the free energy of binding due to the increase in entropy upon release into the bulk solvent.⁵⁷ However entropy gain is bounded, with structural waters being unable to have an entropy change larger than that of ice melting, since the water-protein complex is less ordered relative to that of ice.⁵⁸ The overall enthalpic and entropic contributions of a water molecule to the free energy of binding are difficult to predict, as the individual sign of their contribution can be either positive or negative, resulting in six possible permutations of their individual effects, again demonstrating the varied nature of water in the context of protein-ligand complexes.

These waters and the interactions they form are not solitary, and often form a whole network of interactions between themselves and the receptor/ligand. If this network is disrupted (even without directly impacting protein-ligand contacts) the binding affinity can be diminished by a factor of up to 10^3 (and in other cases it is not associated with an increase in affinity whatsoever).⁵² Water’s effects are not limited to entropy contributions or bonded interactions however, with electronic effects also playing a key role on account of its high dielectric permittivity ($\epsilon = 80$), modulating charge interactions in proteins and even altering the pH within an active site itself, due to the local environment (where ϵ is much lower than the bulk solvent, and can effect the pH by as much as 3 units).⁵⁹

2.2 Cheminformatics

It was noted in 1899 by Hans Horst Meyer that the lipophilicity (a molecule’s ability to dissolve in non-polar solvents) of a compound correlated with its efficacy as an anaesthetic, this being the first time the physical properties of a compound were unambiguously and quantitatively linked to a desired therapeutic effect, was the birth of cheminformatics and rational drug discovery.⁶⁰

Cheminformatics focuses on leveraging chemical information—such as molecular structure and physicochemical properties—to identify relationships between molecular features, termed descriptors (Section 2.2.1), and activity data, whether experimental or *in silico*. These relationships enable predictions of properties such as drug-likeness (Section 2.2.4), the presence of undesirable structural elements (Section 2.2.2), and toxicological profiles (Section 2.2.3).

2.2.1 Descriptors

Descriptors are quantities estimated with a heuristic or parametrised approach that characterise molecules, and are widely used in cheminformatics. Translating an arbitrary

chemical structure to that of a set of numbers is useful for many machine learning and cheminformatics tasks, such as solubility prediction or quantitative structure-activity relationship (QSAR) modelling, a method that relates chemical structures to their biological activities using mathematical models e.g. linear regression.⁶¹ These descriptors can represent structural, physiochemical or even geometric properties and can be as simple as 'total molecular weight' (MW) or 'fraction of sp³ carbons' (useful since these molecules tend to be flexible and oily). More complicated descriptors can be derived from parametrised equations such as $c\log P$ which is essentially the sum of the contribution of non-overlapping fragments to the molecules overall solubility. Electronic quantities can be calculated via Gasteiger partial charges,⁶² and geometrical properties can be calculated via conformer embedding schemes such as ETKDG (Experimental-Torsion Distance Geometry with basic Knowledge, the main conformer generation algorithm used in RDKit - see section 3.3.4). Once a molecule has been embedded, quantities such as moments of inertia or radius of gyration can be easily determined.

There are many standard sets of descriptors which are readily accessible, for example the vast array of (over 200) that are present in RDKit. A superset of these descriptors are Mordred descriptors,⁶³ which contains over 1800 different properties.

2.2.2 Rule of 5 (Ro5)

Despite the ferocious complexity of drug discovery, some simple heuristics are regularly used as tools to aid the development of new treatments. A famous example is Lipinski's 'rule of five'⁶⁴, originally inspired by the observation that only a fraction of approved US drugs violate two or more an acceptable range criteria for a select set of properties. These properties can capture complicated parameters to an acceptable degree of accuracy with readily calculated numerical values, such as: molecular weight (MWt) for physical size, hydrogen bond donors (HBD) and acceptors (HBA) for propensity for polar interaction, and calculated lipophilicity ($c\log P$) for lipophilicity. Each of these quantities has an upper limit beyond which is considered undesirable (MWt >500, HBD >5, HBA >10 and $c\log P$ >5), although combinations of different violations can hover around 10 % in FDA approved drugs).⁶⁵ Exceedingly large values for these metrics are not typically seen in drugs due to the effect they have on key factors such as solubility, permeability and promiscuity. For example, the more HBDs a ligand possesses, the more likely it is to form unwanted interactions with (and subsequently show affinity for) off-target receptors in the body. Molecular weight is taken to be a surrogate for a combination of other properties that correlate with size such as lipophilicity, and the number of rotatable bonds is generally used as a proxy for the degree of flexibility of a ligand, which can be liable to induce significant entropy loss upon complexation, disavouring binding.⁶⁶ Models like these are obviously too simple to be used without caution, and efforts have been made to

move ‘beyond’ these prototypical rules.⁶⁷ Nevertheless these rules have utility as filters, especially for extreme values like in the case of controlling aqueous solubility; solubility of a drug is not a requirement for efficacy, however, with Ibuprofen being sparingly soluble in water (0.02 mg/ml) but fully bound to albumin and various other plasma proteins which distribute the drug around the body.⁶⁸

2.2.3 Absorption Distribution Metabolism Excretion Toxicology (ADMET)

A drug that cannot be absorbed by or is toxic to the body is of no use, regardless of how potent an inhibitor it may be. The ability to optimise pharmacodynamics and pharmacokinetics is achieved by tuning a compound’s Absorption, Distribution, Metabolism, Excretion and Toxicology (ADMET), and is a significant focus of later stages of drug development, as well as a sieve for promising compounds entering clinical trials. This challenging multi-parameter optimisation problem causes a ‘valley of death’ where 80% to 90% of research projects fail before they are ever tested in humans, and for every drug that gains FDA approval there were on the order of 10^3 failures.⁶⁹

Once oral treatments are administered, absorption typically occurs via the small intestinal epithelium and occurs readily for lipophilic molecules, with peak absorbance observed for compounds with a ClogP between 4 and 5.⁷⁰ Hydrophilic molecules can still pass through, but must do so through aqueous pores that cover $< 1\%$ of the intestinal surface⁷¹, highlighting the need for tuning lipophilicity to assure good absorbance by the body. This tuning can generally be achieved by replacing polar moieties with non-polar ones, based on their ClogP values (which have generally been shown to be accurate to within an order of magnitude).⁷² The ability to predict pKa values of ionisable groups is another important facet of adjusting permeability, due to charged compounds at physiological pH generally being unable to diffuse through membranes and requiring active transport to access a cell’s interior.⁷⁰

The most basic approach to tuning toxicological properties is to create a substructure filter with known toxic moieties (for example, epoxides are known to cause mutagenicity and thiophenes are hepatotoxic)⁷³. Promiscuous functional groups that interfere with assays, giving false positives or generally invalidating assays, are routinely included in such filters. These substructures are termed Pan-Assay Interference compounds (PAINS), a group of 480 substructures known to invalidate assays via a whole host of mechanisms such as: reactivity with biological and bioassay nucleophiles such as thiols and amines; photoreactivity with any protein functionality; physicochemical; redox cycling and redox activity; micelle formation, or having photochromic properties that might interfere with absorption and fluorescence.⁷⁴ It is of note that the original (protein-protein) assays used to design PAINS filters were performed at rather high concentrations (50 μM) and so it should be expected that any errant behaviour would be less likely under more

typical drug concentrations found *in vivo*.⁷⁵ This fact helps explain why there are > 80 approved drugs containing such groups (~ 5 % of FDA approved drugs contain at least one PAINS substructure). The ability of these simple observational filters (with no consistent ontology) to identify false positives, whilst being able to not preclude true positives has been criticised and has lead to the introduction of more sophisticated methods that go beyond simple substructure queries, translating filters into more general SMARTS strings, and employing refinement using larger databases,⁷⁶ as well as incorporating machine learning classification techniques.⁷⁷

As always, there is a trade-off between accuracy and complexity, with simple models being fast and highly accurate for smaller and more mundane molecules, but are often wildly inaccurate for both larger and exotic species. Recently, there have been attempts to tackle more complicated compounds by training specific machine learning models that cater to bespoke applications, which are not amenable to generic approaches.⁷⁸

2.2.4 Quantitative Estimate of Druglikeness (QED)

Assessing whether or not a particular molecule possesses a structure that could plausibly be a drug, as opposed to a structure that is optimising some scoring function is a key question in CADD. It is a particular worry with the perennial use of generative machine learning models, which are known to generate structures of questionable validity (analogous to the problem of large-language models (LLMs) ‘hallucinating’). Quantitative Estimates of Druglikeness (QED) have been concocted using empirical distributions of molecular properties and are especially useful in filtering vast libraries of compounds ensuring time is not spent evaluating unlikely candidates that could never exist anyway. A simple number that can be associated with the beauty of a drug was first introduced by Harrington.⁷⁹

$$\text{QED} = \exp \left(\frac{1}{n} \sum_{i=1}^n \ln(d_i) \right) \quad (5)$$

where d_i is a desirability function for a property. The modern implementation is taken as a geometric mean of arbitrary individual functions designed to represent the distribution of properties for 771 approved drugs from the DrugStore database.⁸⁰

2.3 High Throughput Screening (HTS) and Virtual Libraries (VL)

Drug discovery is driven by testing molecules, which has the prerequisite of choosing the molecules that are to be tested. Various commercial libraries exist for this purpose, but are intractably large for practical use. Therefore, being able to extract useful subsets

(such as molecules that are similar to a query based on some similarity metric — for example, molecules similar to a known or predicted inhibitor) is of interest.

The rise of on-demand and combinatorial chemistry (the combination of a smaller number of simple building blocks to form a larger library) in recent years has seen a dramatic increase in the size of the libraries that are available. An example is Enamine REAL which currently has approximately 9.6 billion compounds available for purchase (see Chapter 5).⁸¹ This increase in readily accessible screening compounds, along with the exponential increase in computing power, means that larger and more efficacious virtual *in silico* libraries (VL) can be made. VLs are databases of molecules about which predictions of activity (such as biological or chemical interactions, pharmacological effects, or material properties) are generated through *in silico* methods like docking, QSAR modelling or virtual screening. These can be utilised in conjunction with traditional libraries to increase the efficacy of high-throughput methods, with both HTS (high throughput screening) - where a library of 10^3 to 10^6 compounds is screened using an activity assay against a target in an automated fashion - and VLs are a routine part of a drug discovery campaign.²⁵

Although large datasets are a promising trove of experimental data, a drawback of utilising these large chemical datasets (especially those that are experimentally annotated) is that, if the chemical landscape contained within is formed by collating different experiments together, there can be a remarkable deleterious effect on the quality of any model built from such a training set. This is due to methodological differences in the experiments, or even human error.⁸² This amalgamation is an essentially unavoidable fact of life for large datasets. For example, in the ChEMBL database more than two thirds of the datasets have less than 10 distinct compounds (60,000 of 85,000, as of 2024). There are numerous examples of low correlation coefficients from assays performed in different experiments, for identical compounds (in one case for 1400 assays of 38,000 compounds, as low as 0.3, with the apparent inter-experimental noise not even particularly ameliorated with significant levels of curation).⁸²

2.3.1 Synthetic Access

Once promising hits for a target are established, and the process of optimising them into lead compounds is well underway, molecules are generally no longer available from on-demand libraries. Having the capability of adjudicating how feasible predicted structures are to synthesise allows prioritisation and, therefore, minimisation of the time and effort required for a discovery campaign.

Various approaches have been developed to assign predicted scores of synthetic accessibility (SA). One tactic is to use the complexity of the molecule as a proxy for the difficulty of the synthesis, with specific unfavourable groups incurring a penalty, e.g. non-standard fused rings.⁸³ The SA implemented in RDKit⁸⁴ utilises a library of

common fragments derived from a curated subset of PubChem ($\sim 930,000$) with scores of common fragments scored with positive values and atypical fragments scored negatively. This approach relies on the assumption that the frequency has a causal relationship with synthetic tractability, which can be erroneous due to bias in the dataset. For example, the distribution of fragments used to create the scoring function was skewed, and only $<1\%$ of the fragments were present more than 1000 times, and so a large portion of seen fragments are still relatively rare and potentially difficult to synthesise.⁸³ An alternative to the scoring approach is retrosynthetic planning, first implemented computationally in the late 60s⁸⁵ using a PDP-1 to present graphical representation of 3D structures, based on a pruned ‘tree’ of transformations (similar to the aforementioned method, a naive scoring system was manually devised to reduce the number of possible routes). The specific techniques and implementations of the fundamental ideas behind retro-synthetic planning have become more sophisticated, but the idea itself is very much the same since it was first conceived. First, create a list of possible synthetic routes that is as small as possible, then efficiently search the tree and assign priorities to different paths. Modern approaches have utilised neural networks as a replacement of handcrafted rules in the prioritisation of fragmentation of molecules to explore synthetic routes, and Monte Carlo methods have been deployed for searching the resulting tree, allowing the creation of retrosynthetic routes for thousands of molecules within a matter of hours.⁸⁶

Although strides have been made in the prediction of synthetic accessibility (SA) scores, it remains a challenging problem. Solutions often fail to be found due to the commercial availability of reagents and the complexity of the molecules to be synthesised. Even when a route is found, it is not a necessary nor sufficient condition for it actually existing.⁸⁶ This is an issue in particular for *de novo* molecules designed via newer generative machine learning models, which do not learn the specifics of the underlying chemistry, which when they generate molecules are known to give impossible structures, in addition to those that cannot be synthesised.⁸⁷ Being able to integrate a metric that accurately captures the likelihood of a molecule actually being able to be created is a key part of harnessing these models and advancing *in silico* techniques in general.

2.4 Structure Based Drug Design (SBDD)

The structure of a protein gives insights into how it functions as well as how ligands bind. It is also a source of information that can be exploited to inform what modifications to these ligands should be made based on information acquired from various experimental techniques such as X-ray crystallography, Cryo-EM and NMR.^{88,89} Solved structures of proteins (typically with a ligand, over 10% of the Protein Data Bank (PDB) database is a protein-ligand complex)⁹⁰ are the bedrock of SBDD, they give a detailed 3D look into

a protein's active site and ligand binding mode which can elucidate key interactions that are essential for a drug's high affinity for its target.

The scale of SBDD has increased with the availability of high-quality crystal structures, driven by the accessibility of infrastructure such as the UK's national synchrotron (Diamond Light Source, built in 2001) which has the capacity to generate a large quantity of X-ray diffraction data. This increased availability of protein crystallography is a global trend, with the number of structures available in the PDB increasing by over 40% in the past few years.⁹¹

Crystal structures are not a panacea, however, and are only able to provide a static, time-averaged picture of what is a complex and dynamic object.⁸⁹ X-ray crystal structures are, as the name suggests, created from proteins that are in a crystal lattice which can have impacts on binding — for example, if the active site is blocked by symmetry mates (adjacent proteins in the crystal).⁹² These crystal contacts are not typically strong in the context of biomolecular interactions but have been shown to bias towards higher-energy conformations of proteins.⁹³ Collection of X-ray diffraction data requires cooling (typically to $\sim 100\text{K}$), and this non-physiological temperature can affect the protein's structure by reducing thermal motion as well as biasing against more flexible conformations.⁸⁹

There are alternatives to X-ray crystallography (Figure 3), with the second most popular technique (by PDB submissions) being Cryo-EM. Cryo-EM uses 2D images of a (macromolecular) target which is frozen in vitrified ice. Images are created from transmitted electrons, and from these images a 3D model can be reconstructed. This method can give higher resolution structures than aforementioned X-ray crystallography techniques, but has known limitations when used for SBDD, such as mediocre throughput, poor resolution of solvent-accessible areas, and requiring prior knowledge when fitting small ligands into Cryo-EM maps.⁹⁴ Small proteins can be especially difficult to image, and so larger molecules and complexes are favoured for cryo-EM imaging.⁸⁸

Even if perfectly accurate protein structures could be attained, there may even be uncertainty about the exact nature of the structure *in vivo*, since the structures of proteins are controlled at the transcriptional, translational, and post-translational level⁹⁵ - meaning that there are various similar, but not identical - versions of a protein all created from the same coding gene. These are called proteoforms, and increase the functional capacity of a cell.⁹⁶ Having the structure of a single proteoform therefore, can be insufficient for designing inhibitors if multiple proteoforms are integral to the particular cellular process that is to be disrupted.

SBDD can be accelerated with the use of high-quality chemical probes that allow specific mechanistic hypotheses to be tested by modulating biomolecular targets, assuring the validity of any assays that are run.⁹⁷ Chemical probes are molecules that offer high potency and specificity to a given target without PAINS, and are a known result that

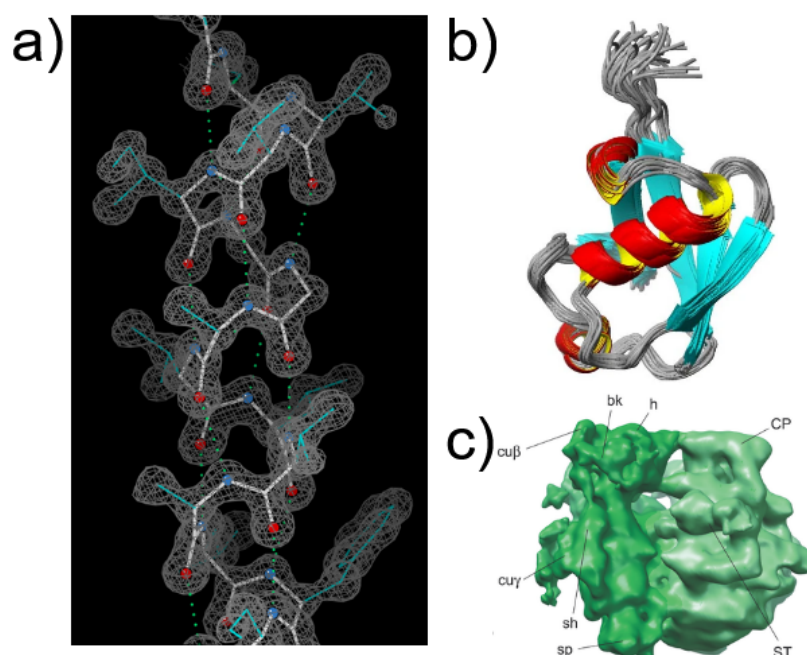


Figure 3: Various methods of protein structure determination used in structure-based drug discovery. a) X-ray electron density, with fitted amino acid residues for a portion of myoglobin (PDB: 2NRL), b) NMR ensemble of human ubiquitin¹, c) Cryo-EM structure of a bacterial ribosome, with common landmarks/domains labelled.

can help identify false positives/negatives. For a given target there should ideally be multiple probes which have mutually orthogonal chemotypes, and having access to effective chemical probes is a key factor in proliferating research for a particular target by facilitating reliable experimentation.

Not all of the human proteome has been explored (genome sequence alone does not provide sufficient information due to post-translational modification of proteins e.g. glycosylation/phosphorylation⁹⁸) with a significant portion being Intrinsically Disordered Proteins/Regions (IDPs/IDRs) which have no stable tertiary structure in their unbound form.⁹⁹ These proteins obviously pose challenges for X-ray crystallography and form part of the ‘dark’ proteome that is inaccessible to classical SBDD techniques.¹⁰⁰

2.4.1 Fragment-based Screening

One method of finding binding pockets is by using a small library of fragments, typically satisfying Lipinski’s rules and containing an assortment of pharmacophoric features and geometric diversity. This fragment-based approach is relatively recent, with its first FDA approved drug, Vemurafenib (a cancer growth blocker) being approved in 2011.¹⁰¹

Small fragments are preferable because of their ability to combinatorially cover chemical space. High-quality library designs are needed and the breadth and complexity

of the chemical space that is spanned is seen as the primary cause of their efficacy.²⁵ The aim is for these small fragments to have significant coverage of a target's binding site, as well as exhibiting different binding modes for same interactions (for example in a different orientation) known as 'pharmacophore doublets'. However, fragment screening is not straightforward, as fragments typically exhibit low binding affinity. Consequently, these fragments must be screened at elevated concentrations using highly sensitive biophysical techniques like NMR, X-ray crystallography, isothermal titration calorimetry, and protein thermal shift assay.¹⁰² They are often difficult to resolve in electron density maps obtained from X-ray crystallography, and can also be easily confused with ordered solvent molecules.¹⁰³

Once structures have been solved, the binding site that is (hopefully) saturated with fragments can be used as a starting point for a SBDD campaign. These screens provide information for initial strategies such as merging fragments that have adjacent binding modes or defining a pharmacophore map of the binding site.

A specific example of this approach is Fraglites,¹⁰³ a library of 31 fragments specifically containing combinations of donor-acceptor and acceptor-acceptor doublets connected by 1-5 atoms. To remedy the aforementioned problems of resolution, the fragments were halogenated with heavier (and therefore easier to resolve) Bromine or Iodine atoms. Fraglites have successfully been used to map cyclin-dependent kinase 2 (CDK2) and has identified orthosteric and allosteric sites.¹⁰⁴

2.4.2 Fragment Growing

Fragment growing involves building up a ligand starting from an initial low potency (μM - mM) scaffold positioned in a binding site. This can be done computationally, for example, with the *de novo* fragment growing software LigBuilder.¹⁰⁵ In LigBuilder, growing is usually performed by iterative hydrogen replacement until the ligand has reached a desired size, a maximum number atoms has been added, or there is no space for further growth. Growth is implemented via a genetic algorithm that uses the best scoring 10 % of the population as a parent for further grown, as scored by an empirical scoring function, all performed in a pre-defined binding pocket. LigBuilder also incorporates retrosynthetic analysis to suggest synthetically accessible ligands, ensuring that the structure can be created from readily available commercial precursors.

There are drawbacks to the fragment growing approach, however. If the active site is relatively large and the goal is for a ligand to bind to two sub-pockets that are far apart, it can be difficult to grow a ligand between them since there is poor interaction in the intermediate space.¹⁰⁶ A limitation of fragment growing is that it assumes a static protein structure, which is not always accurate. This approach overlooks cryptic pockets—binding sites that only become accessible upon ligand binding—thereby limiting the identification

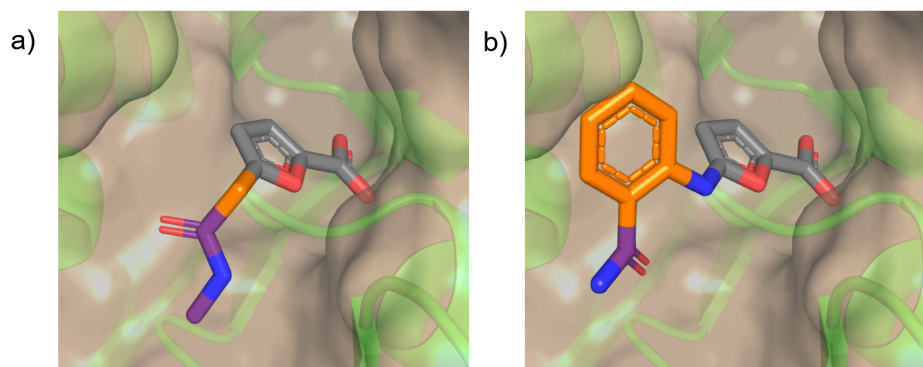


Figure 4: An example of two R groups added to a furan ring core with FEgrow. a) carbon linker, N-Methylacetamide R group; b) phenyl linker, formamide R group. R groups shown in purple, with linkers shown in orange. There are in excess of 1 million possible linker/R group combinations, which can be added to replace either a methyl group or a hydrogen.

of potentially crucial interaction sites.

There are relatively few open-source fragment growing tools, but one example is OpenGrowth¹⁰⁷ (noting that LigBuilder is free but closed-source). OpenGrowth again utilises an iterative growth method to produce hundreds of thousands of potential structures, and then applies search criteria to sort them. Fragment growth is achieved by successively replacing hydrogens, creating a bond between the new moiety and the atom that the removed hydrogen was connected to (adjusting bond length for hybridisation of the two atoms involved in the new bond). The choice of fragments to join is based on their frequency of occurrence in a chemical database (ChEMBL), so the generated fragments will be of similar character to the initial database structures. A "3-mer" screen is also implemented, allowing exclusion of specific combinations of three fragments. To incorporate dynamic information, the ligand can be grown in multiple protein conformations simultaneously. The ligand can be scored with respect to them all using its empirical scoring function, which is derived from the observed frequencies of atom-atom contacts from a curated dataset of protein-ligand complexes, a notable drawback of this empirical scoring function is that it fails to properly account for quantum mechanical effects.¹⁰⁸

Another example of iterative fragment growing is FragPELE,¹⁰⁹ which uses stochastic (Monte Carlo) methods in a parallel fashion for improved sampling, as well as employing a flexible binding site — reducing steric clashes — to generate *de novo* molecules. Monte Carlo (MC) steps have three stages: perturbation (e.g. ligand rotation), re-positioning of side chains after perturbation, and final relaxation of the structure. Once a fragment is grown, a longer MC simulation with an emphasis on side-chain sampling is used to score the ligand. This method achieved good agreement with respect to crystallographic

data and the approach attains similar results to FEP+ (Schrödinger’s proprietary, free energy perturbation software). This approach is computationally expensive however, and FragPELE takes on the order of 1h per fragment on 48 Intel Xeon Platinum 8160 processors.

Recently,¹¹⁰ machine learning was implemented to produce designed multi target ligands (DML) which inhibit two similar hydrolases, the inhibition of which has been associated with anti-inflammatory effects.¹¹¹ Both structure-based and ligand-based approaches were implemented, with the structure-based approach utilising the MOE (Molecular Operating Environment) docking software suite. From a starting fragment, ligands were generated with a random forest approach and encoded using various molecular fingerprint tools.^{112 113} This approach generated 116 ligands, of which 11 leads were cherry-picked using criteria such as synthetic availability or solubility. Although novel inhibitors were successfully generated, human expertise was still required to whittle down the potential candidates and this approach mandated a large number of known active ligands for two similar active sites.

2.4.3 Fragment Linking

Fragments can also be optimised by linking together two (or more) hits in the active site. Fragment binding when linked is typically greater than the sum of their parts, termed ‘superadditivity’, and is often seen with the combination of polar and non-polar fragment interactions.¹¹⁴ This is rationalised by expecting polar fragments to suffer a large desolvation penalty, which is offset by the complementary hydrophobic fragment.¹¹⁵ A tempting starting point for linking is to combine fragments together with alkyl chains and, indeed, this has produced high-affinity ligands.¹¹⁶ Functional groups should be strategically added to the linker though, in order to exploit interactions with the protein and increase affinity. Databases have also been heavily utilised in order to find appropriate linkers either by combinatorial slicing of bonds¹¹⁷ or *de novo* generation via genetic algorithms.¹¹⁸ Database methods have the distinct limitation of not incorporating any initial 3D information in the search, which affects the ability of the method to maintain binding poses, since they have strong spatial constraints.

Fragments can also be linked using virtual coupling. AutoCouple¹¹⁹ is an *in silico* coupling method that can generate libraries of molecules that can be synthesised, ideally, in one step while filtering for undesired functional groups. Virtual compounds are scored using the open-source rDock¹²⁰ and poses obtained were minimised using CHARMM¹²¹ and the CHARMM36/CGenFF force field¹²². Using this diversity oriented approach, novel nM potent and cell-permeable inhibitors were generated.

Machine learning approaches to fragment linking have been used to generate SMILES, or graph representations of linkers, allowing a greater set of novel and diverse ligands

versus using database methods. 3D information can be incorporated in the generation of linkers (for example with DeLinker) where the relative positions and orientations of fragments are utilised to minimise RMSD (root mean squared displacement) of the generated compounds relative to the starting structure.¹²³ This graph-based deep generative model generates novel linkers that outperform database methods with respect to retaining 3D similarity, especially with linkers that have at least 5 atoms and DeLinker can be applied scaffold hopping as well as fragment linking design tasks. After training, the model has implicit knowledge of the 3D structural information of the initial fragments, but no structural information of the binding site is incorporated in the generation of ligands. Building on DeLinker, 3D pharmacophore information was incorporated using DEVELOP (DEep Vision-Enhanced Lead OPTimisation) to try and address this deficiency.¹²⁴

There are also 2D machine learning approaches. The open-source utility SyntaLinker approaches the task of fragment linking as a natural language processing problem.¹²⁵ This approach divides the ChEMBL database into terminal fragments and linkers, and then trains deep conditional transformer neural networks to learn syntactic patterns in the SMILES strings that link fragments together. This method can generate a large number of novel linkers, as well as implement length and pharmacophore constraints. This 2D approach even surpassed DeLinker in terms of top RMSD values and linker novelty (SyntaLinker achieved a top RMSD of 1 Å vs 2.5 Å for Delinker, and produced 20% more novel compounds without losing recovery).

Given the importance of attaining the correct binding mode for *de novo* drug discovery, for example via docking, it is prudent to be able to validate a given predicted pose. Many similar tools exist for simple geometrical and chemical checks,^{126–128} PoseBusters combines both types of validity checks and is designed as a test suite crafted to detect chemically inconsistent and physically implausible ligand configurations produced by protein-ligand docking and molecular generation (e.g. diffusion models¹²⁹) techniques, as well as benchmark them. Currently, no deep learning-based docking method exceeds the performance of standard docking methods when both the physical plausibility and binding mode RMSD are considered. When assessing docking tools, it is crucial to consider physical plausibility. It is possible to attain excellent RMSD scores, yet propose ligand poses that are physically unreasonable. Although metrics like root mean square displacement can yield seemingly laudable scores, they have the potential to produce unrealistic and inaccurate conformations. Due to the inadequacy of straightforward metrics such as RMSD, alternative methods involving energetic or steric calculations should be considered. PoseBusters takes into account not only stereochemistry and both intra- and intermolecular geometry but also protein-ligand clashes.

2.5 Ligand Based Drug Design (LBDD)

In the absence of available structural (3D) information, the design of *de novo* ligands must be done by using the chemical and physical properties of the ligands themselves, combined with assay data to evaluate their effectiveness. This approach is referred to as Ligand Based Drug Design (LBDD). An indispensable tool for LBDD is Quantitative Structure Activity Relationship (QSAR), defined as using chemical data analysis to map empirical relationships between the structure of a ligand (shape, moieties, number of rotatable bonds, etc) and IC_{50}/K_i . This is done by testing various functional groups, substructures and/or linkers to glean information about the relative importance of their contributions to affinity in a rational manner.¹³⁰

One specific example of a QSAR methodology is that of Free Wilson analysis¹³¹, one of the first attempts to mathematically formalise the impact that certain substructures have, for a series of analogues. This R group (an R group is the typical way to abbreviate the structures of molecules, and here is used to mean an arbitrary functional group that is added to a molecule) decomposition method has the benefit of being simple and interpretable due to the clear relationship between the presence/absence of groups and the activity of the compound. Though Free-Wilson analysis is limited to predictions that are within the scope of the initial set of molecules used.¹³² *Hansch et al.* devised a similar approach which incorporated physiochemical properties such as the octanol-water partition (which is a method of estimating lipophilicity via solubilities of non-polar solutes in solvent and itself is not without problems, specifically the neglecting of hydrogen bond donor effects) coefficient and other parameters to an equation to be fit to data creating a predictive model. The aforementioned parameters must be available (or be amenable to calculation) and are often simple quantities that are merely proxies for underlying effects and do not reflect the complexities of what is actually contributing to binding - so causality can be tricky to establish due to the possibility of confounding parameters correlating with another unknown but causative variable.¹³³

One challenge to QSAR lead development is the frequently non-linear relationship between a ligand's structure and its activity, the problem in general is a navigation of high-dimensional activity landscapes composed of molecular representations and activity data.¹³⁴ It is commonly presumed that similar ligands (for some definition of similar) will have similar activity, however this is usually too optimistic. Even a single and relatively minor change, such as an inverted stereocentre, can have unpredictable effects on affinity.¹³⁵ While in simple and narrow regions of chemical space, the activity landscape might resemble smooth rolling hills, where gradual changes in structure result in similarly proportional effects to activity - this is not true in general, and large discontinuities can occur, termed 'activity cliffs'.¹³⁶ These cliffs can be caused by a difference of binding mode or the addition of large groups that no longer fit, or even

disrupting water networks.⁵² An especially egregious occurrence to a QSAR model is when an activity cliff is caused by a feature that is not captured by the molecular representation being used, so molecules that are dissimilar and ought to be far away in chemical space appear in adjacent regions, baffling any attempt at SAR.¹³⁷ Cliffs can be beneficial in early stages of a campaign though,¹³⁸ providing a mechanism to generate potent lead compounds from superficial hits, but are something to be avoided in the latter stages - where more careful multi-parameter optimisation of physicochemical properties, whilst conserving important structural features, is needed.

LBDD campaigns have been characterised as a self avoiding walk through chemical space, where the objective is to traverse novel areas that might afford potent binders, whilst avoiding retracing through regions that have already explored (as defined by a chemical similarity score e.g. molecular fingerprints).¹³⁹ Ligand based approaches are information sparse, and generally less computationally expensive but can suffer from hazards arising from choices made in how to represent molecular data. Ultimately, the task of molecular design is highly non-linear and dependant on factors that cannot be easily modelled, even with structural information.

2.5.1 Representations of Molecules

There are various standard formats for chemical data; this standardisation enables a broader ecosystem of CADD tools to function due to the interoperability that this standardisation provides. However, decisions must be made regarding how chemical structures are represented, depending on the stage of the drug discovery campaign and the intended use of the information.

Firstly, there are simple string (1D) representations like Simplified Molecular Input Line Entry System (SMILES), which encode molecules as chains of atoms and bond types with predefined rules. This format can represent organic molecules, including isotopes, radical species and charges (extensions exist that are used as the basis for ML/generative approaches for *de novo design* with enhanced properties such as chemical robustness),¹⁴⁰ but lack the ability to store 3D information, and so are incapable of modelling protein-ligand complexes or conformational properties. SMILES (or images of 2D structures) are used as the typical human-readable format, with more complete 3D representations (like the SDF format¹⁴¹) that contain full structural and topological information being used for docking or molecular dynamics approaches (various representations of molecules are shown in Figure 5). Less memory-intensive formats, suitable for use in similarity searches of large databases and larger-scale virtual screening, include the most common 2D extended connectivity fingerprints (ECFPs).¹⁴²

These fingerprints reduce the 2D structural complexity of a molecule to a single bit vector (an array of 1s and 0s). Representing molecules as vectors in this way is amenable

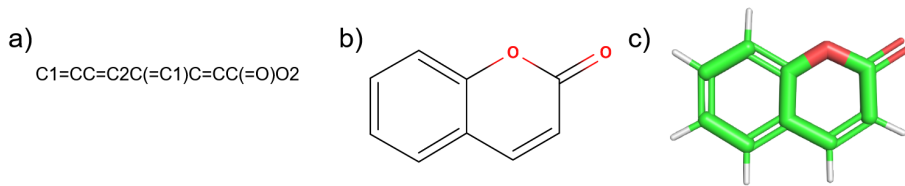


Figure 5: chromen-2-one. a) 1D SMILES representation, b) 2D skeletal structure, c) 3D conformer with explicit hydrogens

to neural network methods due to the high-dimensionality of the data, and deep learning has been shown to be effective in automatically identifying and prioritising salient features for SAR.¹⁴³ In addition to purely structural information, molecular descriptors can also be calculated,⁶³ such as partial charges, total polar surface area, counts of specific types of functional groups, molecular weight, $\log P$, and anything else easily quantifiable about the molecule which can capture relevant physical properties. These quantities can be utilised in conjunction with activity data to perform statistical analyses that can tease out important structural features. Identifying trends in activities is easy enough for simple structural changes and datasets containing a handful of molecules, but in the age of high-throughput screening and large chemical datasets, interpreting compound data (that are not-necessarily from a homologous series) can be challenging.¹⁴⁴ A large change in activity is useless for building a QSAR model if it involves too big of a structural change as it obfuscates the causality between structure and activity. One way round this problem is by identifying pairs of molecules in large datasets via Structure-Activity Landscape Index (SALI) analysis, where any change in affinity is normalised by the similarity.¹⁴⁵

2.5.2 Similarity

Comparing similarities between chemical structures can be done in various ways, but is commonly done using the Tanimoto coefficient between fingerprints; a measure of how many bits share in common between them, with vectors that are exactly the same scoring 1, and any given vector scoring 0 with its opposite (or XOR inverse). The Tanimoto similarity coefficient $T(A, B)$ between two binary fingerprints A and B is given by:

$$T(A, B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

where A and B are bitvectors and \cap is the intersection between them (all elements that appear in both vectors).

Extracting features from a 2D graph and reducing them to a single bitvector comes with an expected set of problems, most notably that of similar molecules not possessing similar fingerprints. This is caused by artefacts in encoding (how the molecule is converted

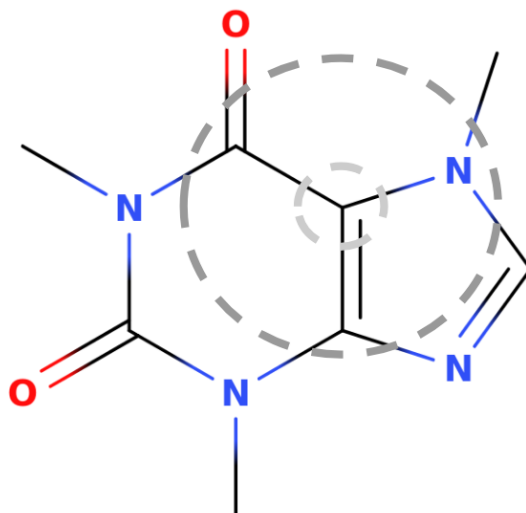


Figure 6: In creating an ECFP fingerprint, each non-hydrogen atom is assigned an identifier, and those identifiers are iteratively combined with identifiers of neighbouring atoms until a specified distance (diameter) is reached. Progressive iterations capture circular neighbourhoods of increasing size around each atom, which are then encoded into integer values via a hashing algorithm, which are collected into a list. A schematic of two iterations of this procedure is shown, with distances of 0 and 1 in light and dark grey, respectively.

to a binary representation) or by the representation itself.

If practical can be circumvented by using more sophisticated methods to capture similarity, such as 3D shape-based methods which can perform better than 2D fingerprint approaches in circumstances when molecules exhibit complex geometries.¹⁴⁶ One such shape-based method is Rapid Overlay of Chemical Structures (ROCS), which aims to quantify the similarity between the volumes of two given molecules by calculating the overlap between atom-centered gaussians.¹⁴⁷ Previous issues notwithstanding, there is no generally agreed upon threshold value for two vectors to be 'similar' anyway, and so it remains an inherently subjective endeavour (typical values for similarity thresholds are generally > 0.6).⁸² Alternative 2D similarity methods, such as substructure similarity approaches, are especially useful for performing matched molecular pair (MMP) analysis. This technique focuses on finding substructures (subgraphs) that differ by a single group. Assuming that the binding mode is identical, any disparity in affinity can then be attributed to the structural changes between the molecules being analysed.

2.6 Experimental Assay Techniques

2.6.1 Fluorescence Polarization (FP) Assay

Fluorescence polarisation (FP) is an assay technique that measures changes in the polarisation of light emitted by a small, fluorescently labelled molecule when it binds to

a receptor. The degree of polarisation of the emitted light is inversely proportional to the rotational mobility of the molecule.¹⁴⁸ When a fluorescently labelled ligand binds to a larger protein (typically >10 kDa), the rotational motion of the fluorophore is constrained, as it must now move with the entire complex. Consequently, the emitted light retains a significant degree of polarisation.¹⁴⁹

Fluorescent labelling is achieved by covalently attaching a fluorophore to a small molecule, often a peptide. The quality of the fluorophore, including its quantum yield, extinction coefficient, and stability, is critical as these parameters determine the dynamic range of the assay.

The modulation of rotation upon complexation enables the measurement of interactions between small molecules and their target proteins.

$$\frac{1}{FP} - \frac{1}{3} = \left(\frac{1}{FP_0} - \frac{1}{3} \right) \left(1 + \frac{RT}{\eta V} \tau \right) \quad (6)$$

Where FP is the observed polarisation, R is the universal gas constant, T is the absolute temperature, η is the solution viscosity, FP_0 is the intrinsic polarisation (polarisation value in the absence of molecular rotation), V is the molar volume and τ is the lifetime of the excited state of the fluorescence. The rotational correlation time θ for a molecule FP is proportional to θ , given by

$$\theta = \frac{\eta V}{RT} \quad (7)$$

To measure FP , a fluorescent sample is excited by polarised light and emission intensities are collected from channels that are parallel and perpendicular to the electric vector of the excitation light

$$FP = \frac{I_{\parallel} - I_{\perp}}{I_{\parallel} + I_{\perp}} \quad (8)$$

where I_{\parallel} and I_{\perp} are the parallel and perpendicular emission intensities, respectively.

FP assays have become standard for high-throughput screens and drug discovery in general and since it is carried out in solution, is nonradioactive and can be performed at low volumes. It is also apt for high throughput screening, and has enabled assays that can elucidate protein-peptide, protein-protein receptor-ligand, and protein-nucleic acid interactions at scale.¹⁵⁰

2.6.2 Surface Plasmon Resonance (SPR)

Surface plasmon resonance (SPR) is a quantitative and label-free assay technique that employs an optical biosensor to measure molecular affinity and binding kinetics. It offers significant advantages over techniques involving fluorescent labelling, most notably the labelling process itself, which can alter a ligand's binding properties.¹⁵¹ Moreover,

SPR provides direct measurements of binding constants and affinities without requiring specialised reagents, such as fluorophores.

The core principle of SPR involves shining coherent light onto a surface comprising a glass slide, a thin gold layer, and immobilised receptors (typically proteins). Under conditions of total internal reflection — due to the disparity in refractive indices between the glass and the analyte solution — incident light excite surface plasmons (oscillations of free electrons, or electron density) on the metal layer, resulting in a measurable reduction in reflected light intensity at a specific angle, known as the resonance angle, $\theta_{\text{resonance}}$.¹⁵²

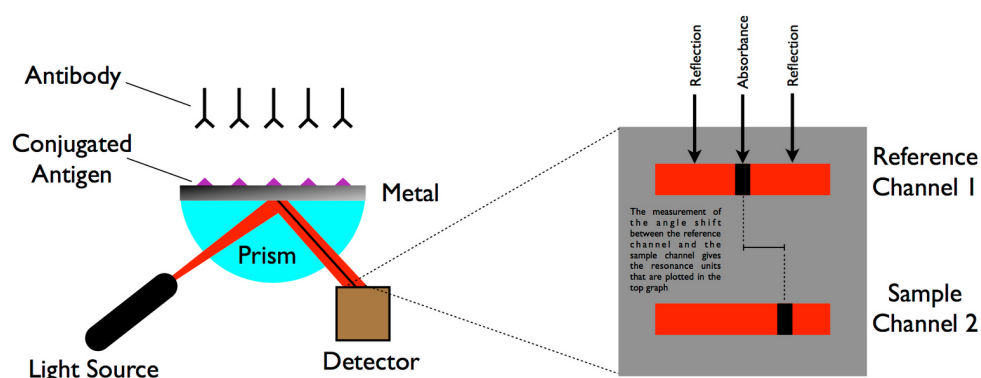


Figure 7: A monochromatic laser is directed at a thin metal surface, exciting surface plasmons (oscillations of electron density) at a specific resonance angle. The reflected light is monitored by a detector, which identifies a distinct intensity dip corresponding to the resonance condition. As analytes, such as antigens, bind to immobilized receptors (e.g., antibodies) on the surface, the added mass alters the local refractive index. This change shifts the resonance angle, which is quantitatively recorded by the detector. The magnitude of this shift is directly proportional to the analyte concentration, allowing measurement of affinities for e.g. binding events.

When analytes flow across the receptor surface at a known concentration and rate, their binding alters the refractive index near the surface. This change shifts the resonance angle and the degree of this shift correlates quantitatively with analyte binding. The incident photons transfer energy to the thin metal layer, generating a standing charge density wave at the conductor-medium boundary. This wave, termed a surface plasmon wave, is sensitive to changes such as ligand-receptor interactions.

SPR experiments typically begin with a baseline measurement under buffer flow. Analytes are then introduced, allowing binding to reach equilibrium, followed by dissociation where buffer flow is restored. Repeating the process at varying analyte concentrations enables calculation of the binding constant, K_D , between the receptor and analyte.¹⁵³

SPR can provide rapid assessment of compound binding, making it invaluable for in vitro studies. It can determine metrics such as residence time, and can discern complex

binding kinetics, including two-step binding mechanisms - which may be challenging for less sensitive techniques. Although SPR cannot distinguish between agonists and antagonists, its ability to detect binding events with high sensitivity makes it a popular tool in FBDD, where smaller compounds with lower affinities are screened at high throughput.¹⁵⁴

2.7 Summary

An overview of the theory and considerations for drug design has been given here, along with the application of computational techniques. The rest of the thesis focuses on these computational approaches in more detail, and their application to drug discovery projects. These projects are primarily focused on SBDD, using X-ray crystal structures, physics-based simulation and machine learning methods (Chapter 4 & 5), in combination with database methods (Chapter 6) for *de novo* design.

3 Theory

3.1 Computational Chemistry

Chemistry is the science of molecules, their potential energy surfaces as a function of their nuclear coordinates, and the behaviour of collections of such molecules, including their macroscopic properties. The application of physical and mathematical models to these molecules via computers is the field of computational chemistry. Computational chemistry enables simulated, *in silico* investigations of a ‘third’ kind (alongside theoretical and experimental approaches). The distinction between experiment and computation is not always clear however, especially when it comes to complex physical measurements, which themselves either heavily utilise computation or are essentially, computers themselves (e.g. nuclear magnetic resonance (NMR) spectrometers). This ambiguity arises because the computational algorithms embedded within these instruments directly influence the acquisition and interpretation of experimental data, blurring the line between computation and experiment. The development of new computational science offers a different but complementary form of investigation where models, theories, and experiments can be tested simultaneously, provided a particular model can be expressed algorithmically. Typically, computational science is focused on numerical and non-exact solutions, as many systems in computational chemistry reduce to that of the generic many body problem (similar problems exist in other fields, such as astronomy), and an inordinate amount of time and energy has gone into addressing this one issue, in its many guises.^{155 156} In general, no closed form solution exists for e.g. predicting the dynamics of more than two interacting particles, and iterative numerical methods are required. These can be performed to an arbitrary accuracy in general and so can essentially be regarded as ‘exact’ in this context.¹⁵⁷ Using these approaches, computational science offers access to time and length scales that are intractable to theory and inaccessible to experiment, which makes them particularly prevalent in biomolecular simulation.

In this chapter, basic quantum mechanical concepts are covered which, although not directly employed in this thesis, are the basis upon which approximate methods such as molecular mechanics/machine learning (MM/ML) are predicated. The principles outlined in this section are presented to provide broader context on the underlying methods used in this thesis, and are referenced in subsequent sections throughout.

The theoretical aspects of molecular dynamics (MD), integral to the FEgrow workflow (a method of evaluating *de novo* compound design *in silico* developed by myself, Ben Cree, and Dr Mateusz Bieniek and employed throughout this work), are then outlined, starting with the fundamental equations that underpin the equations of motion for biomolecular simulation. Applications of MD are presented, including methods for the estimation of free energy of binding such as free energy perturbation (FEP) calculations, which are

employed throughout this thesis and are the ultimate aim for the output of FEgrow.

A brief outline of docking procedures and techniques (used extensively in Chapter 5 in completion of CACHE (Critical Assessment of Computational Hit finding Experiments) Challenge 2) is also given.

An overview of Uniform Manifold Approximation (UMAP) is also provided in this chapter, focusing on its application in molecular visualisation, specifically its usage in reducing the dimensionality of molecular fingerprint data, which allows for a more intuitive visualisation of similarity relationships between molecules.

Finally, the principles of active learning (AL) are summarised with a focus on Gradient Boosted Machine (GBM) and Gaussian Process (GP).

These concepts, theories, and methodologies form the foundation of the computational approaches employed throughout this thesis. Methods mentioned here are revisited in the context of computational workflow design and computational experiments in subsequent chapters.

3.2 Quantum Mechanics (QM)

Writing down the equations that govern quantum mechanical systems (systems of physical objects small enough that their properties are quantised) has been possible for over a hundred years. The Schrödinger equation, which relates a system's Hamiltonian, \hat{H} , an operator representing all physical observables, such as kinetic or potential energy to its Energy, E (eigenvalues).

$$\hat{H}\psi = E\psi \quad (9)$$

where ψ is the wavefunction (eigenfunction), contains all information that it is possible to know about a system

$$\int_{-\infty}^{\infty} |\psi(x)|^2 dx = 1 \quad (10)$$

and \hat{H} is the Hamiltonian operator corresponding to a system of M nuclei and N electrons with no external electric or magnetic fields, shown below with all physical constants equal to unity (atomic units).

$$H = -\frac{1}{2} \sum_{i=1}^N \nabla_i^2 - \frac{1}{2} \sum_{A=1}^M \frac{1}{M_A} \nabla_A^2 - \sum_{i=1}^N \sum_{A=1}^M \frac{Z_A}{r_{iA}} + \sum_{i=1}^N \sum_{j>i}^N \frac{1}{r_{ij}} + \sum_{A=1}^M \sum_{B>A}^M \frac{Z_A Z_B}{R_{AB}} \quad (11)$$

where A, B and i, j represent the nuclei and electrons respectively, M_A and Z_A represent the nuclear mass and charge, ∇^2 is the Laplacian operator, defined in Cartesian coordinates as $\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}$. R_{ij} and r_{ij} represent the distance between two particles and can be written in the form $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$

Quantum mechanics (QM) is able to quantitatively predict many properties of interest, such as electronic structures, reaction transition pathways, and material energy-band structures.¹⁵⁸ Due to problems relating to the lack of an analytical solution, and scaling laws for the memory required (in a real quantum system, each additional electron doubles the possible number of quantum states, doubling the required storage space), various formalisms for approximating quantum systems have been devised. One such approximation is the independence of nuclear and electronic coordinates (and therefore potential energy surfaces), justified physically by the disparity in the masses and time-scales of the motion of electrons and nuclei. This allows the wavefunction to depend parametrically on stationary nuclear coordinates, reducing the number of interactions that need evaluating, and is referred to as the Born-Oppenheimer approximation.

Another routine assumption is considering the mean field of electronic charge of a system as opposed to the effect of each electron individually. This assumption underpins the Hartree-Fock method of solving the electronic Schrödinger equation, where each electron is described by its own single-electron wavefunction, and the total wavefunction is represented as a Slater determinant — a mathematical construct that provides an antisymmetrised linear combination of these wavefunctions, engineered to satisfy the Pauli principle.¹⁵⁹ Modelling the total wavefunction as a product of simple wavefunctions in this way is less computationally expensive, but comes at the cost of naturally neglecting the (negative) electron correlation energy, by treating the electrons in a mean-field way.¹⁶⁰

In the Hartree-Fock method, the total energy of the system depends on the orbitals of all electrons, requiring the equations to be solved iteratively. This iterative procedure leverages the variational principle, which asserts that any approximate wavefunction will have an energy greater than or equal to that of the exact wavefunction. The goal is to find the ‘best’ trial function that minimises the energy and is achieved by constructing matrices of integrals and molecular orbitals. The Fock operator — a one-electron operator that describes the kinetic energy of an electron, its attraction to the nuclei, and its repulsion from other electrons — is then used to determine the energy of the system. Electron densities are updated iteratively until convergence to some specified tolerance is reached.

$$\hat{F}\psi_i = \epsilon_i\psi_i \quad (12)$$

where \hat{F} is the Fock operator, ψ_i are the molecular orbitals, and ϵ_i are the orbital energies.

$$\mathbf{FC} = \mathbf{SC}\epsilon \quad (13)$$

where \mathbf{F} is the Fock matrix, \mathbf{C} is the matrix of molecular orbital coefficients, \mathbf{S} is the overlap matrix, and ϵ is the diagonal matrix of orbital energies.

There are two broad categories for calculations, the first is those that are performed

without being augmented by any experiment, *ab initio* (Latin: ‘from the beginning’) techniques, and the second, semi-empirical methods (referring to the hybrid nature of the calculation), those that have some integrals replaced with empirical parameters fit to experimental data (such as free energies of solvation) for reduced computational complexity.

Other more accurate methods of calculating QM energies exist, one being Density Functional Theory (DFT). In DFT the fundamental object is not a wavefunction, but a density. DFT is formally exact (but empirical in practice) and based on the theorem by Hohenberg and Kohn that states that the electronic energy of a system is uniquely determined completely by the electron density ρ (up to an additive constant).¹⁶¹ The method by which density is related to energy is via density functionals, which convert a density function, $\rho(\mathbf{r})$, to a positive real number. Functionals come in a veritable zoo of forms, but share the desired property that the number of variables - and therefore computational complexity - is independent of the system size. The multitude of extant functionals with slightly different characteristics and performances has made it difficult to accurately adjudicate which is actually best for a particular purpose, and often leads to the continued use of outdated and superseded methods.¹⁶² Another advantage of DFT is that it has a superior ability to capture many-body electron correlation due to it being modelled implicitly in the electron density. These methods are then used as the foundation for simpler and faster methods that are used in force field parameterisation, for example.¹⁶³

As always, there is an intrinsic trade-off between accuracy and computation, with DFT methods such as the B3LYP/6-31G* functional/basis set, giving errors in the 5 - 10 kcal/mol range for heats of formation, large compared to gold standard coupled cluster methods, accurate to <0.5 kcal/mol, but on the other hand they typically require $\sim 10^2$ more wall time to compute.¹⁶⁴ Other properties that are more relevant to biomolecular simulation and ligand binding (thermally accessible conformational states and energies of small drug-like organic compounds under physiological conditions) have fortunately been shown to converge faster with levels of theory, and have adequate accuracy at the DFT level.¹⁶⁵ Another fundamental approximation which is present in the vast majority of *ab initio* methods is that of the basis set, a set of mathematical functions that can be combined linearly to represent a more complicated object, such as a molecular orbitals being composed of simpler atomic orbitals. These atomic basis sets can be expressed as simple Gaussians (for computational simplicity) shown below in polar form

$$\chi_{\zeta,n,l,m}(r, \theta, \phi) = N Y_{l,m}(\theta, \phi) r^{2n-2-l} e^{-\zeta r^2} \quad (14)$$

where N is a normalization constant and $Y_{l,m}$ are spherical harmonic functions, l represents orbital angular momentum quantum number and ζ is a constant related to

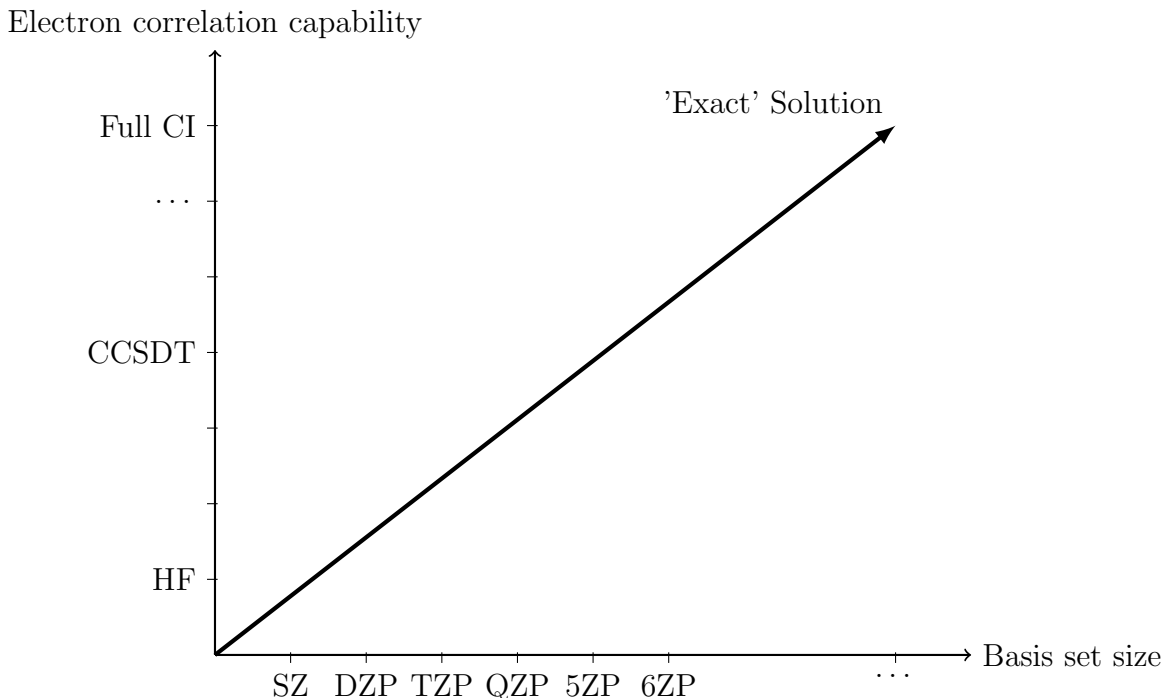


Figure 8: The two main dimensions of accuracy/complexity in modelling quantum systems. Electron correlation capability is the ability for the level of theory to capture the coupling of electronic motion, by increasing the sophistication of the configuration interaction (CI). Methods such as coupled cluster single-double-triple (CCSDT) include configurations that contain single, double and triple excitations configurations. Basis sets are a set of mathematical functions which can be combined in a linear way to represent the total electronic wavefunction. The basis sets shown here are Slater-Type Orbitals (STO) basis function (equation 11), where a SZ basis set has only a single s-function (for second row elements it has two s functions and one p function), DZ has double the number of functions, TZ has triple, and so on. When both the full CI and basis set limit are reached, the exact solution to the Schrödinger Equation, up to the Born-Oppenheimer approximation, is reached.

the effective charge of the nucleus. θ and ϕ represent the polar and azimuthal angle respectively, with r being the radius. The number of basis functions composing the MO is described in terms of zeta, with double zeta (DZ) being composed of double that of a single zeta (SZ) MO, for example.

These quantum calculations are the bedrock of various areas of computational chemistry, such as parameterising force fields for biomolecular simulations and providing reference data for machine learning potentials. They are especially relevant for the FEGrow workflow detailed later in this thesis in the conformer minimisation step, which utilises a machine learning force field to produce accurate conformers for *de novo* compounds grown in the context of a protein receptor. Force fields and conformer generation, along with other concepts, are elaborated on in the following section.

3.3 Molecular Dynamics (MD)

Macroscopic observables of chemical systems are difficult to calculate and were traditionally modelled by approximate methods (such as virial expansion, modelling the departure of other macroscopic quantities from ideal behaviour) that could only be applied to simple systems, such as gases. For non-trivial systems, determination of macroscopic properties like diffusivity, protein conformational dynamics and crystal growth is done by simulation using a physical model.¹⁶⁶ Computational models that accomplish this are a confluence of theory and experiment, and are used to make testable predictions *in silico* for condensed matter systems, by simulating microscopic details, e.g. molecular geometry, defined by some potential energy surface. One way this can be (classically) done is via atomistic Molecular Dynamics (MD) simulations, using Newtonian mechanics to propagate a system of interest. The essence of this technique is numerically solving a differential equation that involves calculating forces, calculating accelerations, velocities and positions iteratively until a specified number of discrete cycles has occurred. These cycles are called time-steps due to each iteration representing a discontinuous step-wise evolution in time equal to δt , which represents the passage of time, typically on the order of femtoseconds. There are many open-source MD engines that exist for this purpose, such as OpenMM¹⁶⁷, GROMACS¹⁶⁸ and ASE¹⁶⁹, which with GPU support can accomplish hundreds of nanoseconds a day for biophysical simulations of solvated protein-ligand complexes, that contain over a million atoms.¹⁷⁰ The scale for routine (e.g. on a single workstation) biophysical simulations is limited due to the time complexity required to run classical MD simulations grows as $\mathcal{O}(N^2)$ where N is the number of atoms. A use of MD in the drug discovery process is real-time modelling of receptor flexibility. An MD trajectory can be used to explore the receptor conformational space, and representative snapshots can be extracted which can be used for docking ligands to an ensemble of structures, which can be more representative of the receptor's structure *in vivo* if the most relevant conformation is not known a priori.¹⁷¹

Initial geometries, \mathbf{r}_i , and velocities, \mathbf{v}_i , as well as typologies (bond connectivity, atom charges) of the system to be simulated must be supplied, as well as defining the functional form of interactions to compute the potential energy V , such as bond stretching or twisting (see section 3.3.3). These initial velocities can be generated randomly according to a Boltzmann distribution

$$p(v_i) = \sqrt{\frac{m_i}{2\pi kT}} \exp\left(-\frac{m_i v_i^2}{2kT}\right) \quad (15)$$

where k is Boltzmann's constant, T is temperature, m_i is the mass of particle i and v_i is the velocity of particle i .

3.3.1 Integrators

If a system contains more than two interacting particles without constraining any degrees of freedom, then there is no closed-form solution for their dynamics. Therefore, approximate numerical methods are needed to solve the differential equations which govern the system's dynamics, and this is achieved with discrete algorithms called integrators. Forces on interacting particles can be calculated in discrete time-steps based on interatomic potentials, with force being the negative derivative of the potential with respect to position. For each atom i

$$\mathbf{f}_i(t) = m_i \mathbf{a}_i(t) = -\frac{\partial V(\mathbf{x}_i(t))}{\partial \mathbf{x}_i(t)} \quad (16)$$

where f_i , m_i and a_i are the force, mass and acceleration of the i th atom. $\mathbf{x}(t)$ is a vector representing an atom's position at time t in Cartesian space and has a corresponding potential $V(\mathbf{x})$ which is calculated via a Force Field (detailed in section 3.3.3).

These forces determine accelerations, and therefore position of the particles in the system by integrating Newton's equation of motion for each subsequent time-step $t + \delta t$.

This integration can be accomplished with various algorithms¹⁷². Shown here is the Velocity Verlet algorithm, a second order virial expansion that acts over a single time-step. It is a commonly chosen algorithm due to its reversible, simple, and numerically stable nature.¹⁷³ The velocities and positions of particles are updated according to the following equations.

$$\mathbf{x}_i(t)(t + \delta t) = \mathbf{x}_i(t)(t) + \mathbf{v}_i(t)(t)\delta t + \frac{1}{2}\mathbf{a}_i(t)(t)\delta t^2 \quad (17)$$

Subsequent atomic positions are contingent on the current accelerations, velocities and positions of the system - and the subsequent velocities are updated according to

$$\mathbf{v}_i(t)(t + \delta t) = \mathbf{v}_i(t) + \frac{1}{2}[\mathbf{a}_i(t) + \mathbf{a}_i(t)(t + \delta t)]\delta t \quad (18)$$

where $\mathbf{x}_i(t)(t)$, $\mathbf{v}_i(t)(t)$ and $\mathbf{a}_i(t)(t)$ represent the velocities of particle i at time t with $t + \delta t$ representing the time advanced by one time step, δt . This procedure can now be repeated until t is equal to some predetermined value equal to the desired length of the simulation. Time steps are an important parameter; if the discrete steps of simulation are too large then fast dynamics will not be properly captured (typically Hydrogens) as C-H bond vibrations occur on the time-scale of ~ 10 fs (or 3000 cm^{-1}) and are typically constrained with various algorithms (LINCS, SHAKE) to enable timesteps (~ 2 fs) similar to the frequency of vibration.^{174 175}

Due to the timescales accessible to MD (up to μs), slower processes such as the folding

of some proteins, or ligand binding events are unable to be accessed and require stochastic techniques such as Monte Carlo (MC) that are useful in estimating kinetically slow processes by probabilistically generating configurations, which can be used to estimate thermodynamic or structural properties.^{176,177}

To further enhance the realism of simulations and control key thermodynamic properties, we can modify the dynamics of a simulation using thermostats and barostats.

3.3.2 Thermostats

Simulated systems should be properly equilibrated — that is, the time and ensemble averages should be equivalent. A simulation should also not be influenced by its starting configuration, which can be prevented by relaxing the system using different conditions, e.g. assuring that the volume of the simulation fluctuates about a constant value, or checking the root mean square displacement (RMSD) between key structures or molecules. Various ensembles can be established by controlling macroscopic parameters and leaving other parameters as degrees of freedom. The simplest ensemble is the constant energy or microcanonical ensemble and the natural ensemble of simulation, the micro-canonical or NVE ensemble, fixes moles, N , volume, V and energy, E , and is achieved without any specific controls on either temperature or pressure. If restriction on temperature is required then the canonical ensemble (NVT, keeping the temperature, T , constant) can be utilised. The temperature of a simulation in general can be calculated using the equipartition theorem, which states that each degree of freedom equally shares all energy at equilibrium.¹⁷⁸ The temperature of the system is proportional to the average kinetic energy and given by

$$\frac{3}{2}Nk_B T = \left\langle \sum_{i=1}^N \frac{1}{2} m_i \mathbf{v}_i^2 \right\rangle \quad (19)$$

However, the (instantaneous) temperature at any given time-step of the simulation will not always be equal to the target temperature. This is expected, but these fluctuations should be centred about the target temperature. To maintain the correct temperature, the Newtonian equations of motion can be augmented and a naïve way of achieving the desired restraint on temperature is by scaling velocities, which is a popular choice for initial equilibrations. This rescaling reduces the temperature artificially but creates a non-canonical ensemble and introduces errors as well as artefacts into simulations, so is avoided in practice.¹⁷⁹ To avoid the problem of discontinuous changes in velocities and the consequent difficulties they bring to simulations, velocities can instead be rescaled smoothly using a relaxation term

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau} \quad (20)$$

Where τ is the time constant of relaxation, T_0 is the given desired temperature and T is the instantaneous temperature of the system. This is the Berendsen thermostat and it does not sample the isobaric (NPT) ensemble due to its exponential decay of deviations from the desired temperature.¹⁸⁰ This thermostat is known to cause unphysical behaviour in simulations, most notably the flying ice-cube effect, where energy is consistently redistributed to low-frequency modes (translation, rotation), violating the equipartition theorem.¹⁸¹ For this reason it is usually applied to initial equilibration and non-production runs.

An alternative thermostat that does sample velocities from the canonical ensemble is the Andersen thermostat, which rescales velocities stochastically.¹⁷⁸ There are also methods that involve coupling the system to a fictitious ‘heat bath’ that can absorb (or dispense) thermal energy which models the exchange of a system’s energy with its surroundings, and maintains Maxwell-Boltzmann statistics, reproducing the canonical ensemble. A drawback of this approach is the expense of structural accuracy due to the rescaling, in a similar fashion to that of the Berendsen thermostat.¹⁸²

Thermostats can ensure correct ensembles and are used to generate equilibrated and production ready configurations of systems with the use of protocols involving ensembles of different types in sequence. Often, simulations are run at specific temperatures or pressures to interrogate particular properties under certain conditions, such as diffusivity or phase change behaviour, as defined by the ensemble. Although they are unphysical, they often have little to no practical effect on the dynamics of the system.¹⁸³ There are completely analogous techniques to correct pressure, i.e. barostats, which is done by acting on positions and volume instead of velocities.

3.3.3 Force Fields (FF)

In order to run a physics-based simulation of a chemical system, one must be able to take a set of atoms (and bond topologies) and map them to a potential energy surface. In biophysical MD, this is typically done with a classical force field (FF). An FF is a collection of simple functional forms and parameters, which are fit to lower level QM or experimental data and applied to molecules using constituent atomic chemical environments directly or indirectly^{184 185}. This process is known as parameterisation.

Historically, indirect parameterisation schemes based on a byzantine library of atom ‘types’ have been used, where each atom’s chemical environment (including tuples of atoms for torsions, dihedrals, etc.) is organised into a discrete set of predefined categories. These atom types are assigned manually, requiring a significant time investment for their (sometimes arbitrary) application, based on chemical intuition of human experts. As a result of this indirect application, atoms in identical chemical environments may even be assigned different atom types leading to an unnecessary proliferation of FF

parameters. Such complexity has notably lead to human errors that have persisted in FFs for decades^{186,187}.

The immutability of FFs, even after the advent of vastly superior hardware is partly due to the heroic amount of effort and man-hours required to create them. However, a new generation of classical force fields are in development¹⁸⁸, spearheaded by the Open Force Field (OpenFF) Initiative. They are being developed using open-source software and data, creating an infrastructure whereby anyone can take an existing generic FF and tune it to their needs, incorporating innovations such as fast charge assignment and functional forms that are more physically relevant.¹⁸⁴ A major improvement in these FFs is the utilisation of direct chemical perception, which employs a more straightforward approach to parameterisation by directly operating on the chemical graph without relying on atom types.¹⁸⁵ This is achieved using SMARTS/SMIRKS patterns, which allow for a more streamlined and chemically intuitive system of FF definitions. Direct parameterisation reduces the number of required definitions and ensures that parameters consistently align with chemical environments. As a result, OpenFF force fields are significantly more concise, comprising hundreds of lines parameters rather than thousands, as seen in widely-used FFs like GAFF2.¹⁸⁹ This advancement leads to easier maintenance and greater accuracy across a broad range of chemical systems.¹⁹⁰ These classical FFs are routinely used in drug discovery, materials science, and polymer modelling due to their low computational overhead.¹⁹¹

The two fundamental types of interactions included in an FF are bonded valence terms (e.g., force constant of bonds, rotational barriers) and non-bonded interactions (e.g., electrostatic, dispersion) which are treated classically using approximate functional forms. Although non-bonded interactions are only summed pairwise, they implicitly account for many-body interactions due to being based on condensed phase data (which contribute appreciably to the behaviour of such phases) and so are referred to as effective pair potentials.¹⁹²

Bonded terms in molecules are the dominant contribution to the potential energy surface, often an order of magnitude greater than non-bonded interactions.¹⁹³ These are parameterised as various potential energy functions that are inexpensive to evaluate, with

force constants fit to QM data.

$$\begin{aligned}
 V = & \sum_{\text{bonds}} \frac{1}{2} k_r (r_{ij} - r_0)^2 \\
 & + \sum_{\text{angles}} \frac{1}{2} k_\theta (\theta_{ijk} - \theta_0)^2 \\
 & + \sum_{\text{torsions}} \sum_n k_{\phi,n} [\cos(n\phi_{ijkl} + \delta_n) + 1] \\
 & + \sum_{\text{non-bonded pairs}} \left[\frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right].
 \end{aligned} \tag{21}$$

Where A, B are parameters describing non-bonded interactions, r_0 is the equilibrium bond length, θ and ϕ are bond angles and q is partial charge. The parameters k_r , k_θ , $k_{\phi,n}$, δ_n , q_i , q_j , A_{ij} , B_{ij} depend on the specific atom type combination being calculated. Although popular, the limitations of classical FFs are numerous. Topologies are unable to be changed once the simulation has started, which prevents the modelling of chemical reactions; electron densities remain static, even though interactions and chemical environments are known to affect partial charges.¹⁹⁴

The functional forms used in FFs are not always physical, with the most prominent example being the use of the 6-12 Lennard-Jones (LJ) potential in modelling dispersion forces, where the repulsive 12 term is simply the square of the attractive portion of the potential. This is a fundamentally quantum effect but the LJ potential has been employed due to a historical lack of computing power, which has promulgated many of these simple parameterised representations (in the classical limit) of more complicated quantum phenomena. Consequently, there has been development of new potentials with more physically sound functional forms to replace the ageing LJ potential and, in newer versions of OpenFF, the LJ potentials were refit by tuning the parameters to better match simulated densities and enthalpies of mixing.¹⁹⁵

In addition to the aforementioned parameters, partial charges are required to calculate the electrostatic potential in MD and are typically assigned with semi-empirical methods. The most popular approach is AM1-BCC,¹⁹⁶ which is computed at the Hartree-Fock level with the 6-31G* basis set. It is a parameterised bond order correction to the AM1 population charge model, itself a parameterisation of the NDDO formalism via experimental reference data.¹⁹⁷ The runtime complexity scales as $\mathcal{O}(N^2)$, where N is the number of atoms and so it cannot be applied to larger molecules like proteins. Moreover, as a QM calculation is required to generate charges, it can take minutes to be applied to small ligands. Recently, graph neural network approaches have been used to accurately assign AM1-BCC charges hundreds of times faster and with linear scaling.¹⁹⁸

Water in biomolecular simulations is integral to accurate predictions and is intimately

related to the force field used. Water is typically modelled using TIP3P,¹⁹⁹ where the geometry of the molecule is based on experimental gas phase data (with the 3 referring to the number of point charges) which are placed on the nuclei and lone pairs omitted for ease of computation. Parameters such as partial charges and LJ parameters were fit to experimental densities and radial distribution functions to create a reasonably accurate water model that remains the most popular to this day.

Although these FFs are ubiquitous, correctly simulating proteins with correct secondary and tertiary structures is not currently done with generic force fields. Instead bespoke protein force fields, such as the AMBER based ff99SB which is fit to (tetra)peptide gas phase QM data, allowing ϕ and ψ torsion parameters to be fit in a more realistic environment that accounts for the conformational influence of neighbouring residues.²⁰⁰ More recent force fields, such as ff19SB have utilised solution-phase QM data to better account for effects like polarisation and are widely used in biophysical simulation.²⁰¹

3.3.4 Conformer Generation

As molecules are three-dimensional objects, being able to (quickly) generate reasonable conformations of molecules (defined as those accessible under physiological conditions) is integral to modelling a ligand's binding affinity and structure based drug design (SBDD) in general.²⁰² Conformations can be generated through various methods, such as Molecular Dynamics (MD) simulations, which yield Boltzmann-weighted conformational ensembles in solution. However, MD simulations are computationally expensive and impractical for large-scale applications like virtual screening, where conformers are needed for millions of molecules.²⁰³ One solution to large-scale conformer generation is the experimental-torsion distance geometry with additional basic knowledge (ETKDG) algorithm.² This is a stochastic (as opposed to systematic) search of conformational space, preferred due to the combinatorial explosion of rotermeric states with molecule size, which makes systematic approaches intractable. Experimentally determined structures (like the Cambridge Structural Database (CSD) and the Protein Data Bank (PDB)) are used to create these libraries.²⁰⁴ Histograms (Fig. 9) were generated for each SMARTS pattern in the range 0° to 360° with a 10° step size. Two functional forms were fitted to these data

$$V(\phi) = K [1 + \cos(d) \cos(m\phi)] \quad (22)$$

and

$$V(\phi) = \sum_{j=1}^6 K_j [1 + \cos(d_j) \cos(j\phi)] \quad (23)$$

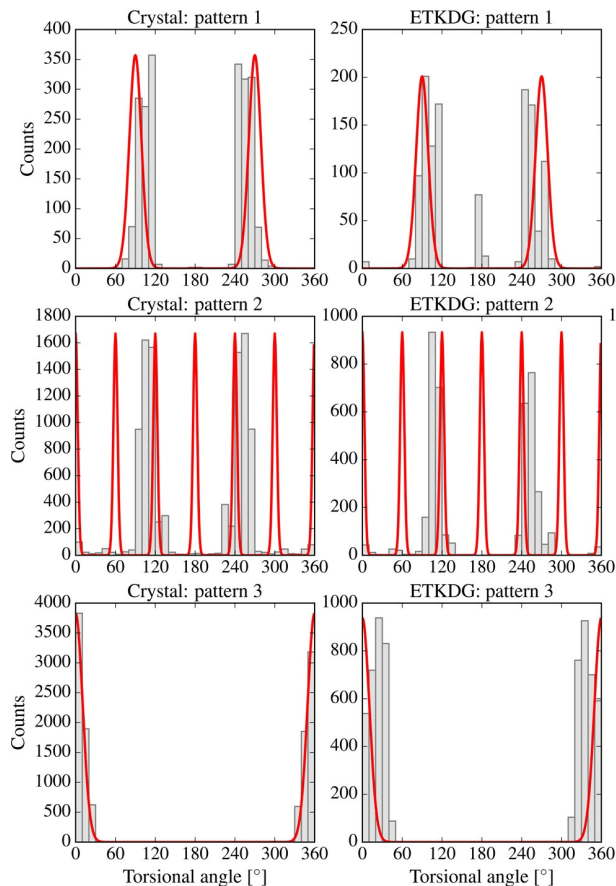


Figure 9: Torsional-angle distributions from the CSD, ETKDG (only acyclic torsion patterns together with the fitted torsional-angle potentials for the first three SMARTS patterns that were fitted.²

where K is the force constant, d is the phase shift and m is the multiplicity. These potentials were further refined manually. The first step in the ETKDG algorithm is the generation of random distance matrices which satisfy basic geometric constraints, bounded by the maximum and minimum atom pair distances (DG), subsequently ‘experimental-torsional’ (ET) terms derived from the CSD and PDB (consisting of approximately 2000 molecules) are utilised in an iterative fashion for coordinate refinement of these initial guess conformations, which is applied to specific SMARTS patterns - for a specified number of iterations. Finally, additional ‘basic knowledge’ (K) terms are used to minimise the embedded conformers using a universal force field (UFF), mostly using parameters as set in ETDG²⁰⁵ with exceptions for torsional angle potentials for bonds in aromatic rings and sp carbons.²⁰⁶ Using ETKDG, 84% of a dataset of 1290 small-molecule crystal structures from the CSD were reproduced with an RMSD of 1.0 Å or better, and 38% within an RMSD of 0.5 Å.² In a review of free conformer generation methods, it was found that the ETKDG algorithm found to more accurately recapitulate experimental structures of small molecules compared to other freely available toolkits, attributed to the stochastic nature of chemical space exploration, but also that it tends to generate

more similar conformers compared to other methods.²⁰⁷

3.3.5 Machine Learning Potentials and ML/MM (Machine Learning/Molecular Mechanics)

The current popularity of Machine Learning (ML) is owed to a paper illustrating the supremacy of artificial neural networks for image recognition relative to the state of the art — principally because of the scale of training data and computing power that had become available — which has subsequently been cited more than 30,000 times.²⁰⁸ This type of image recognition is an example of deep learning; statistical frameworks composed of layers of (non-linear) operations that are often referred to as ‘neural networks’ due to the architecture being modelled after neurons in the human brain.²⁰⁹ Artificial neurons have a simple structure. They are linear mathematical functions that receive input(s), perform operations on those inputs, and then pass the output through some non-linear activation function. The inputs could be anything (e.g. pixel brightness/colour, in the case of image recognition). When utilised in concert with groups (layers) of other neurons whose inputs and outputs are linked, an emergent potency that is suited to complex tasks of prediction and classification previously considered beyond the scope of computers is realised. A non-linear operation (performed on its output) is required to access this power since any linear combination of operations can be represented as a single operation, meaning the omission of non-linearity would essentially make all neural networks single layered.²¹⁰ These models can provide robust and versatile models for broad and complicated tasks, with applications to natural language processing, preference prediction, and applications to basically every field of physical science.^{211–213} Creating networks of simple layers allows a function of arbitrary complexity to be modelled, such as identifying objects in images. The malleability of the network means that these categories do not have to be predefined and can be implicitly learned through data, a useful property for modelling phenomena that possess complicated behaviour from a large set of interrelated parameters.²¹⁴ Models are trained by adjusting the values of certain parameters, known as weights, through a feedback loop that compares predictions to the training data and iteratively updates these weights. This process is facilitated by backpropagation, which employs gradient descent to minimize the error associated with each neuron’s output. An error function, such as the root mean square error between predicted values and ground truth, is defined for each output. During training, this error function is minimised iteratively, with the weights being updated in each step, referred to as an ‘iteration.’ The entire training dataset may be processed multiple times, which is described as training in ‘epochs’ and can be repeated

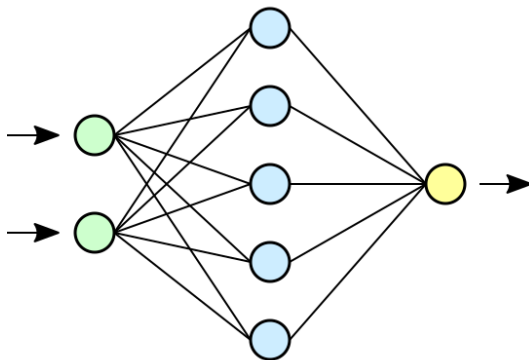


Figure 10: A single-layered neural network comprised of two input neurons (green), five hidden neurons (blue) and one output neuron (yellow). The weights and biases (eq 24) can be adjusted to minimise the error in the output of the network. For example predicting the dissociation energy of a diatomic molecule, where the input neurons represent the constituent atoms, and the output neuron represents a scalar energy value.

until an arbitrary accuracy is reached.

$$w_{ij}^{(t+1)} = w_{ij}^{(t)} - \eta \frac{\partial E}{\partial w_{ij}} \quad (24)$$

where w_{ij} is the weight connecting neuron i to neuron j , t represents the current iteration, η is the learning rate controlling the step size of updates, E is the error function (e.g., mean squared error), and $\frac{\partial E}{\partial w_{ij}}$ denotes the gradient of the error with respect to the weight.

The most powerful modern large-language models (LLMs), for example, have billions of parameters which they use to extract and store meaning from a huge corpus of text.²¹⁵ The rise of ML in scientific research is largely due to its ability to reduce complex tasks to matrix manipulations, which has been enabled by the development of neural networks and the advent of extremely powerful GPUs, designed to handle such operations in parallel. Modern hardware is capable of achieving approximately 10^{14} floating-point operations (FLOPs) per second, facilitating the efficient processing required for data intensive ML applications. ML methods have been routinely applied in the field of computational chemistry for more than a decade, with applications ranging from constructing DFT functionals to drug design through QSAR prediction.^{143,216,217} The most striking recent development is in protein structure prediction, where AlphaFold, a model created by Deepmind, has led to a revolution in the ability to predict protein structures. Generating hundreds of millions of structures (of varying accuracy), providing a platform for insight into the large and previously ‘dark’ sections of the human proteome.²¹⁸

Paul Dirac surmised the following in 1929: “The underlying physical laws [...] of a large part of physics and the whole of chemistry are thus completely known. [...] It therefore becomes desirable that approximate practical methods of applying quantum

mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation.” After almost a hundred years, his aspiration is starting to be realised thanks to the generation of large quantum mechanical (DFT) and molecular simulation datasets for training of ML models. The scalability of their (automatic) generation and their generalisable nature have set the stage for the development of ML techniques, which have the potential to revolutionise the way we approach science.²¹⁹

Specifically, this is a revolution in the field of Neural Network Potentials (NNPs) for MD simulation, an increasingly attractive alternative to classical force fields. These potentials were first conceived in the 1990s,²²⁰ and escape the limited and handcrafted functional forms and bond topologies of classical force fields. They also avoid the need for predefined atom types, which are artificial constructs with well-known limitations.¹⁸⁴

NNPs are parameterised on lower level DFT data (for instance, MACE-OFF23 (an organic NNP for small organic molecules) was trained to reproduce the energies and forces computed at the ω B97M-D3(BJ)/def2-TZVPPD level of quantum mechanics), aiming to map atomic environments (atom identity, local geometry) to a scalar energy value. Mappings are done via complicated functional forms that are learned implicitly by neural networks. These can be much more expressive compared to the restricted forms of classical force fields and are able to more closely match reference energies, allowing much more accurate simulations compared to classical force fields, but at a fraction of the cost of QM calculations.³ State-of-the-art models have been shown to reach 0.1 kcal/mol accuracy on restricted chemical spaces, and below 1 kcal/mol on more diverse datasets.²²¹ In order to achieve these results, high quality datasets must be available and, ultimately, a model is unable to extract useful information if it is not present in the training data, or there is not enough data to learn from in the first place.²²² Although these NNPs are much faster than semi-empirical methods and more accurate than classical force fields, they remain orders of magnitude more computationally expensive (and hence slower) when used for biomolecular simulations.¹⁸⁴ Additionally, there can be problems with numerical stability due to the complex functional forms NNPs possess, which may produce high-energy conformations not represented in the training data, leading to simulation failure.²²³

As mentioned, in order for a NNP to be of practical use, the training data must not only cover relevant chemical space, but also contain a diverse set of conformations,²²⁴ as well as providing forces, which are key in creating accurate models.²²⁵ Once such dataset is the QM9 dataset, which contains over 134,000 small molecules rich in chemical diversity — however, it has been noted that it does not contain conformational diversity since only low energy conformers are present.¹⁸⁴ Datasets like these that are large, easily extensible and versioned will be the engine for future development of NNPs.

NNPs, in lieu of standard parameterisation, embed atomic environments using local

features that can be extracted from simulation trajectories by inferring topologies directly from spatial coordinates. There are numerous schemes to achieve this featurisation (see below), but the key considerations are: that they capture the local environment up to a cut-off distance; are invariant under arbitrary transformations that do not affect the underlying physics (e.g. rotations and translations (equivariance)), and are efficient to calculate.²²⁶ These qualities are essential in creating a robust physical model of practical utility.

If only a specific region of a simulation such as a binding pocket in a protein-ligand complex is of interest, applying NNPs to the entire simulation box is unnecessary, as using more sophisticated methods than required in regions that do not significantly impact the target process is redundant. In this scenario, hybrid MM/ML systems can be created (which can be done in OpenMM²²⁷, for example, and as simply as using a classical force field) whereby a small portion of the simulation is modelled using an NNP which is then embedded into the wider simulation. The ML force field (NNP) describes the intramolecular energetics V_{ML} of the ligand \mathbf{r}_L and the MM force field is responsible for the interaction between the ligand and the environment E .

$$V_{MM/ML} = V_{MM}(\mathbf{r}_E, \mathbf{r}_L) + V_{ML}(\mathbf{r}_L) - V_{MM}(\mathbf{r}_L) \quad (25)$$

According to this scheme, MM-ML coupling is not explicitly included, with a consequent disadvantage of the non-bonded interactions being limited in accuracy to that of the classical force field.²²⁸

There are many implementations of NNPs, with applications to different domains and varying levels of accuracy³ (including in FEgrow (Chapter 6)). A specific NNP that is used extensively in this work is ANAKIN-ME (Accurate Neural network engine for Molecular Energies) or ANI. This general purpose NNP is used for ligand conformer energy minimisation in the context of a protein receptor (see section 4.2.3).²²⁹ It has been shown to achieve DFT level accuracy with the same scaling laws as classical force fields ($\mathcal{O}(N^2)$). ANI has achieved RMSEs (Root Mean Square Errors) of ~ 1 kcal/mol on torsion benchmarks, demonstrating its capacity to accurately model small molecules, in addition to being able to run atomistic MD simulations.²²⁹ More relevant to the work here, it has been shown to be effective in ranking conformers of drug molecules more accurately (relative to MP2/cc-PVTZ²³⁰), according to optimised geometry and relative energies than both MMFF94²³¹ and PM6.²³² ANI’s accuracy, in consideration with its speed (tens of milliseconds to perform single-point calculations on a conformer) demonstrates the capacity NNPs have for utility in large virtual screening tasks, at least within a ML/MM set-up.

In ANI, the total energy of a molecule, E_T , is calculated via E_i , which represents an

atomic number specific NNP using

$$E_T = \sum_{\text{all atoms } i} E_i \quad (26)$$

In this manner E_T represents a sum over all atomic contributions to the total energy. In addition to transferability, the above sum provides the benefit of nearly linear scaling in computational complexity. Modified Behler and Parrinello symmetry functions²³³ (BPSF) are used to create an atomic environment vector. Each element of \tilde{G}_i^X captures specific regions of an atom’s angular and radial chemical environment.

$$\tilde{G}_i^X = \{G_1, G_2, G_3, \dots, G_M\} \quad (27)$$

The vectors \tilde{G}_i^X are fundamental to applying this functional form of the total energy. For an atom i , \tilde{G}_i^X provides a numerical representation of i ’s total local chemical environment i.e. incorporating both radial and angular features. The approximation of the local atomic environment is achieved using a piecewise cutoff function

$$f_C(\mathbf{r}_{ij}) = \begin{cases} 0.5 \left(\cos \left(\frac{\pi \mathbf{r}_{ij}}{r_C} \right) + 0.5 \right) & \text{for } r_{ij} \leq r_C \\ 0.0 & \text{for } r_{ij} > r_C \end{cases} \quad (28)$$

Here, r_{ij} is the distance between atoms i and j , and R_C is the cutoff radius, with $f_C(R_{ij})$ being a continuous function with continuous first derivatives. To probe the local radial environment for an atom i , the following radial symmetry function, introduced by Behler and Parrinello, produces radial elements G_R^m of \tilde{G}_i^X :

$$G_R^m = \sum_{\text{all atoms } j} s_i e^{-h(r_{ij}-r_s)^2} f_C(\mathbf{r}_{ij}) \quad (29)$$

The index m ranges over a set of h and \mathbf{r}_s parameters. The parameter h adjusts the width of the Gaussian distribution, while \mathbf{r}_s shifts the centre of the peak. In ANI, various modifications were made to the original functions representing the angular features to better capture chemical environment, such as the addition of a parameter that enables the angular features to be evaluated within radial shells based on an average for the distance from neighbouring atoms, designed to facilitate extremely fine grained interrogation of regions of the radial environment, which aids with transferability. \tilde{G}_i^X is computed for each atom in the molecule, which are then input into their respective NNP to evaluate each atom’s E_i^X , which are subsequently summed to give E_T .

An alternative atomic embedding scheme is Atomic Cluster Expansion (ACE), where

an atomic property p , as a function of the local environment of atom i is expanded as

$$\phi_i(p) = \sum_{\mathbf{v}} c_{\mathbf{v}}(p) \mathbf{B}_{i\mathbf{v}} \quad (30)$$

with expansion coefficients $c_{\mathbf{v}}(p)$ and $n\mathbf{v}$ basis functions $\mathbf{B}_{i\mathbf{v}}$ (which depend on atomic positions and are ordered hierarchically, they are again translationally and rotationally invariant and based on Clebsch-Gordan coefficients) with multi-indices $i\mathbf{v}$. The energy of atom i can then be evaluated using all properties as follows

$$E_i = \mathcal{F}(\phi_i(1), \phi_i(2), \dots, \phi_i(P)) \quad (31)$$

Where \mathcal{F} is a nonlinear function. The Multi-ACE (MACE) architecture was trained on the SPICE dataset, a quantum chemistry dataset containing 1.1 million conformations for a diverse set of small molecules, dimers, dipeptides, and solvated amino acids, includes 15 different elements and both charged as well as uncharged molecules,²³⁴ calculated at the ω B97M-D3(BJ)/def2-TZVPPD level of theory. The MACE-OFF model maps chemical elements and positions to a system’s potential energy, similar to ANI and like all force fields, but does so by communicating an atom’s local environment using a graph. This graph is constructed by connecting nodes (atoms) and if they are in each other’s local environment, $\mathcal{N}(i)$ (set of all atoms j around the central atom i for which $\|\mathbf{r}_{ij}\| \leq \mathbf{r}_{\text{cut}}$, where \mathbf{r}_{ij} is the vector from atom i to atom j and \mathbf{r}_{cut} is the cutoff hyperparameter). Each node has an array of associated features denoted $\mathbf{h}_i^{(t)}$ which is represented in the spherical harmonic basis

$$h_{i,j}^{(0)} = \sum_z W_{kz} \delta_{zz_i} \quad (32)$$

The superscript (0) in this case indicates the initial (0-th) layer. For each atom, the features of its neighbours are combined with interatomic displacement vectors, \mathbf{r}_{ij} , to form a one particle basis, and the radial distance is used as an input into a learnable radial function $R(\mathbf{r}_{ij})$.

This basis can then be made to be equivariant and combined (essentially including many-body interactions, pooling over n neighbours incorporates n bodies) with a message-passing neural network. The ‘messages’ sent on this network are then a learnable linear combination of the equivariant many-body features. Two message-passing layers are used in the model, and the atomic site energy is simply the sum of various functions applied to node features from the first and second layers, with forces being calculated by taking the analytical derivatives of the total potential energy with respect to the positions of the atoms. MACE has been shown to produce NNPs that are capable of sub-chemical accuracy for the hydration free energies of organic molecules, as well as showing promise

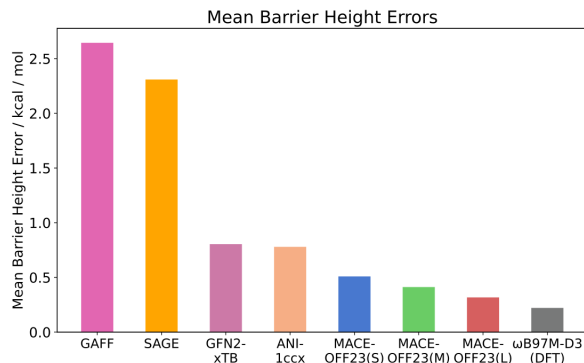


Figure 11: The competitive accuracy of a ML forcefield, MACE-OFF³ relative to DFT at the CCSD(T) level of theory. Barrier height was calculated for the torsion angle between two aromatic rings in the biaryl torsion benchmark,⁴ which contains 78 molecules.

for condensed phase alchemical free energy simulations.²³⁵ After significant progress in recent years, it is no longer the case that energies and forces are the limiting factor in the utility of NNPs.

Speed and stability have usurped these old considerations for improvement, and the design of the next generation of NNPs should focus on addressing the balance between speed and accuracy, to fully take advantage of their power. This means there is room in the liminal space between the simple, fast and inaccurate classical force fields and the more complex, but slow NNPs that is yet to be exploited.²²¹

The issue of out of domain applicability of ML models is a general one, and not limited to computational chemistry and NNPs.²³⁶ Recent studies however have shown early signs of these kind of models beginning to generalise the underlying physics they were trained on, with *Duignan et al.* showing that molten ion models have the ability to form correct crystal lattice structures (even with correct unit cell parameters to within 10 %).²¹⁹ As training datasets become larger and richer, and the models will become more sophisticated, and the frequency of these emergent phenomena will increase, precipitating a wealth of real-world applications to currently insoluble problems.

3.3.6 Free Energy Calculations

Gibb’s free energy, G , is the natural thermodynamic quantity under normal experimental conditions and dictates the spontaneity of (molecular) processes, such as those in chemical reactions or in protein folding. This quantity can be computed for transfer processes such as conformational change, solvent transfer/partition coefficient, membrane insertion and relative binding free energy of one molecule to another.²³⁷ In the context of drug discovery, one such quantity that is of interest is the change in free energy change, ΔG , of protein-ligand complexation which can be used to aid the rational design of compounds.³⁶ This free energy change can be used to determine the effect that structural changes have on

the potency of a molecule. These changes can be set up as a transformation network, with nodes being molecules and each edge a transformation. The ΔG between molecules can then be used to inform and rank compounds that are to be tested experimentally, e.g. via assay.

The binding free energy, $\Delta G_{\text{bind},L}$, for a ligand L, binding to a receptor R, can be approximated by the Hemholtz free energy of binding $\Delta A_{\text{bind},L}$, given by:

$$\Delta A_{\text{bind},L} \approx -k_B T \left(\ln \left(\frac{P(RL)}{P(R)P(L)} \right) + \ln(c^\circ N_{\text{Av}} V) \right) \quad (33)$$

where $P(RL)$ and $P(R/L)$ is the probability of finding the complex in the bound and unbound states respectively, k_B is Boltzmann’s constant, T is temperature. The last term involving N_{Av} , Avagadro’s constant and c° , the reference concentration corrects for the simulated concentration being different than the standard concentration. Assuming $\Delta V \approx 0$ for the process,²³⁷ then

$$\Delta G_{\text{bind},L} \approx \Delta A_{\text{bind},L} \quad (34)$$

These probabilities are given by the Boltzmann probability density function for a given configuration, \mathbf{q}

$$P(\mathbf{q}) = \frac{\exp(-\beta V(\mathbf{q}))}{\int_{\Gamma} \exp(-\beta V(\mathbf{q})) d\mathbf{q}} \quad (35)$$

where β is thermodynamic temperature, defined as $\frac{1}{k_B T}$ and $V(\mathbf{q})$ is the potential energy of configuration \mathbf{q} and Γ is simulation box volume. A thermodynamic cycle of a binding event can be formed, which must sum to 0 (Fig. 12).

$$\Delta G_{\text{b},A}^\circ + \Delta G_{\text{b},B}^\circ - \Delta G^{\text{w}} - \Delta G^{\text{p}} = 0 \quad (36)$$

$\Delta G_{\text{b},A}^\circ$ and $\Delta G_{\text{b},B}^\circ$ represent physical binding events, and ΔG^{w} and ΔG^{p} representing alchemical transformations from molecule A to molecule B in the unbound and bound states, respectively. The free energy change associated with the virtual downwards transformations in 12 are relatively easy to calculate, so utilising eq. 36, $\Delta\Delta G$, the difference in ΔG can be calculated

$$\Delta\Delta G_{\text{bind}} = \Delta G^{\text{p}} - \Delta G^{\text{w}} \quad (37)$$

Calculating entropic quantities in absolute terms can be challenging, but the calculation of differences in these quantities is a more efficient task due to the cancellation of systematic errors caused by deficits in the FF used (errors of equal magnitude might occur in identical terms for each simulation but with opposite signs), or insufficient sampling of configuration

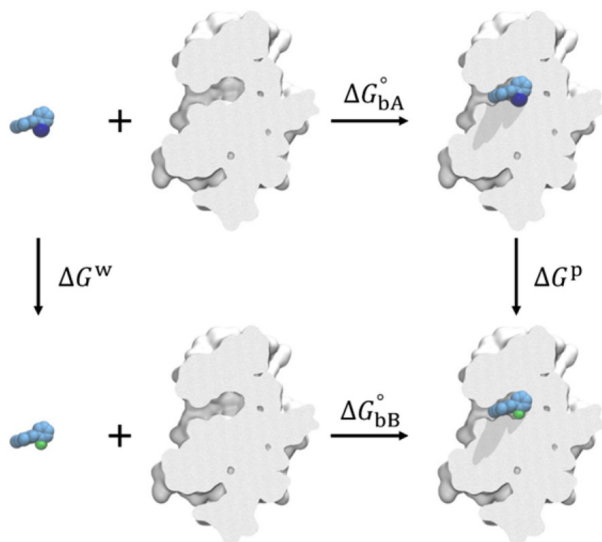


Figure 12: Thermodynamic cycle used to calculate relative binding free energies ($\Delta\Delta G_{bind} = \Delta G_{bA}^\circ - \Delta G_{bB}^\circ$) between congeneric ligands. The horizontal legs correspond to the physical binding process, whereas vertical legs indicate the unphysical transformation of ligand A (blue) into ligand B (green) performed in bulk solvent (left) and in the protein binding site (right).

space.²³⁷ The difference $\Delta\Delta G_{bind,AB}$ (eq. 37) can be used to quantitatively estimate the priority of a congeneric series of ligands, or of different conformations.

Determining the relative values of the free energy change of binding for a series of (similar) molecules requires construction of the topology network for alchemical transformations. The specific structure of this network is an important consideration: for a library of n potential compounds there are on the order of n^2 possible transformations, most of which are not viable (for example, if the structures are too dissimilar) — and those that are may be not be desirable. It is therefore prudent to be able to create sensible transformation maps of a minimum size that take into account not only the desired design goals of a project, but also useful graph features such as the creation of cycles (that should have a ΔG value of 0, the difference from 0 is termed closure error) that can be used to test convergence and various tools exist for this purpose.²³⁸

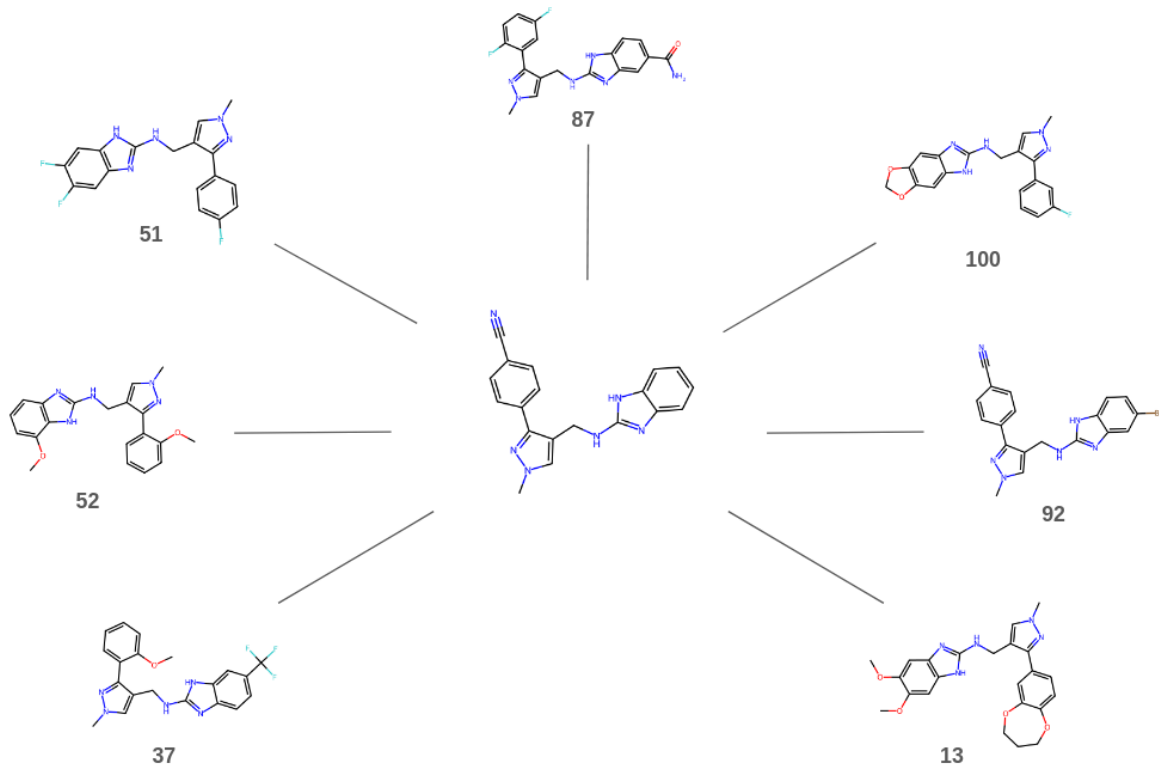


Figure 13: An example network of alchemical transformations used for calculation of relative binding free energies of 7 inhibitors for a helicase of SARS-CoV2 (section 5.4.1).

There are two general methods for determining ΔG : by perturbation or integration methods. The first is Free Energy Perturbation (FEP), whereby the intermolecular potential of one system is perturbatively changed to another to provide a potential energy surface (PES) bridge between two physical states, so that rigorous free energy changes can be calculated using the Zwanzig equation

$$\langle \Delta A_{AB} \rangle = kT \ln \langle e^{(E_B - E_A)/kT} \rangle_A \quad (38)$$

where $\langle \rangle_A$ denotes an ensemble average. Because the energy difference must be small compared with $k_B T$, and the efficiency of free energy estimators is heavily dependent on phase-space overlap,²³⁹ This overlap can be quantified using an overlap matrix used in reweighting estimators, discussed below. Since the timescale of the kinetics of binding events often exceed that of current MD simulations²⁴⁰, methods have been devised to estimate the free energy change of binding without observing multiple binding events. The process of transforming from A to B generally involves breaking the transformation into several intermediate steps or simulations. Since free energy is a state function and therefore path independent,²⁴¹ these small intermediate simulations (called λ windows) can be alchemical in nature. That is to say, ligands have their structure modified in

unphysical ways, in-place, to calculate free energy change. Different proportions of the initial and final states potential can be ‘turned on’, which is characterised by a λ parameter (0 meaning entirely A, 1 meaning entirely B), with the total free energy change determined by summing the changes in each step (simulation).

$$E_\lambda = \lambda E_A + (1 - \lambda) E_B \quad (39)$$

The second class of methods, Thermodynamic Integration (TI), uses a similar scheme starting from the partition function given by

$$Q = \sum_{i=0}^N e^{-E_i/k_B T} \quad (40)$$

where N represents the number of particles and E_i represents a single energy state. The partition function allows calculation of all macroscopic functions in statistical mechanics, giving the free energy of state λ , $A(\lambda)$.

$$A(\lambda) = -kT \ln Q(\lambda) \quad (41)$$

Differentiating this expression with respect to λ yields

$$\frac{\partial A}{\partial \lambda} = -\frac{kT}{Q} \frac{\partial Q}{\partial \lambda} = \left\langle \frac{\partial V}{\partial \lambda} \right\rangle \quad (42)$$

replacing the right-hand side by an ensemble average and integrating over λ gives

$$A(1) - A(0) = \int_0^1 \left\langle \frac{\partial V(\lambda)}{\partial \lambda} \right\rangle d\lambda \quad (43)$$

the left-hand side is the Helmholtz free energy difference and the right-hand side may be approximated by a discrete sum.

$$\Delta A = \sum_i \left\langle \frac{\partial V(\lambda)}{\partial \lambda} \right\rangle \Delta \lambda_i \quad (44)$$

The use of this equation for calculating ΔA is Thermodynamic Integration (TI). The main distinction between these methodologies is that TI averages over a differentiated energy function, whilst FEP averages over finite differences. The computational expense lies in the generation of the ensembles, not the averaging, and so both methods can easily be used on the same simulation trajectories.

From eq. 34 ΔG can then be calculated from ΔA . Statistical estimators such as

Bennett’s acceptance ratio (BAR)²⁴² are used to compute the free energy as

$$\Delta G_{\text{env}} = k_B T \sum_{k=0}^{K-1} \Delta f(\lambda_k, \lambda_{k+1}) \quad (45)$$

where Δf is the unitless free energy difference

$$\Delta f(\lambda_k, \lambda_{k+1}) = f(\lambda_{k+1}) - f(\lambda_k) = -\ln \frac{Q(\lambda_{k+1})}{Q(\lambda_k)} \quad (46)$$

The similarity of these states can be exploited by multi-state estimators such as Multistate Bennett acceptance ratio (MBAR) that utilise configurations sampled at all alchemical states to compute free energy differences, reducing time-complexity of estimation and allowing statistical uncertainty to be determined.²⁴³ FEP methods have been shown to be accurate to within ~ 1 kcal/mol,²⁴⁴ and are routinely used in modern drug discovery workflows.^{245,246} FEP calculations can fail to accurately predict $\Delta\Delta G_{\text{bind}}$, though, if the free energy change is dependent on a process that is not captured by the simulation. Induced-fit effects can additionally cause inaccurate predictions.²⁴⁷ Modern research into the application of FEP has been centred around automating them at scale,²⁴⁸ incorporating NNPs in calculations as opposed to classical force fields²⁴⁹ and using ML methods to reduce the number of transformations needed to find top scoring molecules²⁵⁰ (for example, by using active learning methods, discussed in Chapter 6).

These advancements enable more efficient and accurate predictions of binding affinities. A ligand’s binding can also be calculated computationally using MD simulations through other methods such as Molecular Mechanics with Generalised Born and Surface Area solvation (MMGBSA) and docking. Docking is the the least sophisticated of the methods mentioned here and is discussed in the following section.

3.4 Docking

Docking is a widely used tool²⁵¹ that aims to predict the position and orientation of a ligand within a protein binding site, as well quantify the binding affinity it has with its target. It necessitates approximations in order to do this at low computational cost, which makes it a practical choice for applications such as virtual high throughput screening (HTS).²⁵² There are two fundamental components to docking: the search or sampling algorithm and the scoring function. Search algorithms explore the translational, rotational, and conformational space of the ligand within the binding pocket in an effort to find native binding modes or poses, which the scoring function then adjudicates.

Two common types of algorithm used are: 1) systematic, such as that used in Glide²⁵³, and 2) stochastic Monte Carlo methods, as in LigandFit.²⁵⁴ A systematic

algorithm examines each degree of freedom of the ligand incrementally and, as such, the computational complexity increases exponentially with degrees of freedom and are liable to be trapped in local minima.²⁵⁵ Stochastic methods, in contrast, randomly accept or reject changes to the ligand with an acceptance probability proportional to the Boltzmann factor, to better sample the entire space.

For scoring functions, there are four general types: 1) forcefield-based, 2) empirical, 3) knowledge-based, and 4) machine learning methods.

1. Forcefield, or physics-based, scoring functions use the sum of energy terms (e.g. van der Waals, electrostatics, dihedral potential energy) to approximate binding free energies between the ligand and protein, with each term representing a quantity of physical significance. These sorts of scoring functions can be expensive, however.
2. Empirical scoring functions are based off QSAR/binding-affinity data and the idea was first formalised by Hansch and Fujita in 1964.²⁵⁶ The premise of an empirical scoring function is using linear regression on a set of known binding affinities, parameterising the weights of the contributing energy terms. This reduces the computational complexity whilst retaining accuracy.²⁵⁷
3. Knowledge based scoring functions use the frequency of pairwise interactions from experimentally determined 3D structures to create a potential via Boltzmann inversion, knowledge based methods afford a great compromise between accuracy and computational expense.²⁵⁸
4. Machine-learning based scoring functions are a more recent development, and can outscore classical methods²⁵⁹ - but their scoring power is heavily dependent on the training set used to create the model²⁶⁰ and in practice, a lot of their success has been found in rescoring classical scoring functions.^{261 262}

In general, there is no ‘best’ scoring function, and the type of scoring function desired is dependent on the particulars of the target system.

3.4.1 gnina

The main docking scoring function used in this thesis (both in FEGrow and for separate docking tasks) was gnina, which is a deep-learning docking framework that is a fork of SMINA. It is an ensemble of four CNN models using two different model architectures and training sets, and these ensembles each contain five models that were trained with different seeds.⁶ The models were trained on the PDBbind database and DUD-E dataset, by creating 3D grids of voxels of protein binding sites which capture pharmacophoric data (such as the positions of 14 different ligand/receptor atom types corresponding to various

functional groups e.g. aromatic/aliphatic). The constituent models have been trained to use this conformational data to predict experimental affinities as well as the probability a pose has a low RMSD to the actual binding pose. Prospective poses of a ligand are generated and evaluated using Monte Carlo sampling and accepted or rejected using a Metropolis acceptance criterion.

gnina has been shown to outperform similar models (that use simpler chemical descriptors) fit to the same training data as well as classical empirical scoring methods, with the RMSE of affinity prediction found to be 1.5 pK units across a curated set of 290 complexes from PDBbind. gnina has also been shown to outperform Vina across 89 DUDE-E targets when directly compared.

In this work, gnina was also deployed successfully (independent of FEgrow) to predict active compounds for the CACHE#2 Challenge, attaining the highest score out of any other protocol (see section 5.4.3). CNN models, although propitious on specifically curated benchmarking tasks, still lack the ability to properly generalise beyond the domain of their training data, and so special attention is needed to provide robust training sets that better represent real-world use cases. One such example is CrossDocked2020, which contains over 10 million ligand poses, in protein-ligand complexes.²⁶³

3.5 Uniform Manifold Approximation (UMAP)

High-dimensional datasets, characterised by a large number of features or variables, are common across various fields but present significant challenges for visualisation and analysis; for instance, in the case of visualising molecular fingerprints, a 1024-bit would require 1024 dimensions to visualise. Dimensionality reduction techniques aim to transform these datasets into lower-dimensional spaces, often for visualisation or further analysis. Principal Component Analysis (PCA) is a widely used method that employs eigenvectors to identify the principal components, which capture the directions of greatest variance in the data. However, PCA is limited in its ability to model non-linear relationships between variables.²⁶⁴

Uniform Manifold Approximation and Projection (UMAP) is an alternative dimensionality reduction technique that scales well with larger datasets and preserves both local and global data structures. UMAP takes in points (a dataset) and returns a two dimensional representation of those points. It does this via constructing a weighted graph (in the original number of dimensions) to represent data relationships and optimises a lower-dimensional projection that maintains a similar topology, which can then be visualised as a regular two dimensional plot. This is a stochastic process and so care has to be taken to make sure random noise is not interpreted as actual structure in the data and can be achieved by tuning hyperparameters in the UMAP algorithm. These include the ‘number of nearest neighbours’, which balances the focus between local and global structures, and ‘minimum distance’ which controls the compactness of data points in the reduced space.

UMAP is based on algebraic topology, when presented with a dataset it initially forms simplices between data points (the simplest shape that can be made in n dimensional space e.g. a triangle in 2D, and a tetrahedron in 3D). These simplices define the connections between data points and their neighbours in the original high-dimensional space, and is termed a simplicial complex - a graph made up of simplices. UMAP aims to project this simplicial complex into a reduced-dimensional space (usually two dimensions), maintaining the interrelationships among the simplices to the greatest extent feasible, to enable visualisation.

UMAP begins constructing the simplicial complex by forming a k -nearest neighbour (kNN) graph for the dataset. In this representation, every data point is linked to its k closest neighbours. This kNN graph is subsequently converted into a fuzzy simplicial set by attributing a weight to each edge within the graph. The weight calculation depends on the local connectivity of the data points and is given by

$$w_{zo} = \exp \left(-\frac{d(x_z, x_o) - \rho_z}{\sigma_z} \right) \quad (47)$$

where $d(x_z, x_o)$ is the distance between data points x_z and x_o , ρ_z is the local connectivity of data point x_z , σ_z is a scaling factor based on data density and w_{zo} represents the weighted connection between points z and o .

The construction of fuzzy simplicial sets enables UMAP to encompass both the local and global structures within the dataset. The edge weights in the graph represent the strength of the connections among data points.

$$C(P, Q) = \sum_{z,o} \left(p_{zo} \log \left(\frac{p_{zo}}{q_{zo}} \right) + (1 - p_{zo}) \log \left(\frac{1 - p_{zo}}{1 - q_{zo}} \right) \right) \quad (48)$$

Where $C(P, Q)$ is the cross-entropy cost function p_{zo} is the probability of an edge between points z and o in the high-dimensional space q_{zo} is the probability of an edge between points z and o in the low-dimensional space calculated via Student's t-distribution kernel, with $\sum_{z,o}$ indicating summation over all pairs of points z and o . This cross-entropy formula measures how well the low-dimensional embedding preserves the relationships between points from the original high-dimensional space and can be iteratively optimised by the UMAP algorithm to optimise the layout of data in a low dimensional space, thus minimising the error between the two topological representations.

3.6 Active Learning (AL)

Active learning²⁶⁵ is a subset of machine learning that is based on iteratively labelling data points from an unlabelled dataset (in our case, *de novo* compounds that are built into protein binding pockets and scored). The aim is to pick the most useful samples for training a surrogate model whilst ultimately minimising the potentially expensive computation needed to find instances that maximise an objective function. There are two main components to an active learning workflow: the regression model and the acquisition function. Every scored instance is used to train a specified machine learning model with more examples refining the model accuracy, which is then used to select new molecules to be built. In this work, we consider and benchmark two models.

3.6.1 Gradient Boosted Machine (GBM)

The first approach is gradient boosted machine (GBM), which is a random forest-based technique, utilising ensembles of decision trees. These trees are created from random subsets of features (fingerprints) that are then used to make predictions. GBMs expand on traditional random forests by using the gradient of the error to construct trees specifically designed to minimise this error. Gradually increasing the number of relatively poor individual trees additively increases their predictive power (hence ‘gradient boosted’). Given a molecule i with a vector of descriptors x_i , \hat{y}_i , the predicted output of the GBM model is defined as

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) \quad (49)$$

where K is the number of trees and $f_k(x_i)$ for the output of the k th tree.

3.6.2 Gaussian Process

The second model is Gaussian Process (GP) regression, which is a Bayesian approach that makes predictions by assuming observations can be modelled by the probability distribution over possible reasonable (Gaussian) functions.²⁶⁶ These Gaussian distributions are iteratively refined by the observation of new samples. Because model prediction is performed via a probability distribution, it natively incorporates uncertainty and other useful quantities, such as estimates of expected improvement of a given new sample²⁶⁷. An arbitrary function $f(x)$ is approximated as a linear combination of kernel functions

$$\hat{y}(x)_i = \sum_{m=1}^M c_m k(\mathbf{x}_i, \mathbf{x}_j) \quad (50)$$

where x_i and x_j represent input vectors in the feature space, the basis functions, k are placed at arbitrary locations and typically are gaussian.

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2l^2}\right) \quad (51)$$

\mathbf{x}_i and \mathbf{x}_j represent input vectors in the feature space, $\|\mathbf{x}_i - \mathbf{x}_j\|^2$ is the squared Euclidean distance between the input vectors and l is the length-scale parameter, controlling the smoothness of the kernel function.

The acquisition function defines the method by which new samples are picked at the start of each active learning cycle, with the simplest example being a ‘greedy’ approach, which directly selects the (currently predicted) highest scoring potential samples (for example predicting a binding affinity for molecular dataset). However, an acquisition function has to balance the selection of the best compounds with the need to further refine the accuracy of the machine learning model. Picking the best scoring samples in descending order might initially increase the objective function, but the algorithm will have the propensity to get stuck in local maxima and to be sensitive to the initial selection of samples.

There are a variety of alternatives that aim to avoid the problems of a simple greedy approach, and the approach used here is the upper confidence bound (UCB) uncertainty-based acquisition function.²⁶⁸ UCB considers not just the value of the objective function, but also the variance of the prediction (model uncertainty), effectively biasing towards the selection of samples about which the model is the least certain of the predicted score. The UCB function is defined by

$$UCB(m) = \mu(m) + \beta\sigma(m), \quad (52)$$

where $\mu(m)$ and $\sigma(m)$ are the mean and standard deviation of the regressor for molecule m , and β is a parameter controlling the degree of exploration (high β increases the chances that an observation with moderate score but high uncertainty will be made).

4 FEgrow

De novo design is a complicated and at times a subjective process; the goal of designing an inhibitor for a target does not have a single ‘right’ answer and requires iterative feedback with experiment.

FEgrow was developed to automate and simplify the computational *de novo* process by integrating a suite of CADD tools, such as OpenMM (MD), PDBFixer (protein preparation), gnina (docking scoring function) and ANI (a neural network potential for conformer energy minimisation). The creation of this package/workflow gives non-expert experimentalists the ability to rapidly test *de novo* designs, while offering computational practitioners the ability to set up in a semi-automated fashion more rigorous calculations, such as FEP.

In this chapter, the development of FEgrow and its ability to recapitulate crystallographic poses, rank order designed ligands, and produce aligned congeneric series of ligands that can optionally be exported for further simulation is detailed.

This chapter is a reprint of a published paper.¹⁷

4.1 Introduction

Computational structure-based molecular design, in particular aiding the discovery of novel chemicals with desired biological activity, plays a crucial role in the modern drug discovery pipeline. High-throughput virtual screening is widely used in hit discovery²⁶⁹, but relies on pre-defined libraries of compounds. *De novo* design software packages aim to construct a model of a ligand in a target binding pocket using growth algorithms, either starting from a scaffold of a known hit compound or entirely from scratch. Such approaches can be beneficial as they do not rely on a (physical or virtual) library, and molecules can be tailored specifically to the problem at hand. Advances in *de novo* design software have been extensively reviewed²⁷⁰, and examples include both rule-based generation methods such as OpenGrowth²⁷¹, AutoGrow²⁷², and LigBuilder²⁷³, and recently deep generative methods for molecule design²⁷⁴.

With advances like these described above, much progress has been made in the important problem of optimising a molecular design within the context of a pre-defined scoring function and binding pocket. However, whether the designed molecule indeed has high biological activity is crucially reliant on the accuracy of the methods that are used to generate and score poses of the designed molecules, as well as other assumptions, such as a rigid receptor, that might be employed. Furthermore, the generated molecules can be quite esoteric, which may be advantageous with regards to arriving at novel intellectual property, but may not be ideal from a synthetic tractability viewpoint²⁷⁵. More commonly, a drug discovery effort may have identified a hit compound with a

well-defined binding mode and wish to explore structure-activity relationships amongst a small library of synthetically accessible analogues. In this case, it would be beneficial to make use of prior knowledge about the binding mode when generating poses of designed compounds. One example of this approach is the E-novo workflow²⁷⁶, which was made available in Pipeline Pilot or Discovery Studio. The available conformations of added chemical functional groups (R-groups) were enumerated with a rigid core, and scored using a CHARMM-based docking method. The physics-based molecular mechanics-generalised Born with surface area (MM-GBSA) was then used to provide a more accurate score. Further, more recent, examples include FragExplorer²⁷⁷, which aims to grow or replace fragments to optimise molecular interaction fields generated by the GRID software²⁷⁸, DeepFrag²⁷⁹, which predicts appropriate fragment additions using a deep convolutional neural network trained on thousands of known protein–ligand complexes, and DEVELOP²⁸⁰, which uses deep generative models to output 3D molecules conditional on provided pharmacophoric features of the binding site. However, the employed approximate physics- or knowledge-based approaches to scoring the designs will limit to some extent their ability to predict and optimise binding affinity.

On the other hand, free energy methods are much more computationally expensive approaches to molecular design that employ rigorous thermodynamics and carefully parameterised force fields to compute (relative or absolute) protein–ligand binding free energies. As such, they overcome many of the accuracy limitations of *de novo* design workflows, and are commonly employed in prospective design efforts to explore and prioritise relatively small perturbations in the hit-to-lead stage²⁸¹. Many excellent tutorials and best practice documentation are available^{282–286}, but most start from the assumption that the user has already built initial poses of the ligands in the binding pocket. For simple R-group additions, input coordinates may be built from maximum common substructure alignment, for example, but it may be difficult to resolve steric clashes or decide between two alternative 3D poses in more complicated cases²⁸⁴. Some widely-used graphical user interfaces, such as Maestro²⁸⁷ and Chimera²⁸⁸, are also available for building R-groups, but these can be proprietary and/or difficult to build into automated workflows and modify according to user needs.

Notable successful computer-aided design efforts have used free energy calculations in conjunction with *de novo* design tools to build (and maybe score) new molecules. Jorgensen and co-workers have pioneered this approach for many years, linking *de novo* design through the biochemical and organic model builder (BOMB) software with free energy perturbation (FEP) through the MCPRO software²⁸¹. BOMB builds ligands into a binding pocket by linking user-defined R-groups to an existing core. Functionality is available for conformer searching, structural optimisation and scoring, using a custom scoring function trained via linear regression on > 300 experimental activity data

points²⁸⁹. Once hits have been built and scored, hit-to-lead optimisation may be further refined through free energy calculations. Such an approach has yielded extremely potent series of leads against HIV reverse transcriptase²⁹⁰, macrophage migration inhibitory factor²⁹¹ and the SARS-CoV-2 main protease¹⁶. In other drug discovery programmes, as part of the recent COVID Moonshot open science effort to crowd source design of SARS-CoV-2 main protease inhibitors²⁹², the Omega toolkit by OpenEye²⁹³ is used for constrained conformer generation, and optimal binding poses are then taken through to free energy calculations using the perses software²⁹⁴. The evident importance of input structure to the reliability of free energy calculations²⁸⁴ means that open-source tools to automate this step are crucial.

Inspired by the BOMB/MCPRO approach to molecular design²⁸¹, we introduce here the FEgrow open-source workflow for growing functional groups, chosen by the user, from a defined position on a core compound. To account for the multi-objective nature of molecular design, we output simple rule-of-five indicators of oral bioavailability, as well as flags for undesirable substructures and synthetic accessibility estimates. For the designed ligands, we enumerate 3D conformers of the added R-group, with options for additional flexibility if desired, within the context of the protein (discarding conformers with steric clashes). A common issue with generating docked poses is inaccuracy in the molecular mechanics force fields used to refine them, particularly for uncommon chemistries. To overcome this, we employ a hybrid machine learning / molecular mechanics (ML/MM) approach to optimisation, whereby the ligand is (optionally) described by the ANI neural network potential^{229,295}, and non-bonded interactions with the static protein are described by traditional force fields. The binding affinities of low energy poses are predicted using a deep learning based scoring function. Finally, FEgrow outputs binding poses in a form suitable for input to free energy calculations, and here this process is illustrated with a case study, using the SOMD software²⁹⁶ to retrospectively compute relative binding free energies of several inhibitors of the SARS-CoV-2 main protease¹⁶.

In this way, the goal is to integrate medicinal chemistry expertise in the FEgrow design workflow, with state-of-the-art methods for pose prediction, scoring and free energy calculation. By building ligands from the constrained core of a known hit, we maximise the use of input from structural biology, and reduce reliance on docking algorithms. The overall aim is for an open-source, customisable, fast and easy-to-use (accessed through Jupyter notebooks) workflow that can adapt to community needs and advances in molecular design.

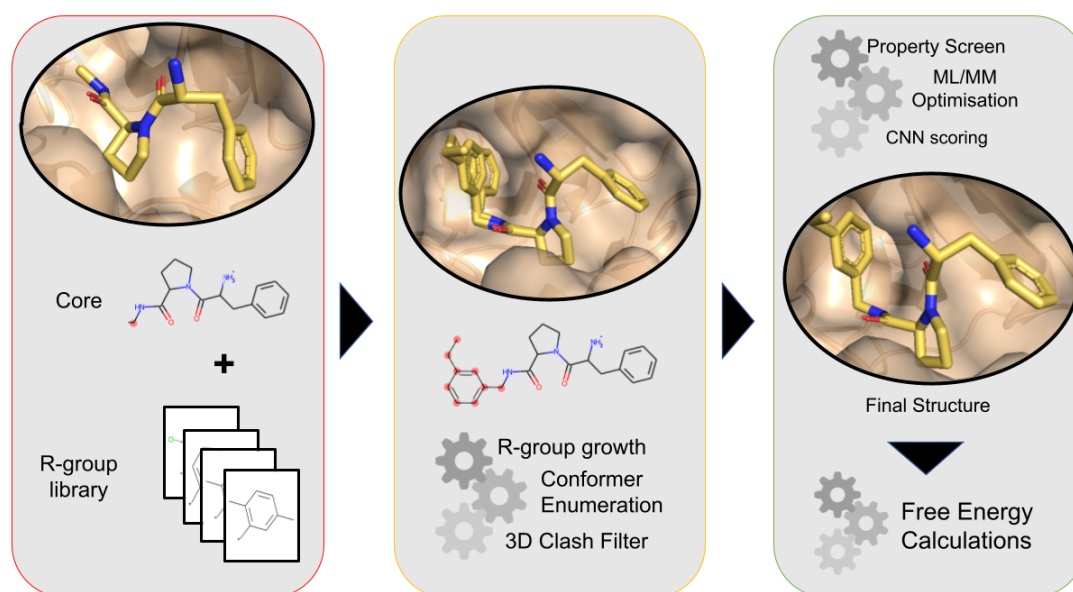


Figure 14: The FEgrow workflow. (left) The user specifies the receptor, ligand core, and a list of functional groups, along with their attachment points. (centre) RDKit⁵ is used to attach the selected R-group(s) and enumerate the available conformers with a rigid core. (right) Possible bioactive conformers undergo structural optimisation using a hybrid ML/MM potential energy function. The binding affinity is predicted using a convolutional neural network scoring function⁶ and molecular properties are assessed. Final structures are output for further free energy based binding affinity assessment.

4.2 Results

4.2.1 Workflow Design

The FEgrow package is written in Python, and supports Jupyter visualisation at each stage using py3Dmol²⁹⁷. Underneath, the main unit in the package is RMol which extends the RDKit class `rdkit.Chem.rdchem.Mol`⁵ with additional functionalities, such as visualisation, molecule merging, conformer generation, as well as storage of energies and other metadata. A convenience class RList is provided with the same functions for operating on a set of molecules, which allows also for future parallelisation. A modular workflow allows for addition/removal of functionality, such as new scoring functions or optimisation algorithms. FEgrow is freely available at <https://github.com/cole-group/FEgrow>, along with a tutorial. Figure 14 shows the overall design of the FEgrow workflow, and the component methods are described in the following sections.

4.2.2 Input and Constrained Conformer Generation

The first task is to define the receptor and the ligand core, along with an attachment point for growth (currently only growth from hydrogen atoms is supported). Users may download receptor and ligand structures directly from the protein databank (PDB), or upload pre-prepared structures. In this study, we used the Open Babel software²⁹⁸ for parsing input structure files and ligand protonation (at pH 7).

Merging the ligand core with a new R-group requires that both the linking atom on the template core and on the attachable R-group are specified. The merging is carried out with the RDKit editable molecule⁵. RDKit is further used to generate 3D conformers using the ETKDG method²⁹⁹. The generated conformers are aligned, and energy minimised using the Universal Force Field³⁰⁰. Harmonic distance restraints to their initial positions are applied to atoms in the common core (identified by a maximum common structure search) using a stiff force constant (10^4 kcal/mol/Å²). In this way, we can enforce the conformations of the generated molecules to only vary from the core in the region of the added R-group. This region may additionally be extended by adding further atoms into the “flexible substructure” of the template. For convenience, we provide a minimal set of around 500 R-groups that are commonly used in medicinal chemistry optimisation³⁰¹.

R-groups can be interactively selected from the library using the mols2grid package³⁰², or the user may prepare their own molecules for attachment (see **Tutorial**).

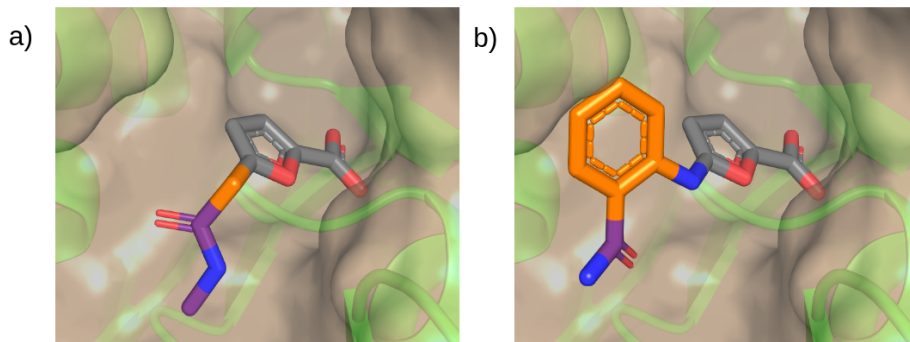


Figure 15: Two different R groups (purple) and linkers (orange) grown from a common core (grey) in an example protein receptor. The core is restrained but the added groups are kept flexible to be optimised (optionally with a machine learning potential) to produce low energy conformers, which can be input into free energy calculations.

4.2.3 Geometry Optimisation

The constrained conformer generation described above aims to enumerate all accessible, physically-reasonable conformers of the added R-group (and any other flexible regions) in vacuum. However, most of these conformations will be incompatible with the protein binding site. Hence, a 3D filter and geometry optimiser aim to find the bioactive conformers of the designed ligands.

The protein is treated with PDBFixer³⁰³ to add any missing atoms, residues and hydrogen atoms. Water molecules (and other non-protein residues) are stripped by default, but can be optionally retained as part of the receptor, for example if they are thought to form an important part of the hydrogen bonding network within the binding pocket (an example is shown later in Case Study I). A simple distance filter removes any ligand conformers that form a steric clash with the protein (any atom–atom distance less than 1 Å). Next, the remaining conformers are refined in the context of a rigid receptor via energy minimisation using OpenMM³⁰³. All atoms of the protein, and any retained water molecules, are kept fixed during the optimisation in the positions provided by the user.

The energy minimisation uses the AMBER FF14SB³⁰⁴ force field for the receptor and either GAFF2¹⁸⁹ (General AMBER force field) or the Open Force Field 1.0.0 (‘Parsley’)³⁰⁵ general force fields for the ligand, with the choice left to the user. Optionally the intramolecular interactions of the ligand can be modelled using the ANI-2x ML potential²²⁹ in a hybrid ML/MM simulation. In this so-called “mechanical embedding” scheme, the total potential energy of the ML/MM system is composed of three terms³⁰⁶:

$$E^{tot} = E^{MM}(R) + E^{MM}(RL) + E^{ML}(L), \quad (53)$$

where R, RL, and L correspond to receptor-receptor, receptor-ligand and ligand

intramolecular interactions, respectively. The second term acts as the coupling term between the ML and MM subsystems and consists of the standard Coulomb and Lennard-Jones 12-6 non-bonded interaction energies. Thus, a general force field (here, GAFF2 or Parsley) is still required for the ligand to model the non-bonded interactions with the receptor. The use of ANI helps to avoid known deficiencies in the potential energy surfaces predicted by force fields, while ensuring that the optimisations are significantly faster than could be achieved with full quantum mechanics. For example, it has been shown that the description of biaryl torsions, which are commonly found in drug-like molecules, is one area where ANI-2x performs better than contemporary general force fields⁴. The hybrid ML/MM approach has also been shown to predict binding poses that overlap well with crystallographic electron density maps of bound ligands, even for those that contain charged moieties that were not included in the training of the ANI potential³⁰⁷. As such, in FEgrow, users may turn on the hybrid approach for binding pose refinement provided the molecule contains only elements covered by the model (H, C, N, O, F, S, Cl), else the selected classical force field is used for the entire ligand.

The lowest energy optimised conformer, and all conformers within 5 kcal/mol, are output as PDB/SDF files for further analysis and scoring.

4.2.4 Binding Pose Scoring

Once the geometry optimisation is completed, the remaining (low energy) conformers are scored to predict their binding affinity. There are many choices available for scoring binding poses and their corresponding binding affinities, and these are usually classified as either force-field, empirical, or knowledge based. In the latter case, input features (such as atom-atom pairwise contacts) are used to train models to reproduce data for known protein–ligand complexes. Recently, machine learning models have emerged, in which an arbitrary, nonlinear relationship between input and target prediction is learned. One such approach is the gnina convolutional neural network (CNN) model³⁰⁸, which takes as its input features a 3D grid-based representation of the protein–ligand complex and the atom types. The model has been jointly trained for binding pose and affinity prediction on a cross-docked set containing examples of ligand poses generated by docking into multiple receptors³⁰⁹. The resulting models are competitive with other grid-based CNN models, and outperform the traditional empirical Vina scoring function³⁰⁹. They are available as part of the gnina docking software package⁶, which is a fork of Smina³¹⁰ and AutoDock Vina³¹¹. Here, we use gnina only for re-scoring the output ligand 3D structures, using the ‘score_only’ flag and the default ensemble of CNN scoring models. Gnina CNNaffinity scores (predicted pK) are output, and compared with experimental binding affinity (where available).

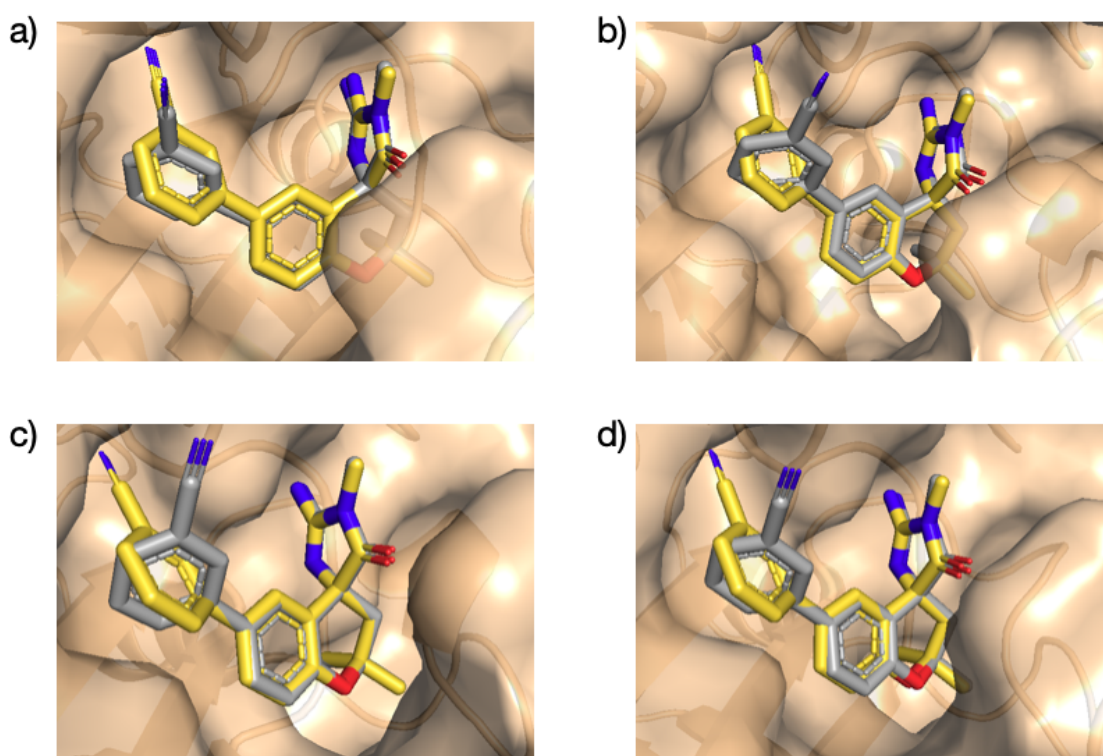


Figure 16: Overlay of protein–ligand benchmark dataset structures for the BACE(Hunt) target (PDB: 4JPC). Crystal structure in yellow and grown compound in grey. a) including water in the binding pocket as part of the receptor, b) using ANI for optimisation, c) using GAFF for optimisation, d) setting relative permittivity (ϵ) and the Lennard-Jones radii scaling factor to 1.0.

4.2.5 Molecular Property Filters

Having assembled the 2D and 3D structures of the core and user-defined R-groups, we include some simple tools for assessing the drug-likeness and synthetic tractability of the designed compounds. Several sets of rules exist to investigate the likelihood of a molecule displaying “drug-like” behaviour. While there are many examples of approved drugs which violate these considerations, they still provide a useful indication of whether a molecule is worth testing (that is, if it disobeys all of the conditions discussed below, it is most likely a poor candidate). FEgrow reports Lipinski’s rule of five (Ro5) counts,⁶⁴ the synthetic accessibility score (SAScore)³¹² and flags describing whether the proposed molecule is Ro5 compliant and if it contains undesirable features based on the PAINS,³¹³ NIH^{314,315} and unwanted substructure³¹⁶ filters. Our implementation is adapted from the TeachOpenCADD³¹⁷ Talktorials 2 and 3, using functionality from the Descriptors and FilterCatalog modules of RDKit.⁵ Further details are provided in **Supporting Information Section S2**.

4.2.6 Analysis of Lennard-Jones and Electrostatic Scaling Factors.

Following the procedure recommended in BOMB, Lennard-Jones radii are scaled by a factor of 0.8 during optimisation. This is intended to mitigate to some extent the rigid protein approximation, by allowing extra space in the binding pocket to accommodate ligand growth. Furthermore, to account in an implicit manner for the neglected dielectric response of the protein and solvent molecules, the atomic charges are reduced by a factor of $\frac{1}{\sqrt{\epsilon}}$, where ϵ is the relative permittivity, in this case taken to be 4. Analysis of the effect of these scaling factors on structural and affinity predictions is shown in Fig 17. We follow the procedure recommended for the BOMB *de novo* design software³¹⁸ in scaling the Lennard-Jones radii and atomic charges when calculating the intermolecular energetics during geometry optimisation. This is to attempt to mitigate the rigid receptor approximation by i) allowing extra space in the binding pocket to accommodate ligand growth and ii) screening electrostatic interactions using an effective dielectric medium.

Table 1 shows the effect of these choices of scaling factor on the correlation between gina predicted binding free energies and experiment (R^2) for the set of thrombin inhibitors, and on the RMSD between the output R-group coordinates and crystal structure (see Case Study I in the main text for a full description). The correlation between gina and experiment is very similar when using optimised structures either with no scaling, or with a scaling factor of 0.8 (and dielectric of 4). The correlation is lower using a scaling factor of 0.9, but on closer inspection this is due to a single outlier (Figure 17), and removing this point increases the correlation to 0.77.

It is interesting to note that, as shown in Figure 17, the predicted binding affinities

become more favourable as the scaling factors tend towards one. This might be expected, as the van der Waals radii of the atoms approach their physical values, the structures are closer to the optimal binding poses as recognised by the CNN scoring function. This is corroborated by the observation that the agreement in structure between the FEgrow output and the known crystal structure (PDB: 2ZFF⁸) improves as the scaling factors are removed (RMSD decreases from 1.4 to 0.8 Å, Table 1). Despite these improvements in the absence of the scaling factors, we prefer to retain their use in the default behaviour of FEgrow to allow the possibility of more hits to be identified during prospective design. Nevertheless, the user is free to adjust them during run time.

Table 1: Effect of the LJ radii scaling factor and relative dielectric permittivity (ϵ) used during optimisation on the accuracy of predicted affinities and structures from FEgrow for the set of thrombin inhibitors (see Case Study I).

LJ scaling	ϵ	R^2	RMSD / Å
0.8	4	0.68	1.37
0.9	2	0.27	1.05
1.0	1	0.70	0.76

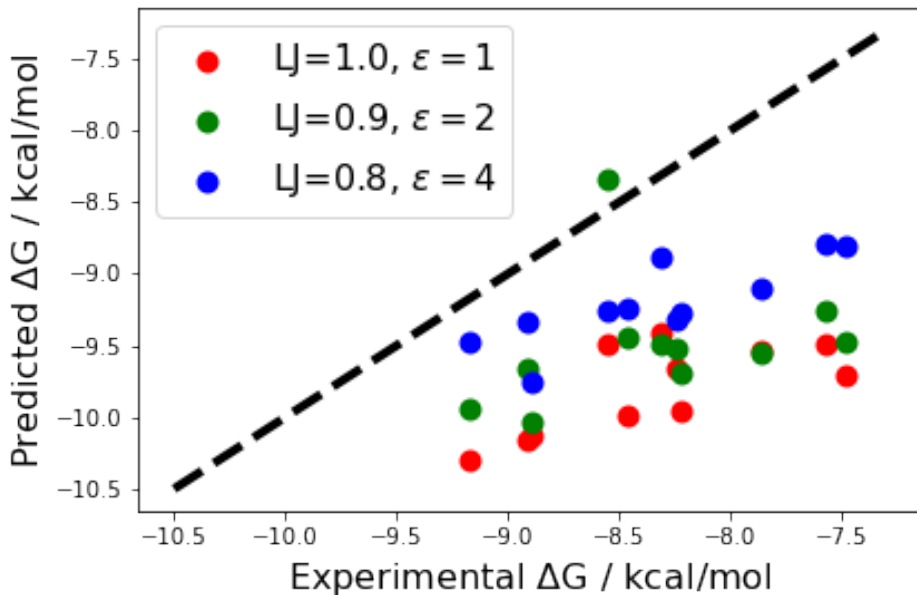


Figure 17: Effect of the LJ radii scaling factor and relative dielectric permittivity (ϵ) used during optimisation on the correlation between predicted and experimental binding free energies for the set of thrombin inhibitors (see Case Study I).

4.3 Case Study I: Protein-Ligand Benchmarks

The protein–ligand benchmark of Hahn et al. is an open, curated set of high quality structural (e.g. high similarity between crystallised and simulated ligands and no missing atoms) and bioactivity (e.g. taken from a single data source with adequate dynamic range) data, which has been collected with the goal of assessing the accuracy of free energy methods³¹⁹. For each target, modelled structures of the protein in complex with a congeneric series of ligands are provided as starting points for free energy calculations, but the methods used to position the R-groups are, to our knowledge, not necessarily consistent or documented.

Here, we apply the FEgrow workflow to ten targets from the protein-ligand benchmark set. Starting from the crystal structure of each target, we truncate the bound ligand to a common core, which is shared across the congeneric series to be modelled. A summary of the targets, the crystal structures used, the number of R-groups grown, and their common

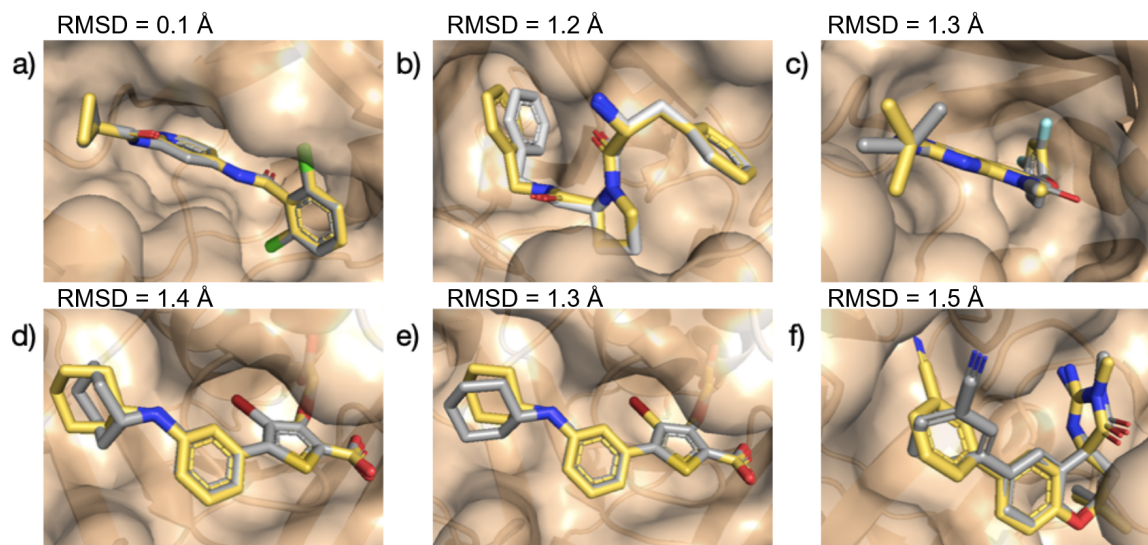


Figure 18: Overlay of protein-ligand benchmark dataset structures (crystal structures in yellow and grown compounds in grey). a) TYK2 (PDB: 4GIH⁷), b) Thrombin (PDB: 2ZFF⁸), c) P38 (PDB: 3FLY⁹), d) PTP1B with force field optimisation (PDB: 2QBS¹⁰), e) PTP1B using ML/MM optimisation, and f) BACE(Hunt) (PDB: 4JPC¹¹). Root-mean-square distances (RMSD) between predicted and experimental coordinates of atoms in the built R-groups were calculated using RDKit⁵.

core and net charge is provided in **Tables S2** and Figure 19.

We use the methods outlined in the previous section to re-grow the congeneric series of ligands in the binding pockets, including enumeration and optimisation of possible R-group conformers, and scoring of final poses. Figure 18 shows overlays of the modelled and crystal structures (where there is an exact match between the crystallised ligand and one of the modelled R-groups), as well as the measured root mean square deviation (RMSD) between the predicted and experimental coordinates of the heavy atoms of the functional groups.

For the targets, TYK2 (Figure 18(a)) and Thrombin (Figure 18(b)), we obtain a good overlap between the grown R-groups and the crystal structures. In the former case, the dihedral angle formed between the grown cyclopropyl C2 and C3 carbons and the amide carbonyl oxygen core (30° and -37°) are in agreement with those reported experimentally⁷. For Thrombin, although the added R-group here is a rigid phenyl moiety, we make use of the option to add atoms from the core to the flexible region and allow the linking $-\text{CH}_2-$ group to freely rotate during structural optimisation. This added flexibility leads to a rotation of the phenyl group of less than 10° , compared to the corresponding crystal structure⁸.

The P38 benchmark set includes a series of alkyl amino substitutions originally investigated as part of a structure-activity relationship study into kinase inhibitors⁹. Here, the added amino group is correctly positioned to form a hydrogen bond with the

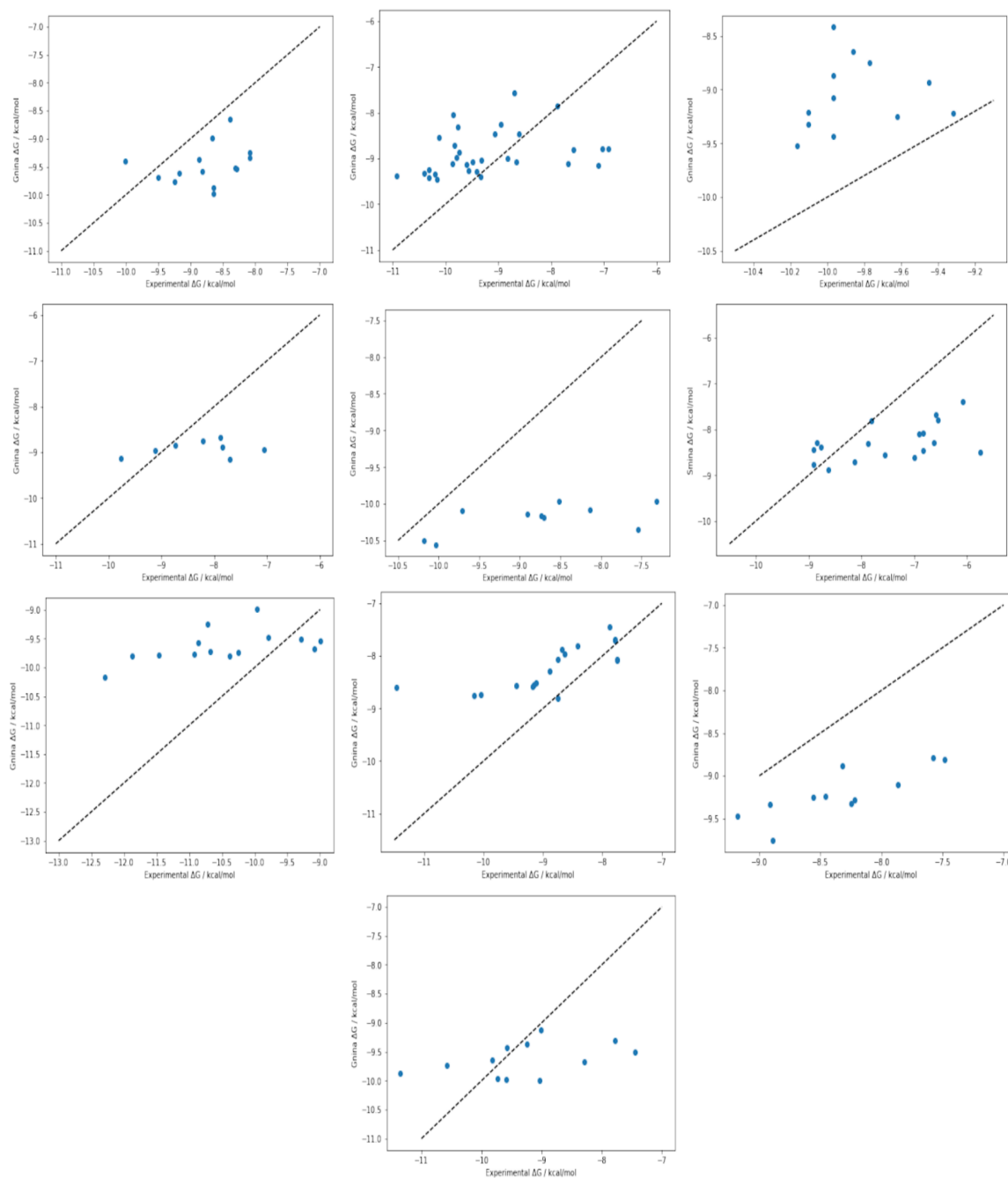


Figure 19: Absolute binding free energies of congeneric series of ligands taken from the protein-ligand benchmark set, using the glna CNN affinity, compared with experiment. Protein targets from top left: BACE, BACE(Hunt), BACE(P2), CDK2, JNK1, MCL1, P38, PTP1B, Tyk2.

protein backbone, though the *i*-Pr group is rotated by around 60° compared to the crystal structure (Figure 18(c)). In PTP1B, the grown cyclohexyl substituent is able to rotate quite freely, with many conformations predicted to lie within 5 kcal/mol of the minimum. The minimum energy structure shows good overlap with the crystal structure¹⁰ with low RMSD, but connects to the core at the axial position of the cyclohexyl group (Figure 18(d)). In this case, the core structure contains a Br atom, so we are unable to optimise with the ANI-2x potential (the workflow defaults to the Parsley force field for the ligand). Interestingly, if we remove the Br atom, and re-run the workflow using hybrid ML/MM optimisation, we recover the equatorial connection as the lowest energy conformer, in agreement with the crystal structure (Figure 18(e)). This demonstrates the potential advantages of employing hybrid ML/MM structure prediction methods in binding mode determination.

Finally, the BACE(Hunt) target, includes a series of substituted phenyl additions to a spirocyclic core. Here, the grown cyanophenyl group is rotated by approximately 90°, relative to the crystal structure¹¹, which shows the *meta*-CN group accommodated in the binding pocket (Figure 18(f)). An exact match with the crystal structure is also output, but it is predicted to be around 3 kcal/mol higher in energy. Closer examination of the experimental structure reveals a crystal water molecule, close to the binding pocket, that is capable of forming a hydrogen bond with the -CN group, and a further network of water molecules that would be displaced by the conformation shown in Figure 18(f). **Figure S2** investigates the effects of including the hydrogen-bonding water molecule in the rigid receptor structure, and changing the force fields used, but no input settings recover the crystal structure.

As discussed, we include with the FEgrow workflow the option to score the output poses of the designed ligands with a scoring function. In particular, we use the gnina convolutional neural network score, which has been trained on both binding pose and affinity prediction³⁰⁹. While accurate recovery of experimental binding affinity is not necessarily expected for current scoring functions, it is useful to evaluate to what extent they can be used to provide guidance in early stage design, ahead of more rigorous physics-based scoring methods. The root-mean-square error between gnina CNN affinities (converted to free energies) and experiment is quite acceptable (**Table S4**), ranging from 0.9 kcal/mol (BACE(P2)) to 1.7 kcal/mol (Jnk1), which indicates that the CNN scoring function is able to predict the affinity range of most of these series. In fact, the errors may be lower than typically expected³⁰⁹, because we are using here additional information from experiment (the binding pose of the core) and not relying on the scoring function to determine the bioactive conformation.

The R² correlation coefficients between the predicted and experimental affinities are more variable (**Table S4**), however, ranging from close to zero (the BACE targets) to 0.68

(Thrombin). The full set of CNN predicted binding affinity data is plotted in **Figure S3**, and reveals that most of the predictions lie in quite a narrow range, compared to the experimental data. We note that this is quite a challenging test for the scoring function, since the modifications made to the core are relatively small and cover a smaller dynamic range in affinity than most test sets. Nevertheless, it seems that current scoring functions have some utility in guiding design, but that more accurate physics-based scoring is required to accurately discriminate between structural changes in the hit-to-lead stage.

4.4 Discussion

We have introduced here FEgrow, an open-source molecular builder and free energy preparation workflow. Taking as input a receptor and ligand core structure, FEgrow aims to build a user-defined library of chemical functional groups of the sort that would typically be used to explore structure-activity relationships with free energy calculations. Inspired by the BOMB approach to molecular design²⁸¹, we grow from a fixed ligand core in order to maximise the use of binding mode information from structural biology sources, and rely on the user’s medicinal chemistry expertise to suggest functional groups that improve binding affinity whilst remaining synthetically tractable. Alternative, generative methods for fragment growth^{279,280} could be incorporated in future, but testing of expert medicinal chemist designs still remains popular today and FEgrow aims to automate this process.

The modular workflow of FEgrow allows us to experiment with functionalities, such as new optimisation or scoring methods. With the use of hybrid ML/MM structural optimisation, in particular, we aim to obtain reliable coordinates for the added R-groups. In this respect, the ANI neural network potential (within the ML/MM approximation) has already been shown to be capable of predicting protein-ligand binding poses in agreement with electron density distributions determined by x-ray crystallography³⁰⁷, and should be significantly more reliable than the general purpose force fields (such as UFF) that are typically used for structure refinement in *de novo* design packages. Updated machine learning potentials or semi-empirical methods³²⁰ can readily be included in future versions of FEgrow.

Ligand designs are evaluated for simple molecular properties, and their binding affinity predicted using the gnina CNN scoring function. Despite the challenge of discriminating between relatively small functional group modifications, the scoring function performs quite well and is useful in providing initial guidance for a number of targets from the protein–ligand benchmark set used here. Nevertheless, we envisage the primary use of FEgrow being as a source of input structures for more rigorous free energy based affinity predictions. We demonstrate this functionality here, using SOMD to calculate the relative

binding free energies of 13 uracil-based inhibitors of the SARS-CoV-2 main protease. Using only a single crystal structure as input (PDB: 7L10) and the FEgrow workflow to build the remaining structures, we obtain excellent agreement with experimental binding affinities (MUE = 0.45 kcal/mol, $R^2 = 0.53$).

We envisage future improvements including the use of a flexible receptor for the growth phase, and future use cases including seeding free energy calculations with multiple low energy conformers. The BACE(Hunt) target in Case Study I highlighted the difficulty of accurately including the energetics and effects on binding affinity of displacing water networks in hydrated binding pockets. There does not currently appear to be a satisfactory means to include water networks into the optimisation or scoring phases of FEgrow, but output structures could be passed to molecular dynamics or Monte Carlo based simulations to assess optimal hydration sites for predicted poses^{321–323}. FEgrow is available for download from <https://github.com/cole-group/FEgrow>, and we welcome suggestions from the community for added functionality.

4.5 Computational Methods

4.5.1 Free Energy Calculations

Structures of 13 inhibitors of the main protease (M^{Pro}) of SARS-CoV-2 were built using the FEgrow workflow and taken through to free energy calculations for accurate physics-based scoring. The PDB structure, 7L10, was used for the receptor. Missing residues (E47 and D48) were added using MODELLER³²⁴, which uses optimisation of a pseudo energy function for loop modelling, and hydrogen atoms were added using Chimera²⁸⁸, which includes options for optimisation of the hydrogen bond network. The BioSimSpace package³²⁵ was used for free energy setup, along with a relative binding free energy protocol described previously³²⁶. The lowest energy conformer for each ligand was parameterised with the GAFF2 force field, using the AM1-BCC charge model. The AMBER FF14SB³⁰⁴ force field was used for the protein, along with the TIP3P water model. Each ligand was then solvated in a 35 Å cube, or 90 Å cube in the presence of the protein. The bound and unbound structures then underwent a short equilibration using the default procedure in BioSimSpace³²⁵. Namely, the structure was minimised, then heated to 300K in the NVT ensemble over a period of 10 ps. It was then equilibrated for a further 10 ps in the NpT ensemble at 300 K and 1 bar, using the Langevin thermostat and Berendsen barostat. Atoms in the protein backbone were restrained to their initial positions throughout, and a 8 Å nonbonded cutoff was applied.

The network of alchemical transformations was built manually to include cycle closures for error analysis, and is shown in **Figure S4**. **Table S6** shows that the absolute cycle closure errors are typically less than 0.5 kcal/mol, and less than 1 kcal/mol for all cycles. The overlap for each perturbation was determined using a maximum common substructure search to determine the atoms to be morphed. Each transformation leg was simulated using the SOMD software package²⁹⁶ for 4 ns, and the first 400 ps were discarded as equilibration. Eleven equally-spaced λ windows were employed between 0 and

1, along with the default soft core. The time step was set to 2 fs, with constraints applied to unperturbed hydrogen bonds. Simulations were performed in the NpT ensemble, using an Andersen thermostat with collision frequency of 10.0 ps^{-1} and a Monte Carlo barostat with a frequency of 25 time steps. Periodic boundary conditions and a tapered nonbonded cutoff distance of 10 \AA were applied. Electrostatic interactions were calculated using the reaction-field method with a dielectric constant of 78.3 outside the nonbonded cutoff³²⁷. All transformations reported here were run in both forward and backward directions, and in duplicate. Free energy changes and their errors were calculated from the output with MBAR using the asymptotic covariance method²⁴³. Final free energies and their associated error bars (Figure S5) were calculated from the network with the freenrgworkflows package³²⁸, using the method of Yang et al²⁰.

5 CACHE (Critical Assessment of Computational Hit finding Experiments)

The Critical Assessment of Computational Hit finding Experiments (CACHE) organisation is a public-private partnership non-profit entity that runs triannual collaborative computational hit-finding competitions ('Challenges'). CACHE Challenges aim to benchmark a range of computational approaches to hit-finding for targets with different available experimental data.

Each Challenge focuses on a particular protein that meets the desirable criteria for drug design. Specifically, the chosen protein should present sufficient difficulty, be experimentally accessible, and show promise for generating new biomedical discoveries. The target may or may not have experimental structures or known small molecule inhibitors available.

CACHE Challenges are centrally organised and use similar experimental assays, the homogeneity of which has been shown to be important in producing reliable datasets.⁸² Central administration also provides financial benefits, with rigorous experimental validation estimated to cost approximately US \$ 25,000 and enabling participation from poorer segments of the drug discovery community via subsidies. This low cost (which includes compound purchase, quality control, equipment time, protein purification, primary biophysical assays and hit confirmation using orthogonal assays) is enabled in part by the advent of ultra-large on demand *in silico* (virtual) libraries, which not only increases the quantity and diversity of compounds available but also reduces the cost on a per molecule basis.

Upon the conclusion of a Challenge, large amounts of high-quality experimental data will be generated. Drug discovery data is typically retained privately, but CACHE releases open-access datasets to the benefit of the whole community.

CACHE's central organisation, purchasing, and testing framework assures standardisation across each Challenge, and they are poised to be a powerful benchmarking technique for the effectiveness of state-of-the-art computer-aided drug design (CADD) methodologies, while providing valuable data for subsequent discovery campaigns.

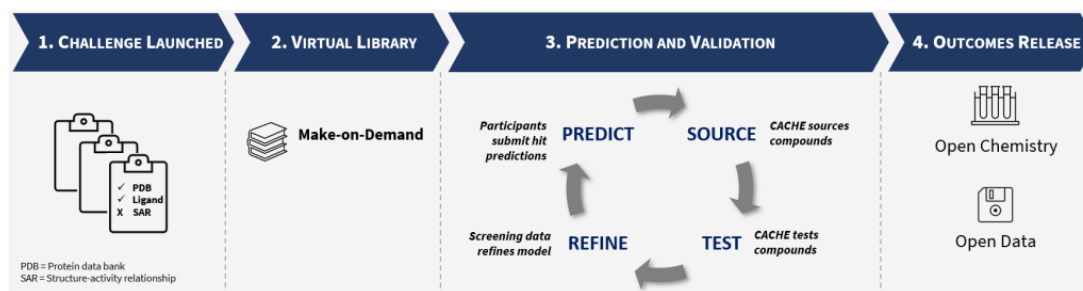


Figure 20: A schematic flowchart of the four constituent phases of a CACHE challenge. This figure is licensed under CC BY 4.0.¹²

5.1 CACHE (Critical Assessment of Computational Hit finding Experiments) Challenge 2

Over the span of two 8-week periods beginning in 2022, CACHE ran CACHE Challenge 2 (CACHE#2), a fragment-based *de novo* design campaign to discover hit compounds targeting SARS-CoV-2's helicase. A multidisciplinary team from Newcastle University, comprising myself, Ben Cree, Rachael Pirie, Mat Bieniek, Josh Horton and Daniel Cole (computational chemistry), Roly Armstrong and Kate Harris (medicinal chemistry), and Natalie Tatum (structural biology), undertook this challenge. All work discussed in this thesis is my own unless otherwise stated.

CACHE#2 comprised of two 8-week sprints consisting of a hit-finding (fragment growth) round based on various experimental data, followed by a hit expansion round.

During the initial fragment growth round of CACHE#2, X-ray crystal structures of fragments in complex with the target were provided. Any hits from the initial selection were returned with a binding affinity determined via surface plasmon resonance (SPR) assay (no crystal structures were provided). These were then used to inform the second, hit expansion round. All experimental assays were run by the CACHE organisation.

For the first, hit-finding portion of the challenge (weeks 1 - 8), the Newcastle University team grew compound molecular designs from hand-picked cores and R groups (completed by Ben Cree and Dr Daniel Cole, henceforth Cree and Cole). Regular meetings with computational and medicinal colleagues were carried out to agree which molecules would be selected to be submitted for similarity searches of the Enamine REAL database. SMILES strings returned from the database were embedded and docked, with the best docked molecules being manually curated. The team then submitted the chosen SMILE strings to CACHE for purchasing and assaying, along with two compounds that were synthesised in-house by Roly Armstrong and Kate Harris (henceforth, Armstrong and Harris).

During the second, hit expansion round (9 - 16 weeks), the team assessed molecules grown during the hit-finding round, and similar molecules (those using the same core but

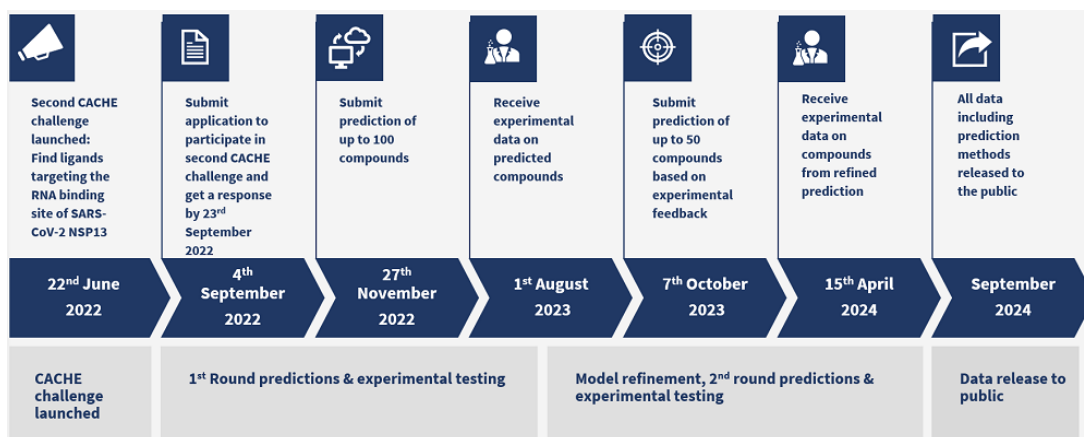


Figure 21: Timeline of this CACHE challenge.

different functional groups, selected using chemical intuition) were grown using FEgrow (Cree). Free energy perturbation (FEP) simulations were then run on the molecules grown (Cree), the results of which were subsequently used to inform compound selection for the round. The group agreed on a further set of similar molecules selected from Enamine REAL, which were then docked using gnina (Cree). The compounds were ranked by their docking scores and also assessed subjectively, considering factors such as overlap with fragments and polar interactions.

The details of Newcastle University team’s process for CACHE#2 is elaborated on in the sections that follow, beginning with a discussion of the target (SARS-Cov-2), followed by an explanation of Newcastle University team’s process for Round 1, including the design strategy, docking process, and synthesis. The subsequent section covers the work in Round 2, detailing the hit data and the docking process, before the chapter is concluded.

5.2 Target Background

Coronaviruses (CoVs) are enveloped, positive-sense, single-stranded RNA viruses.³²⁹ SARS-CoV-2, a coronavirus, caused a global pandemic in 2020 that caused the deaths of millions, and will remain a relevant global health concern for years to come. The development of anti-viral compounds for such a virus is therefore, sagacious.

SARS-CoV-2 has 16 nonstructural proteins (NSPs) that are integral for various processes related to transcription and replication of the virus. One of these proteins is NSP13, a 5-domain 67 kDa helicase. This motor protein is involved in the unwinding and separation of RNA in a 5’ to 3’ direction powered by ATP hydrolysis, and its 5-domain architecture is common to other Nidovirus helicases (to which the family Coronaviridae belongs).³³⁰ During replication, NSP13 associates with other viral replication proteins forming a multi-protein complex on viral template RNA which allows it to perform its function.¹³ (Fig. 22)

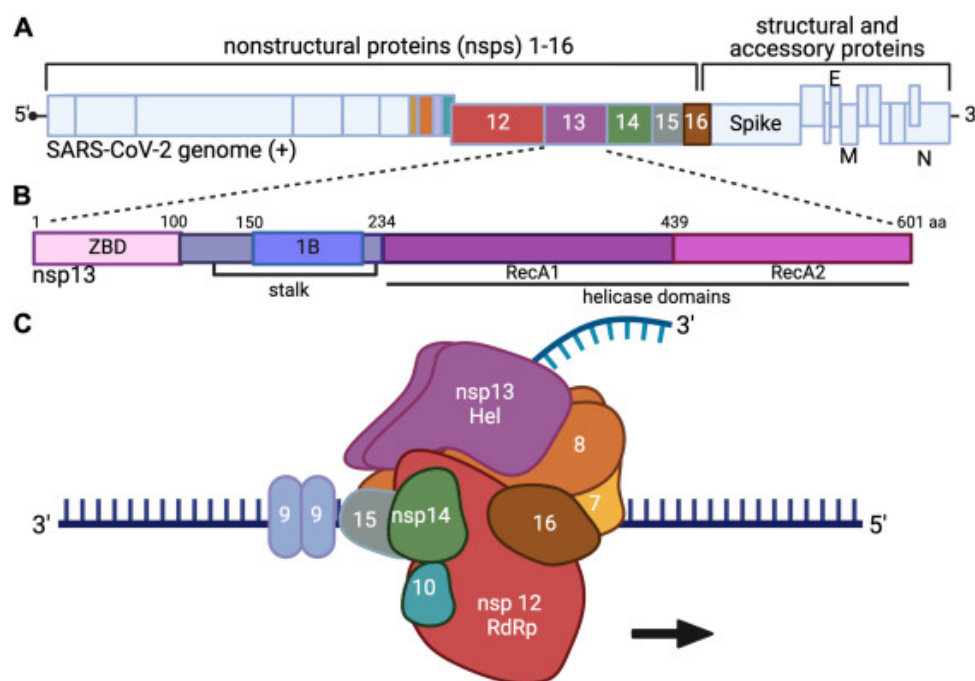


Figure 22: CoV genomic architecture and nonstructural proteins (NSPs). A. Schematic of the SARS-CoV-2 genome with the NSPs and the structural and accessory proteins. B. Schematic of SARS-CoV-2 nsp13-HEL. C. A model of probable NSPs 7–16 assembly in a multi-protein complex on viral template RNA, based on biochemical and structural studies. NSP13 helicase [HEL] shown in purple. This figure is licensed under CC BY 4.0.¹³

Inhibiting the activity of a virus' helicase can prevent access to genetic material, causing a cessation of viral replication and potentially offering an avenue to treat acute infection. Although NSP13 is known to be involved in replication, the exact role this NSP performs in the virus in general is not well understood, with the enzyme having been shown to be involved in a broad range of activities, not just the canonical ATPase and unwinding functions. The investigation of these hitherto undiscovered functions is a fertile area of research that is just beginning to be explored.¹³

NSP13 has been found to exhibit two distinct conformations ('open' and 'closed') which represent distinct states in its catalytic cycle.³³¹ The inhibitory activity of the fragments found to bind in allosteric sites is hypothesised to be due to locking the protein into one part of its catalytic cycle, preventing conformational change that is integral to its function from occurring.¹⁵ A particularly attractive feature of NSP13 for antiviral development is its remarkably conserved nature, being the most conserved NSP in the SARS-CoV-2 genome.¹⁵ Targeting a less conserved protein would be undesirable due to their propensity to mutate, and therefore potential to develop resistance to any inhibitor that may be discovered.³³²

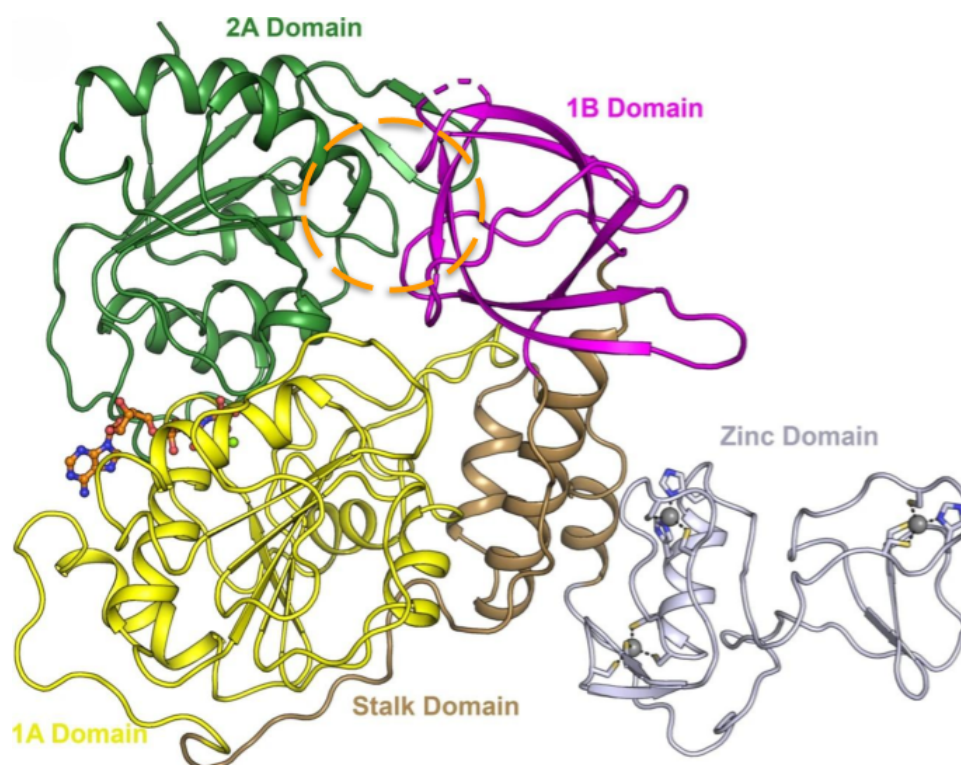


Figure 23: Structure overview of NSP13 (PDB: 6ZSL) with domains labeled and coloured individually, the interface of the 2A and 1B domains (circled in orange) form the binding site for this challenge. Figure is licensed under CC BY 4.0.¹⁴

There are multiple binding sites of interest,³³³ but here the focus is on the RNA binding site (Fig. 23). There is also an ATP binding site (not targeted) which powers the catalytic action, shown in sticks). This is because the residue makeup of the binding site is a particularly conserved portion of the sequence, with one study identifying 87% of the residues being identical across 27 sequences analysed in the Uniprot³³⁴ database.¹⁵ Additionally 100% of the side-chains that were forming direct interactions with two of the initial X-ray crystal structure fragments that were used to inform design were found to be conserved.¹⁵ These factors, along with a favourable druggability score³³⁵, has made NSP13 a target of interest for SARS coronaviruses.³³⁶

Part of the rationale for the choice of target for CACHE#2 was the fact that no inhibitors were known, but at the conclusion of the challenge, this was no longer the case, and there are now multiple examples of hit compounds.³³⁷ For example, a recent screen of a diverse library of approximately 650,000 small drug-like compounds³³⁸ yielded a confirmed hit-rate of 0.14% (881) for the initial screen, which were subsequently validated by titration assay. No structural or assay data other than the initial fragment screen were used to inform the design of inhibitors.³³⁹

5.3 Round 1: Hit and Free Energy Calculations

5.3.1 Design Strategy

The basis for our design approach was fragment growing executed using our open-source FEgrow software package, using crystallographic fragments that occupy the binding site shown in Fig. 24.

These fragments were used as both a starting core for growth and also for their pharmacophoric information, with R groups recapitulating interactions seen in experimental structures prioritised in the final selection.

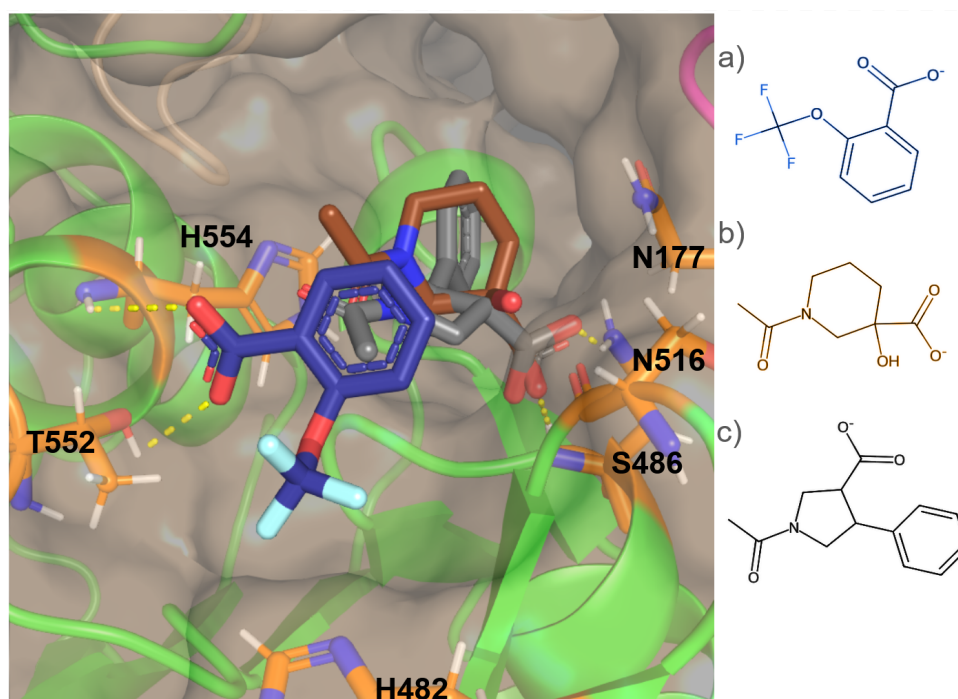


Figure 24: All fragments crystallographic fragments used in the design process; a) 5RMM (blue), b) 5RLH (brown), c) 5RLZ (grey) overlaid, with interacting residues identified (orange).

During Round 1 of CACHE#2, weekly design meetings were held, during which a few dozen *de novo* structures were agreed upon. Added R groups possessed between 2 and 15 atoms.

Structural waters within 5 Å of the initial fragments were collated and combined from 52 NSP13 structures from Newman *et al.*. All of the water molecules from each structure were extracted, producing a 'water map'.¹⁵ This map showed both areas of frequent hydration and predicted polar contacts seen in the X-ray crystal structures. During the design process, it was leveraged as a tool to suggest sub-pockets and/or residues that were not exploited by the original fragments, which could then be incorporated into *de novo* designs.

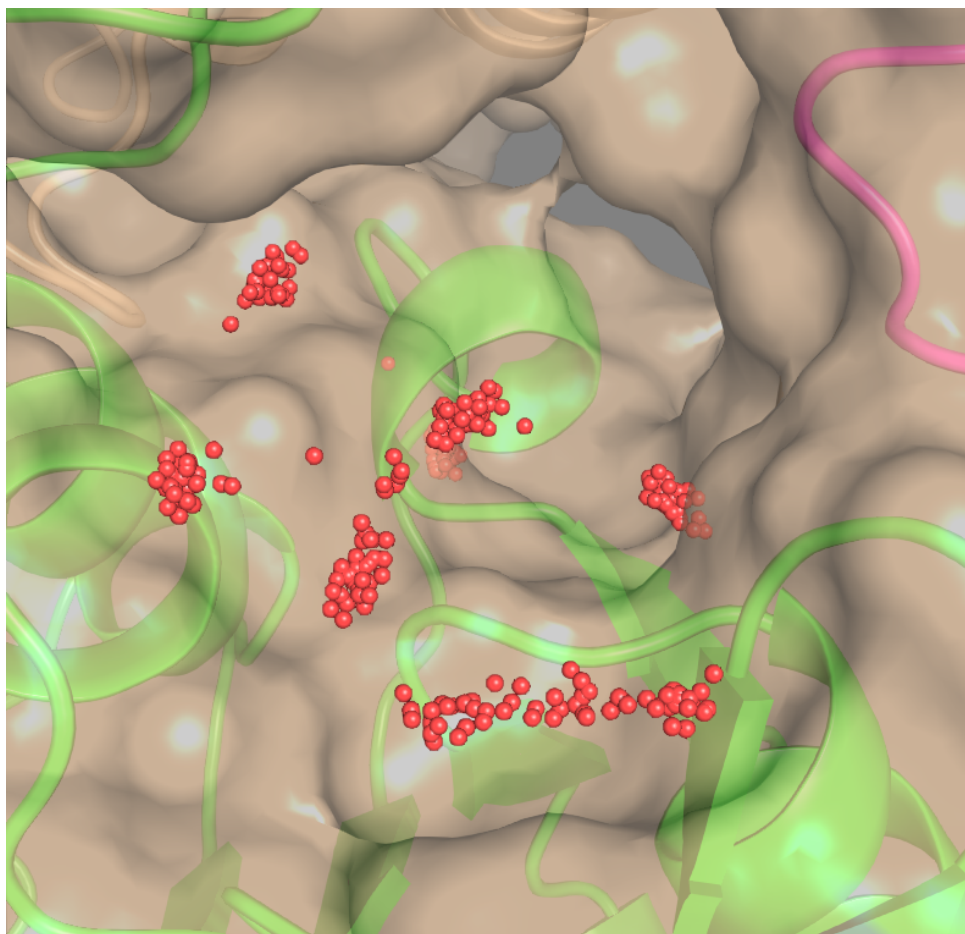


Figure 25: Water map constructed from 52 Protein Data Bank (PDB) entries of NSP13. Structures from Newman *et al.* that had waters in the RNA binding site were collated and overlayed to give areas of the pocket that were frequently hydrated.¹⁵

The topography of the pocket was defined relative to our initial *de novo* design strategy, where 5-membered rings were grown off of a carboxylate core (interacting with N516) common to two crystallographic fragments (5RLH, 5RLZ (Fig.26)), in the same fashion as the ring seen in Fig. 24 c. This approach gave three promising growth vectors in the R_1 , R_2 and R_3 positions (Fig. 27) which facilitated access to three distinct areas of the binding site.

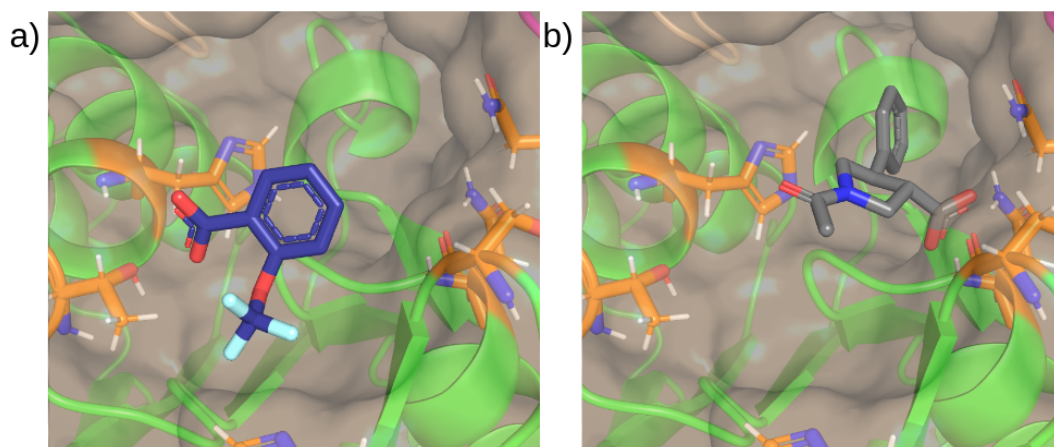


Figure 26: Crystallographic fragments: a) 5RLH (blue), with the trifluoro group occupying the R_3 sub-pocket, b) 5RMM (grey), with the phenyl group occupying the R_1 sub-pocket. Key residues shown in orange.

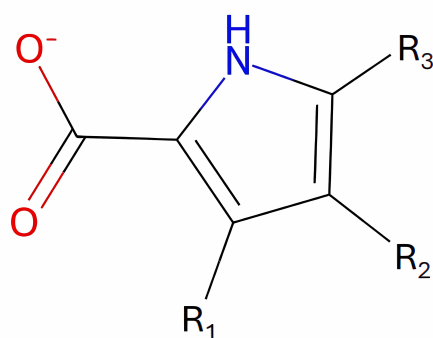


Figure 27: An exemplar 5-membered ring added to the carboxylate core from 5RLH with FEGrow showing three main vectors used for growth, similar to 5RMM.

Our overall strategy was to grow a range of linkers and functional groups off of each vector (and combine independent vectors), as well as investigate different 5-membered rings and carboxylic acid bioisosteres, looking at predicted affinities, polar interactions, and corroboration with the water map. The best scoring example (Fig. 28) possessed similar poses and groups to that of the crystallographic fragments in both the R_1 and R_3 sub-pockets, and was scored by gnina to be $1.4 \mu\text{M}$.

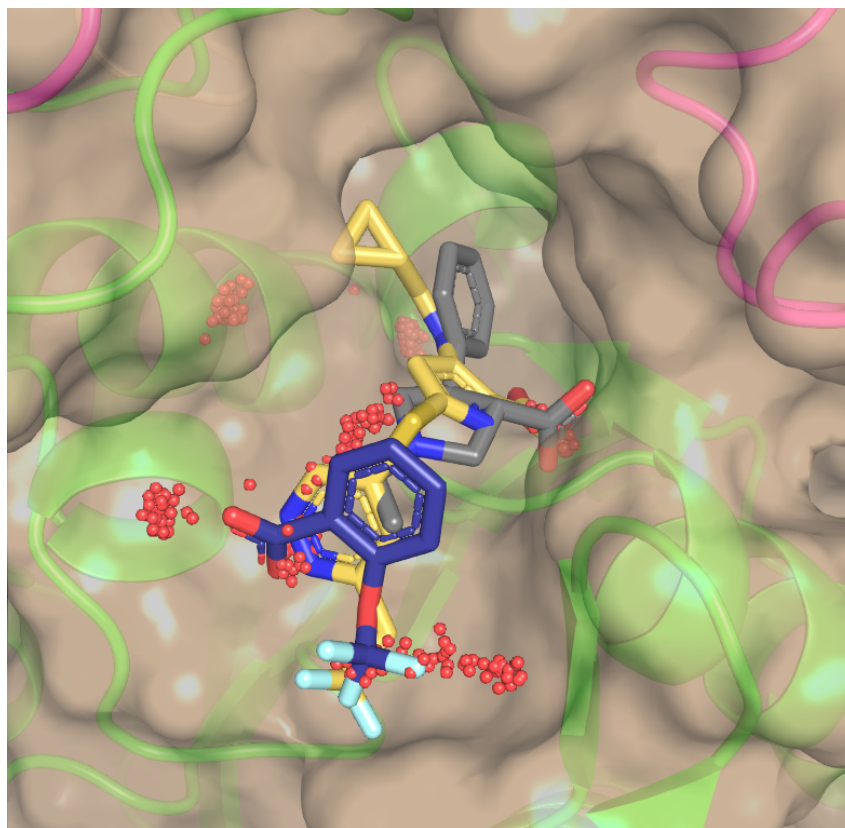


Figure 28: Best scoring *de novo* design (yellow), with a predicted gnina affinity of 1.4 μM , with crystal fragments 5RLH (blue) and 5RMM (grey), along with the water map.

5.3.2 Compound Designs

The first iteration of the design process was focused on choosing a similar ring to that of 5RMM to add via FEgrow, to act as a hub for further expansion.

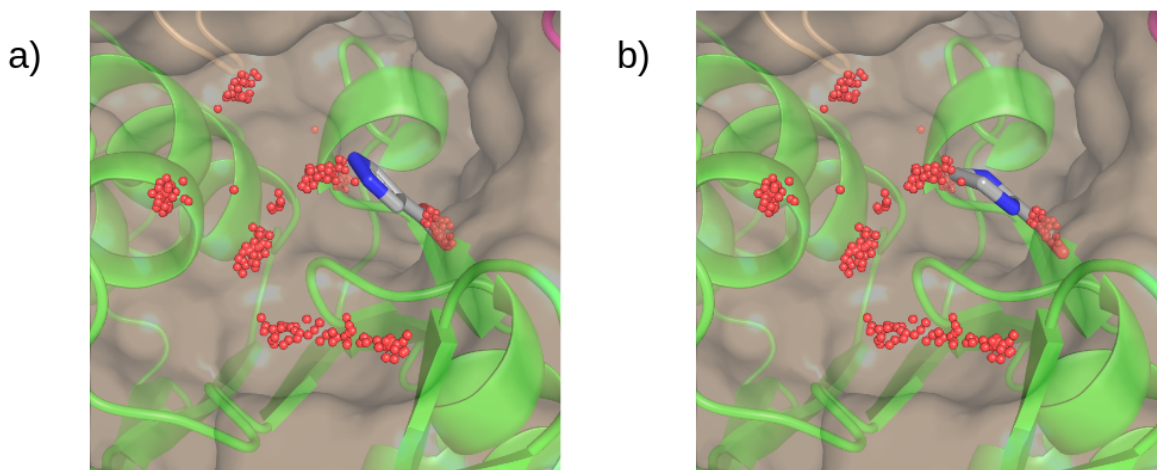


Figure 29: Exemplar 5-membered rings added to the carboxylate fragment (of 5RLZ) with FEgrow: a) pyrazole, b) imidazoline

An initial set of 5-membered rings was chosen for scoring and, from the various rings tested, those that achieved both of the previously outlined criteria (growth vectors in both R_1 and R_3 sub-pockets that were aimed at other X-ray crystal structure fragments) and had high predicted affinity were used in a further expansion round, in which the R_1 vector was chosen for growth. Growth into the R_1 pocket, which is occupied by a phenyl group from the 5RMM fragment, had the aim of recapitulating the same hydrophobic interactions towards the R_2 subpocket seen in the 5RMM crystal structure. A selection of R-groups for this purpose were chosen based on consultation with synthetic and medicinal chemists (Harris and Armstrong). Groups of a hydrophobic nature tended to score well, in agreement with the experimental structures (Fig. 30).

The R_2 vector was explored early on in the design process, however this vector consistently scored poorly. This is unsurprising, considering the distance needed to traverse the binding site (the nearest polar residue on that side of the pocket was over 10 Å away from any 5-membered ring). As a consequence, few groups, if any, formed any interactions with the receptor and as a result was not considered further in the design campaign (Fig. 31)

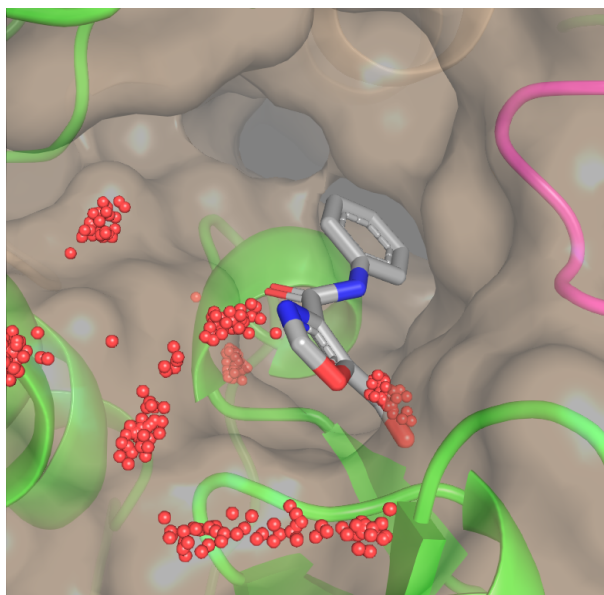


Figure 30: An example of a hydrophobic R group design off the R_1 vector that scored favourably. Oxazole ring, predicted glna affinity $95 \mu\text{M}$, amide linker, phenyl R group.

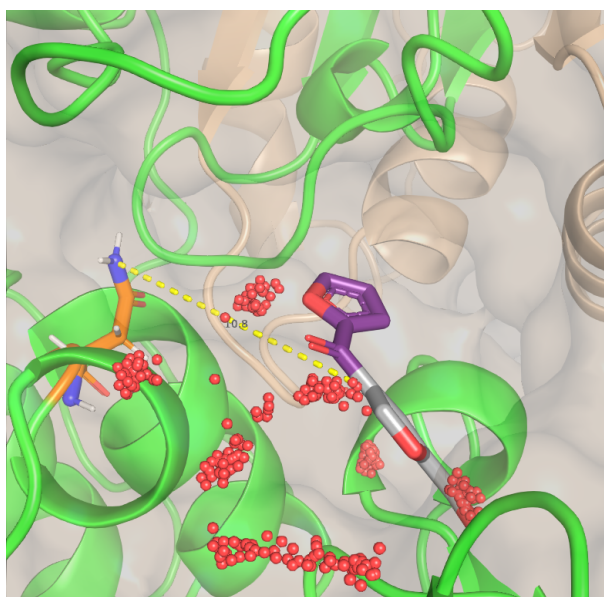


Figure 31: Furan ring. $360 \mu\text{M}$, R_2 vector, ketone linker, furan R group (purple). The R_2 vector was not well suited to R group addition due to the positioning of the core that was used, with the nearest polar residue on the other side of the pocket (orange) at a distance of 10.8 \AA .

Refocusing efforts to the R_3 vector, many groups displayed an inability to reach the residues deeper in the groove (H482, T552) to form interactions seen in the X-ray crystal structures of 5RLH and the water-map (Fig. 32), especially those that were small and contained aromatic rings and tended to point towards solvent.

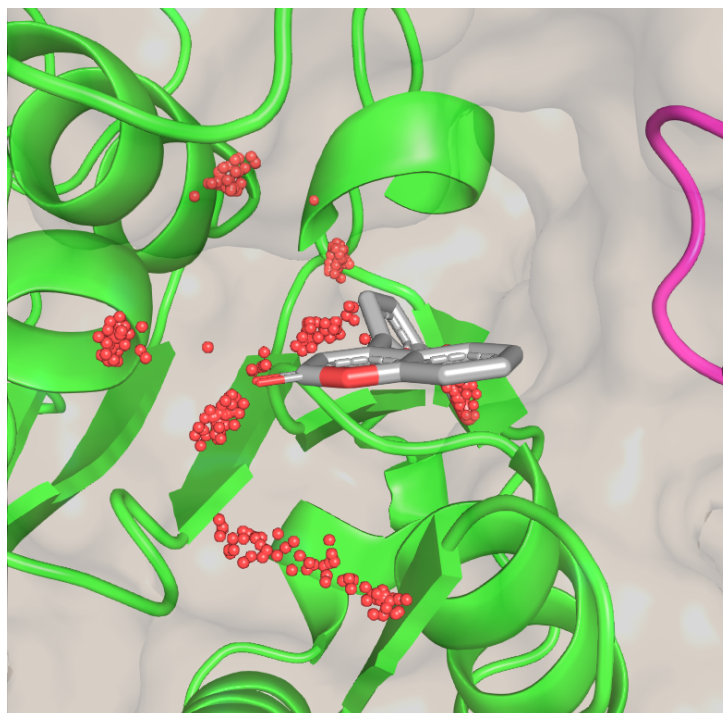


Figure 32: Furan ring. 28 μM , R_3 vector, no linker, benzopyran R-group. The lack of a linker group restricted the ability of the benzopyran to reach the groove.

The 5RLH fragment was used to inform the initial strategy, prioritising hydrophobic groups that recapitulate the positioning of the trifluoro group in the groove. To address the previously outlined issue of groups being unable to form interactions seen in 5RLH (off the R_2 vector), a test set of 6 linkers (Fig. 33) were grown from a pyridazine ring to coarsely assess the potential of, and whether any particular linkers should, be prioritised, based on predicted affinity and geometry. Larger linkers was found to overshoot the intended subpocket, so the test set was restricted to single atoms of various different elements affording different bond lengths and angles.

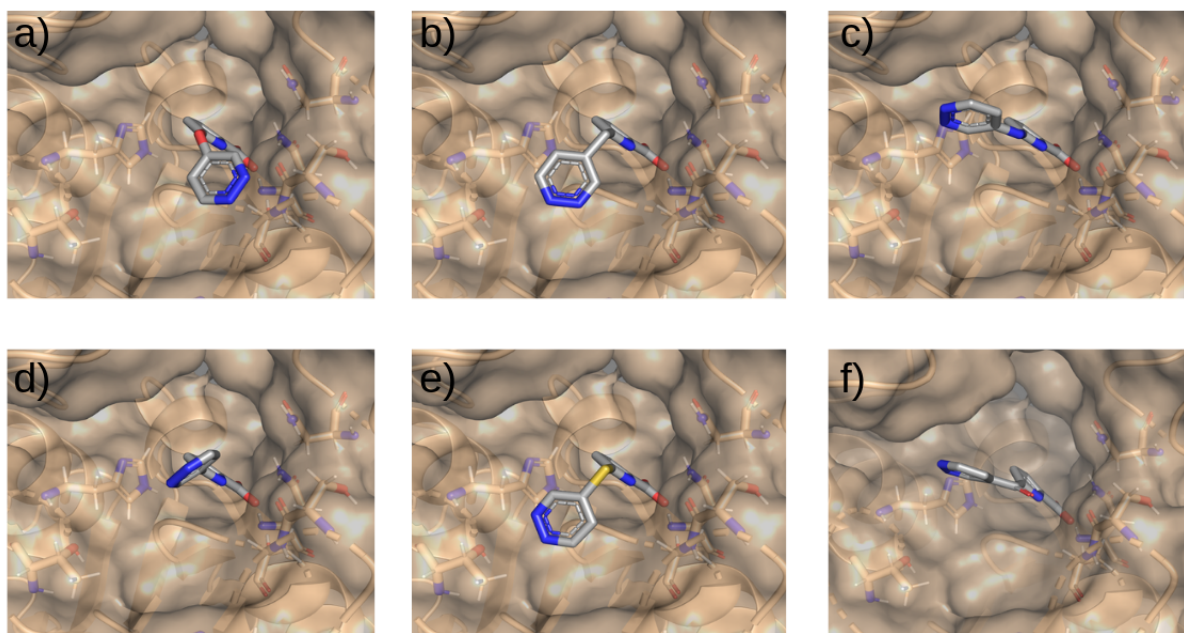


Figure 33: Test set of linker atoms from a pyrazole ring with pyridazine R-group. a) Oxygen, b) Carbon, c) Nitrogen, d) No linker, e) Sulphur and f) Ketone.

Both the amine and ketone linkers had no conformations that pointed in the direction of the protein (Fig. 33 c, f), with all others having low energy conformers pointed towards the groove, as desired. The geometry of the thio-linker most closely matched the phenyl group of 5RLH due to the geometry of the C-S-C bond angle ($\sim 90^\circ$) dictated by the lone pair repulsion, although it had a lower gnina predicted pK (Fig. 33 e). The dynamic range of the predicted affinities were not especially significant (roughly $70 \mu\text{M}$) and most conformers ended up in a similar orientation, which meant that most single atom linkers fulfilled the desired criteria, perhaps with a slight preference for Carbon/Nitrogen/Sulphur, which were prioritised for further growth into the R_3 sub-pocket (Fig. 34)

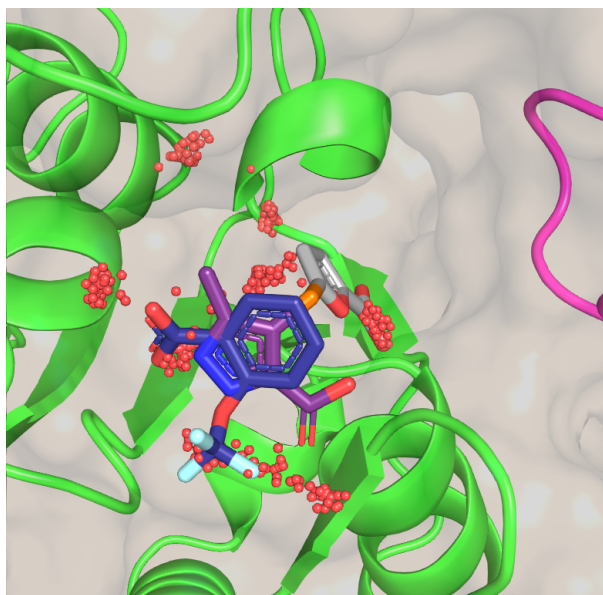


Figure 34: An example of a *de novo* design utilising the C linker (orange), overlaid with 5RLH (blue). Predicted gnina affinity, 43 μ M

Further optimisation (via trial and error and with variations of previously identified R groups) led to these 3 vectors being expanded simultaneously, yielding some high scoring (but very large and unwieldy) molecules. This reached the limit of design, with molecules' molecular weights approaching 600+ Da (Fig. 35), and so more efficient use of molecular weight was needed. This is due to the fact that the larger the molecules become, the less reasonable the assumption of an identical binding mode becomes. Similar designs of the same vectors were tested, prioritising smaller R-groups and filtering molecules above a maximum molecular weight.

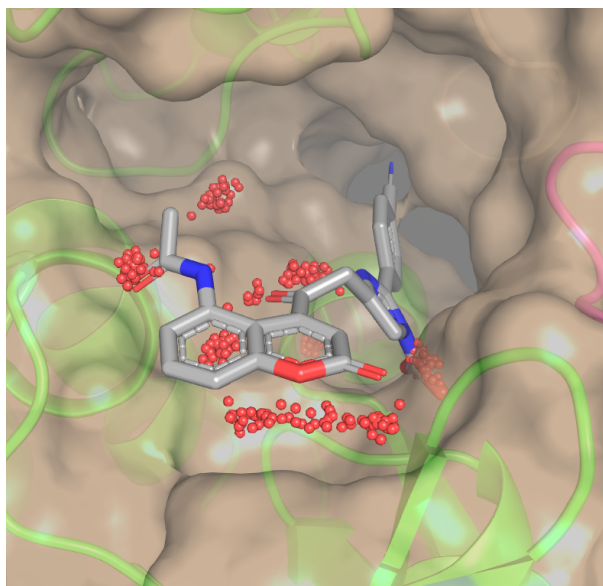


Figure 35: A large molecule grown with FEgrow, incorporating all three vectors with a predicted gnina affinity of $1.2 \mu\text{M}$.

Fig. 35 shows a compound which has both the R_1 and R_3 vectors expanded. The benzonitrile group occupies the same position as 5RMM and the benzopyran group had two rounds of expansion, with a methyl amide group being added aiming towards H554.

5.3.3 Docked Enamine Compounds

Designs that scored well were brought to the weekly team meetings, during which molecules were manually selected to be submitted for similarity searches in the Enamine REAL database. A selection of top Enamine compounds, retrieved from these ECFP4 fingerprint similarity searches are illustrated below. Retrieved compounds were docked with gnina.

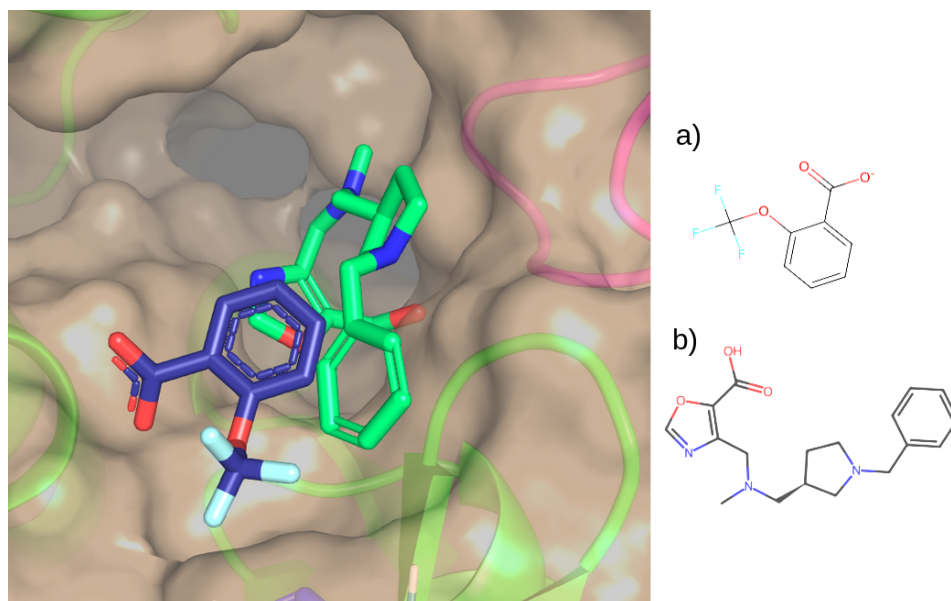


Figure 36: a) Example docked enamine structure (green), b) 5RLH (blue), with a predicted gnina affinity of $17 \mu\text{M}$.

Fig. 36 shows the general shape was analogous to that of a combination of 5RLH and 5RMM; its oxazole ring occupies the same position as the rings seen in the query molecules, albeit in a slightly twisted fashion. The phenyl group of the molecule is positioned in the groove towards N177, similar to that of 5RLH.

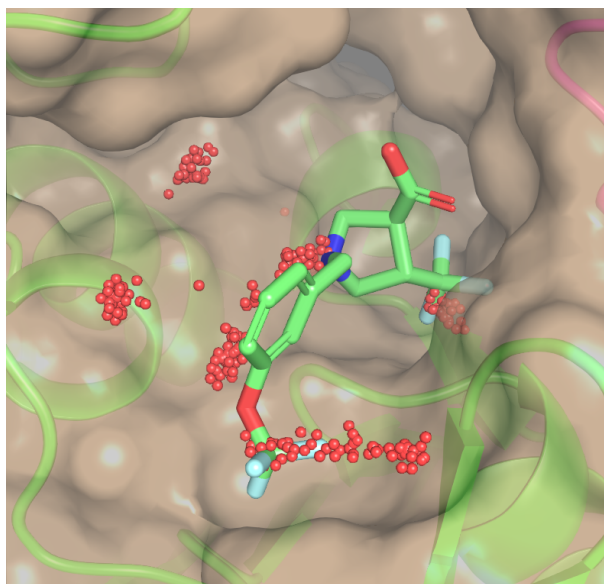


Figure 37: A top-ranked Enamine compound, with water map shown in red. gnina predicted affinity $8.7 \mu\text{M}$

Fig. 37 exhibited a carboxylate group forming identical interactions to the crystal structures, as well as a trifluoromethoxybenzene in the R_3 sub-pocket in addition to

exhibiting significant overlap with the water map. Both of these moieties are connected to an azepane ring, which is similar to the piperidine ring seen in the 5RLZ fragment.

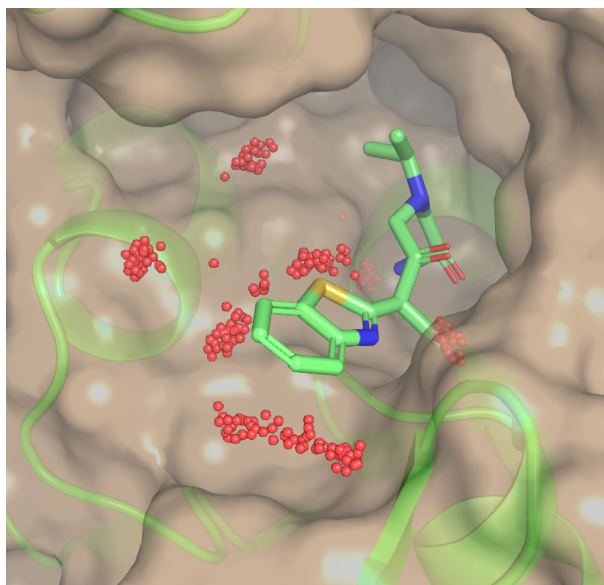


Figure 38: Top docked Enamine compound. predicted gnina affinity $0.85 \mu\text{M}$

The top-scoring docked enamine compound is shown in Fig. 38. Here, gnina recovered the binding mode of the carboxylate anion with a cyano isostere (interacting with N516). The general shape of the ligand additionally matches that of the combined crystallographic fragments (Figure 24), which has been shown to be important for binding.³⁴⁰

5.3.4 Synthesis

Due to the fact that exact matches for the similarity searches for a given designed query molecule were sparingly found, 7 compounds (Fig. 39) were considered for synthesis in-house to better test predicted designs. All compounds incorporated the trifluoromethoxybenzene of 5RLH and carboxylate of 5RMM, with pyrazole and pyrrole rings used as a hub from which groups were grown towards other subpockets. Compound **1** is based on a simple merging of 5RMM and 5RLH, and compound **2** is for probing the necessity of a ring in the R_1 sub-pocket. Compound **3** and **4** were testing the effect of various halogens and compounds **5** - **7** were intended to investigate alternative rings of compound **1**.

Due to budget and time constraints of the challenge, only compounds **1**, **4**, **5**, **6** were made by Armstrong and Harris.

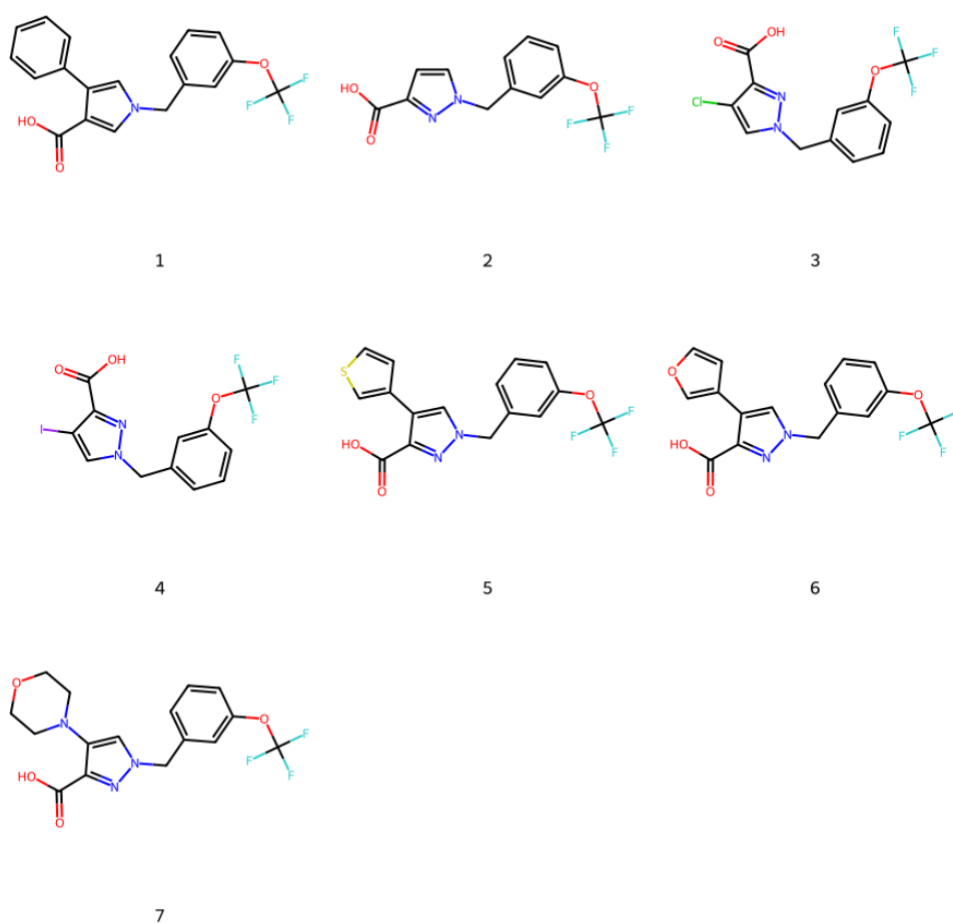


Figure 39: Compound designs of custom molecules to be synthesised in-house.

5.3.5 Round 1 Results

A list of 192 Enamine compounds were compiled by visually inspecting the docked poses from similarity searches of selected grown molecules. ECFP4 fingerprints were generated from the chosen molecules and used to search Enamine REAL. The top 10 compounds were returned for each SMILES string. Compounds were then ranked by gina predicted pK and independently scored by eye (complex were visualised in PyMOL). An ordinal 1-5 scale was used, with 5 indicating a compound should be moved higher, 3 kept where it is, and 1 moved down.

After manual independent ranking, compounds that were identified as needing repositioning were inspected and moved if judged to be necessary. After this final sifting process, a list of 150 compounds was submitted for experimental assay. This was later filtered down to ~100 compounds, based on compound availability and price. This final list of 100 compounds was submitted for experimental testing by CACHE.

All compounds were screened at 50 μ M against NSP13 using an inhibition assay by CACHE, and those that showed a dose-response between 0.5 and 2 RU (response unit

over maximum expected signal) were selected for dose-response testing via SPR. Primary hit compounds were tested for aggregation and solubility (by dynamic light scattering, up to 200 μM), and for binding to NSP13 using an SPR assay. ATPase inhibition was also performed but little correlation was seen between the assays, so only SPR was used to adjudicate whether a hit should progress due to enzymatic assays being prone to false positives.³⁴¹ Of the 100 submitted compounds, 19 fell within the range for follow-up experimental validation. Compounds where NSP13 K_D (μM) < 150 and NSP13 % Binding < 150 and NSP13 % Binding > 30 were advanced to Round 2. Using this criteria, a single compound was confirmed via dose-response assay from the initial selection of 100 compounds and advanced to the second round, (CACHE_1438.39, Fig. 40). In total, 46 compounds selected by 18 participants progressed.

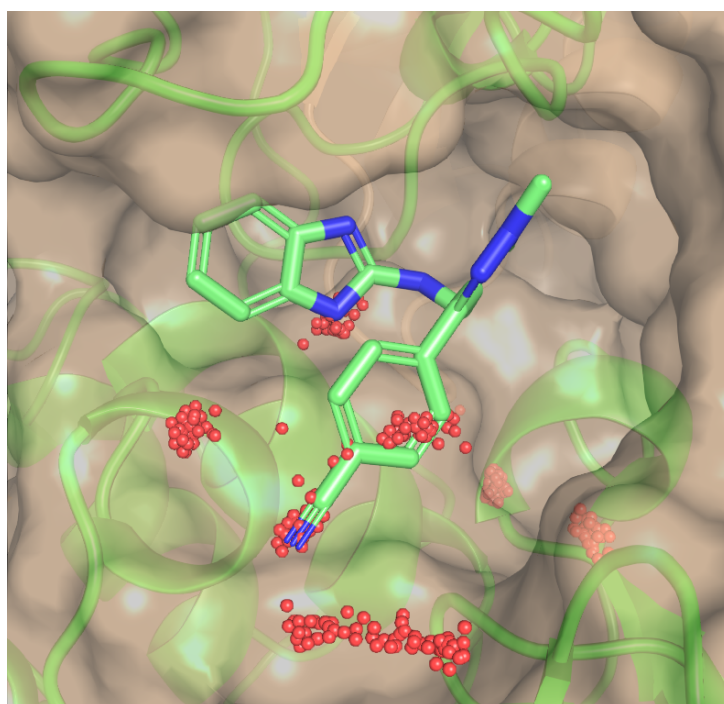


Figure 40: Original 9 μM hit structure as predicted with gmina, and watermap overlayed.

The binding affinity of the hit compound (CACHE_1438.39, Fig. 41) for NSP13 was measured by SPR in a dose-response experiment to be 9 μM . It was ranked 18th in the 100 compounds submitted.

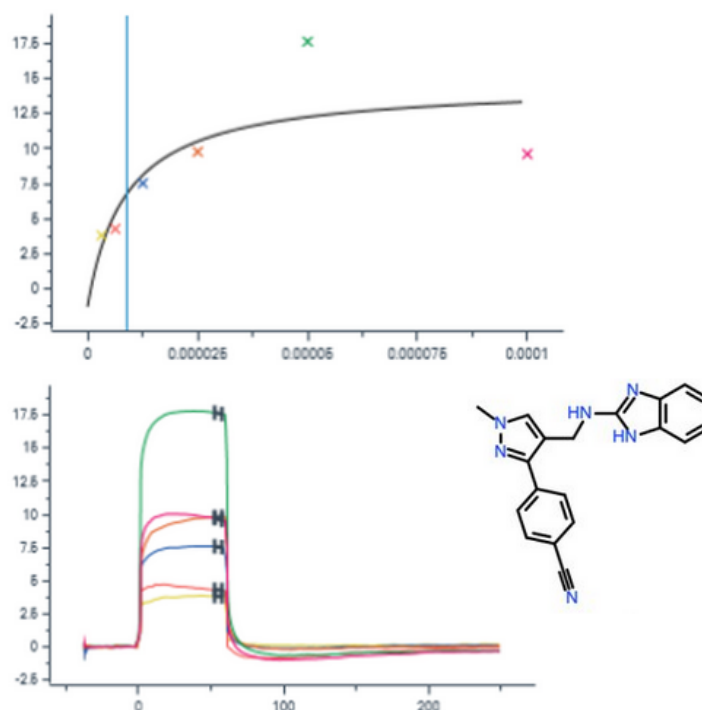


Figure 41: Steady-state SPR affinity data for the round 1 hit compound, 9 μM .

5.4 Round 2: Hit Expansion

A total of 45 analogs of CACHE_1438_39 were submitted for Round 2. Of these, 4 compounds (see section 5.4.1, not including re-supplied parent molecule) showed a binding response by SPR with a tendency to reach saturation. Compounds were both tested for selectivity to WDR5 and for ATPase activity by CACHE, but as in Round 1, ATPase activity did not correlate with binding data and it is unclear if compounds binding in a specific cavity in the RNA site would inhibit activity, hypothesised to be due to the complex structural dynamics of NSP13.³⁴² One hit compound (Fig. 43 c, 87 μM) was considered confirmed, as it was orthogonally validated by ^{19}F NMR.

5.4.1 Round 2 Hit Data

In the rankings for the primary metric of the challenge (hit quality), our team ranked 9th out of 23. This was a combined normalised sum of all scores from members of a hit evaluation committee based on biophysical properties as well as a subjective evaluation by expert medicinal chemists. Biophysical activity was focused on qualities such as number of actives, number of analogues, SPR confirmation and number of orthogonal assays; whilst expert medicinal chemistry evaluation criteria was focused on visible SAR, structure (subjective attractiveness, PAINS, reactive moieties), synthesis, as well as physical chemistry parameters like solubility, permeability and lipophilicity. Ultimately, at the end of the second round, 5 hits were found, 1 parent and 4 actives with 42

inactives. The main deficit seen in the hits is that there is no obvious SAR (especially so, considering that only a single confirmed hit was found). The one confirmed hit (by ^{19}F NMR) was ranked 5th out of a total of 40 molecules. A large proportion of the score attributed to this compound was due to the hit being especially well validated, as well as possessing desired drug-like qualities, in addition to reasonable low micromolar affinity (structural confirmation of the binding mode is being pursued by the organisers).

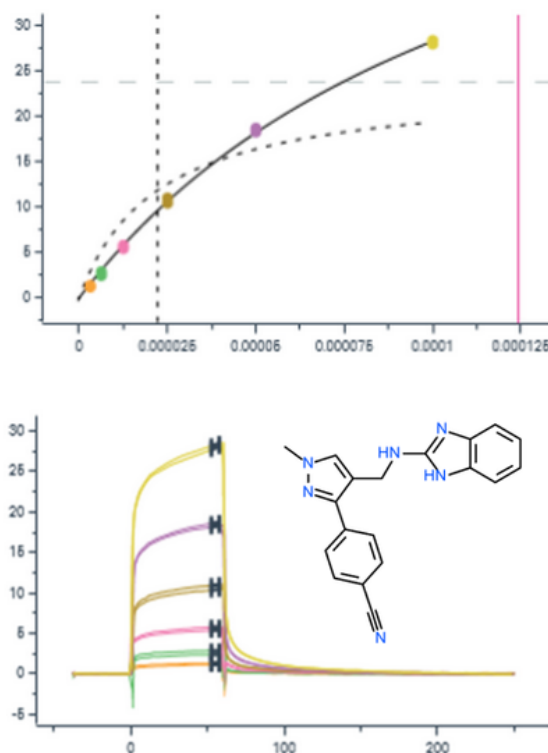


Figure 42: SPR steady-state affinity measurement for the hit from round 1, 124 μM .

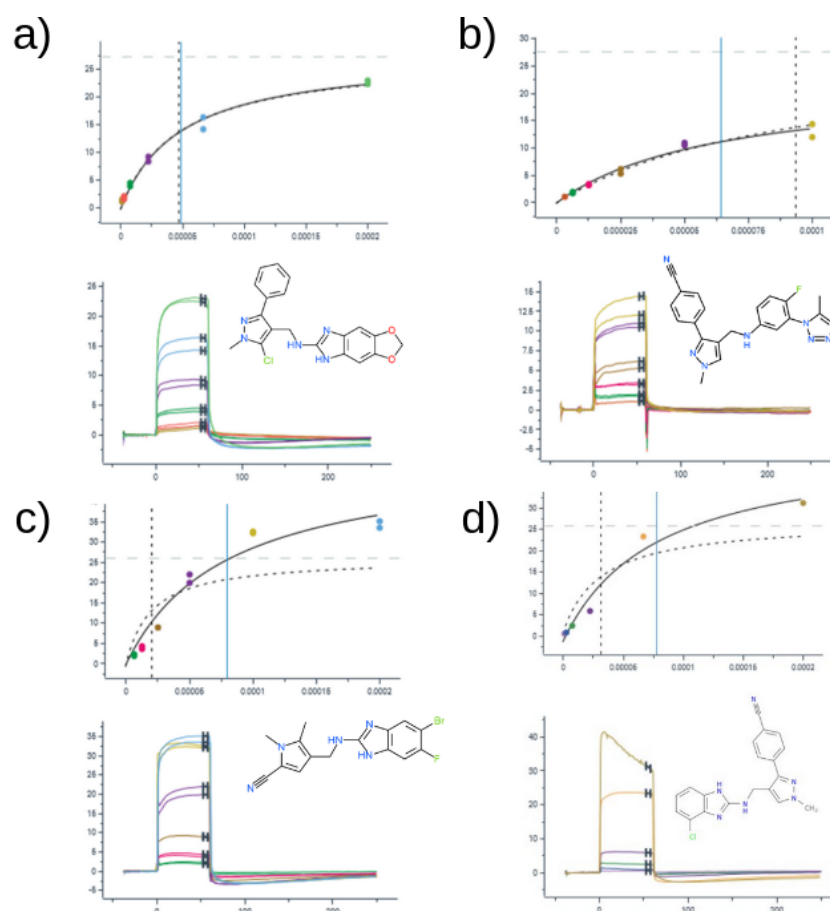


Figure 43: SPR steady-state affinity measurements for all round 2 hits: a) 49 μM , b) 64 μM , c) 87 μM (orthogonally confirmed via ^{19}F NMR), d) 78 μM

5.4.2 Free Energy Perturbation (FEP) Calculations

Using the experimental results (a single hit) from the first round, a second round of hit expansion was carried out utilising free energy calculations. A substructure search of Enamine REAL was carried out to produce a list of compounds with substructure overlap that could be templated and modified via a SMILES template interface for FEgrow. When given a SMILES string of a core (substructure overlap), this interface grows groups off any vector by treating the maximum common substructure as the core, and groups not common to both molecules as R-groups to be added. Initial binding mode was obtained from docking the original hit with gnina, and the free energy protocol used was identical to that of Chapter 4. A manually curated set of 7 grown structures based on promising functionalisations suggested by the substructure based growing were taken through to free energy calculations for accurate physics-based scoring. The network of alchemical transformations was configured in a star configuration and is shown in Fig. 44.

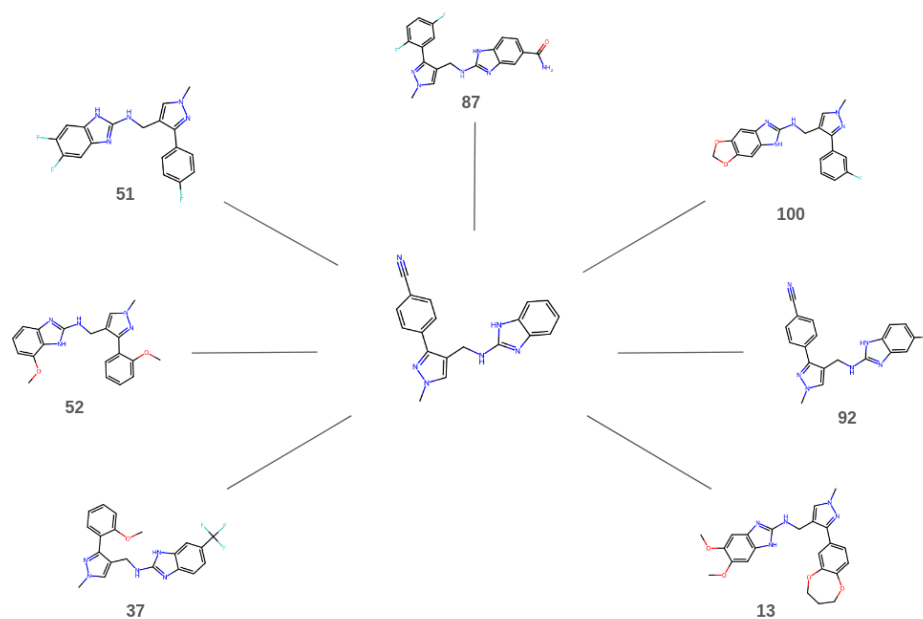


Figure 44: FEP transformation network for investigating follow-up compounds to the Round 1 hit.

In general, the first set of free energy calculations did not have good agreement between the forward and backwards trajectories, possibly due to poor λ -state mixing, caused by a lack of configuration sampling.²³⁹ However, compound 100 showed good agreement between the forward-reverse runs and a strong improvement in affinity, relative to the original hit. As a result, this was the only compound derived from free energy calculations that was used in a similarity search of the EnamineREAL database. A second set of free energy calculations was planned, but omitted due to time constraints.

Compound	$\Delta\Delta G$ (Forward) / kcal/mol	$\Delta\Delta G$ (Backward) / kcal/mol
13	-0.14 ± 0.015	0.59 ± 0.0032
37	-0.45 ± 0.00015	-0.20 ± 0.016
51	-0.085 ± 0.0022	0.31 ± 0.0043
54	-0.70 ± 0.0074	-0.034 ± 0.0062
87	1.6 ± 0.31	0.26 ± 0.073
92	-0.68 ± 0.00096	0.27 ± 0.0060
100	-0.80 ± 0.0022	0.84 ± 0.012

Table 2: Forward and Backward SOMD Free Energy Differences (MBAR) in kcal/mol. Errors are MBAR estimates.

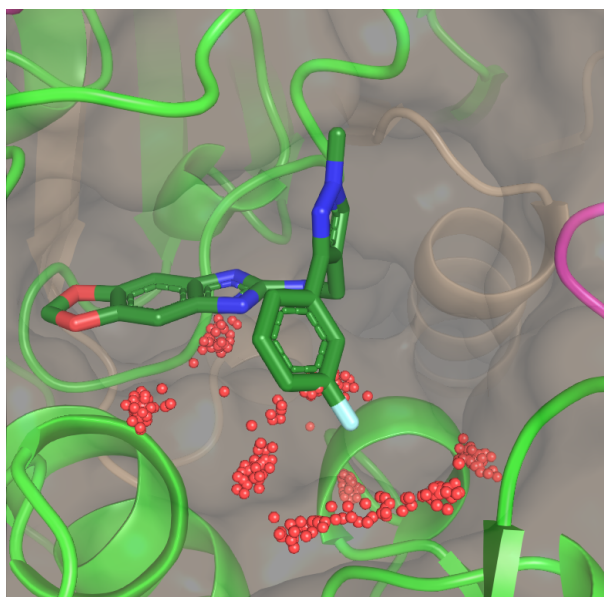


Figure 45: Compound 100, predicted ΔG to be -0.8 kcal/mol relative to the hit compound.

5.4.3 Docking

As a secondary evaluation metric for the challenge, the aggregated compounds selected (for round 1) from all participants were classed as either active or inactive by each participating team. This metric represents the ability of a protocol to judge a given molecule’s potential activity. A merged selection of 2200 compounds from all participants was docked using gnina and ranked in order to assess our ability to predict hit compounds from those submitted and as a post-challenge validation of the gnina docking protocol. 3D conformers were generated from SMILES strings using RDKit ETKDG conformer generation.

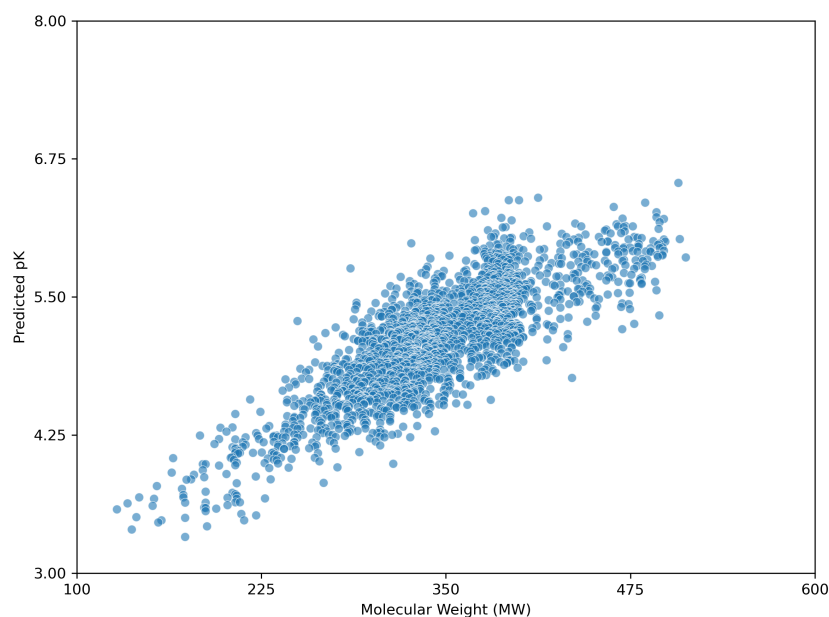


Figure 46: Scatter plot of 2200 compounds submitted from all participants, coloured by number of molecules. Predicted pK determined by gnina.

Most compounds submitted had a molecular weight of between 300 and 400 Da, and predicted affinity increases linearly with molecular weight. Our protocol was the highest scoring among all participants (Fig. 47), demonstrating the efficacy of the scoring function used in FEgrow. Fig. 46 demonstrates the relative lack of dynamic range for predicted compounds, with virtually all compounds lying in the range 4.5 - 6.0. Despite this, gnina was able to successfully (compared to other teams) predict active compounds by ranking docked structures by affinity.

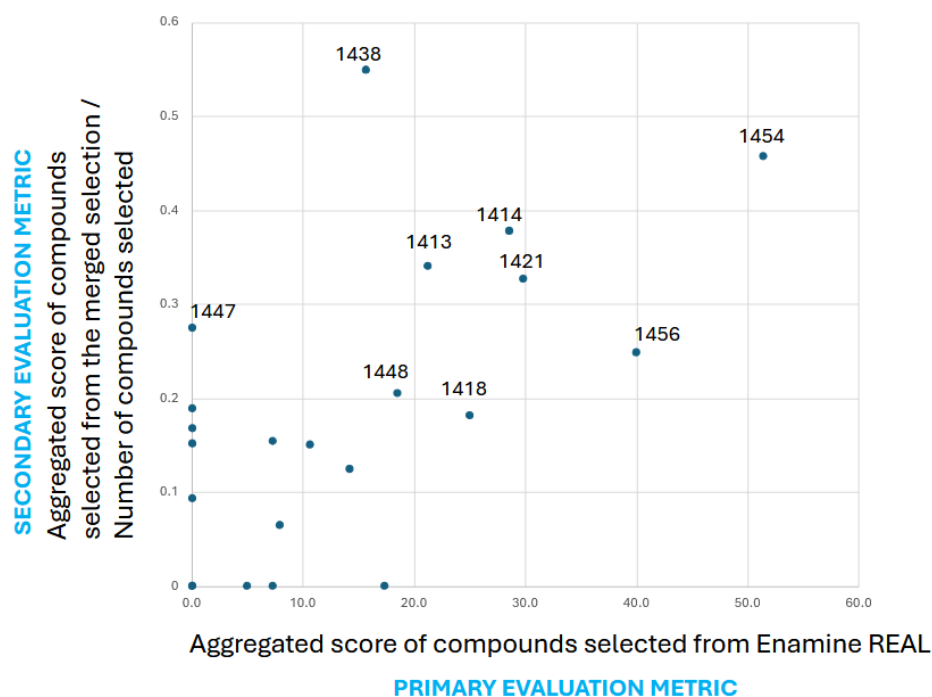


Figure 47: Comparison of performance according to the two evaluation metrics. This work is 1438. Y axis represents the ability of a protocol to accurately predict active compounds for NSP13. X axis is the combined score for all hit compounds, for each team.

5.5 Conclusion

Our combined and normalised sum of all scores from members of the hit evaluation committee, based on biophysical properties as well as medicinal chemistry evaluation, placed us 9th out of 23 participants. This was based on a single confirmed hit, which scored relatively highly - being the 5th best molecule submitted overall.

Although successful in finding hit compounds, there are multiple areas that can be improved in our protocol. Firstly, the manual growth of FEgrow was a significant bottleneck for design throughput, due to both the time required to select R-groups as well as adding them. Free energy calculations while useful, took a significant amount of time relative to the length of the campaign, making them prohibitively expensive to utilise extensively in such a short design timeline. To better leverage these rigorous calculations, more efficient methods of simulation are needed.³⁴³ All compounds submitted had a predicted affinity (via gnina’s CNN scoring function) of micromolar or better, so there was a relative dearth of hits compared to our prediction. This suggests that the scoring function is more suited to judging relative affinities, rather than absolute (as seen in Fig. 46’s minimal dynamic range and effectiveness of the scoring function for hit classification).

One of the biggest obstacles was the fact that any compounds that were designed could not be directly purchased, and instead similar compounds (as defined by Tanimoto similarity of Morgan fingerprints) had to be retrieved from a large on-demand library, which limited the efficacy of the design process. A large proportion of the designs that were chosen for similarity search had only a poor resemblance to those that were returned, and rarely contained the exact moieties that were predicted to form key interactions.

As a result of these issues, subsequent improvements were made to FEgrow to enable active learning-driven compound design, with custom scoring functions and automated seeding of molecules available from on-demand chemical libraries which obviates the need for selecting unevaluated molecules that dilute the power of the *de novo* process and are merely superficially similar to a designed query molecule (see chapter 6).

6 Active Learning

This chapter is a reproduction of a preprint on ChemRxiv.³⁴⁴ All computational work is my own except the database search (Section 6.2.2), which was performed by Dr Mateusz Bieniek. Fluorescent activity assays were performed by Siddique Amin (Section 6.3.3).

6.1 Introduction

Recent advances in structural biology, from sample preparation, to synchrotron infrastructure and data analysis pipelines, have transformed the throughput of protein-ligand complexes available to inform drug discovery campaigns³⁴⁵. When soaked with carefully designed compound libraries³⁴⁶, the numbers of small molecule (or fragment) structural hits can reach 10s or 100s against a single therapeutic target³⁴⁷. A frequently employed next step is to attempt to grow and/or link the hit compounds, using either custom synthesis³⁴⁶ or ordering from catalogues of purchasable compounds^{21,348}. However, chemical space is vast such that even choosing follow-up compounds for purchase from on-demand libraries, such as the readily accessible (REAL) Enamine database³⁴⁹ (> 5.5 bn compounds in 2022), becomes highly non-trivial³⁵⁰.

As such, attention is turning to cheminformatics and machine learning based algorithms for structure-based *de novo* hit expansion, linking and merging³⁵¹. A wide range of approaches are available to build from initial structural biology data, including DeepFrag²⁷⁹ that identifies promising fragments for addition to an input bound ligand, using a deep convolutional neural network, and DEVELOP³⁵² that combines 3D pharmacophoric constraints from the binding pocket with a graph-based deep generative model for R-group and linker design. The SILVR method enables an equivariant diffusion model to be conditioned to generate molecules based on a reference structure, such as a fragment from a crystallographic screen³⁵³. The V-SYNTHES approach makes use of on-demand libraries for hit-finding by decomposing compounds from purchasable databases into reactive scaffolds and synthons, and using the highest scoring docked fragments as seeds for further growth³⁵⁴. One particularly noteworthy example is the use of fragment merging to design hits against the nonstructural protein 3 (NSP3) of the severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2)³⁵⁵. Fragments from a crystallographic screen were merged using the Fragmenstein package³⁵⁶, ensuring placement of molecular substructures onto the original fragments, and subsequently used as templates for searching on-demand chemical space. In this way, fragments were rapidly elaborated into a 0.4 μ M hit (representing a >400-fold improvement in affinity).

While extremely promising, all of the above *de novo* design approaches suffer from some combination of the following issues: i) reliance on an approximate classical molecular mechanics force field or knowledge-based algorithm for generating and optimising binding

poses, ii) use of an approximate objective function (usually a docking score) as a surrogate measure of binding affinity, iii) approximation of a rigid target receptor structure, and iv) limited synthetic tractability of the designed compounds. We therefore developed the FEgrow software as an open-source, interactive Jupyter notebook based workflow for building user-defined congeneric series of ligands in protein binding pockets to start to address some of these open questions (Figure 48)¹⁷. FEgrow grows user-defined functional groups (R-groups) off a constrained core of a known hit compound, thus incorporating input from structural biology and the expertise of the user in selecting synthetically tractable elaborations. Since publication, we have added functionality for connecting R-groups to the core via a flexible linker, which can be chosen from a library of those common to bioactive molecules³⁵⁷. In this way, users can choose from 1M+ combinations of linker and R-group from our distributed libraries (or upload their own R-group modifications). The modular workflow allows for the incorporation of state-of-the-art molecular modelling algorithms, such as the use of hybrid machine learning / molecular mechanics potential energy functions to optimise the ligand binding pose^{227,229}, and the gnina convolutional neural network scoring function to predict the binding affinity⁶. We plan to expand the range of available optimisation algorithms and scoring functions as they become available (see Methods Section). While interactive work is useful for small-scale studies, we have found it useful to automate the workflow for use on high performance computing (HPC) clusters, and since publication have added an application programming interface (API) to FEgrow (Figure 48B). This enables us to build virtual libraries with a common core, for example, using reaction-based generative scaffold decoration with LibInvent³⁵⁸ or substructure searching of compound libraries³⁵⁹, and then rapidly build the compounds into the protein binding pocket with FEgrow. However, unless the libraries are designed using information from the binding pocket, time is wasted building and scoring compounds that are unlikely to be beneficial and it is still not feasible to routinely scan all possibilities. Hence, rather than exhaustive or random searches of chemical space, we investigate here the use of active learning to elaborate compound design with FEgrow. The general idea behind this approach is that a subset of compounds is evaluated using an expensive design objective function (in this case the molecular growing and scoring algorithms in FEgrow) and used to train a machine learning model (Figure 48)²⁶⁵. The machine learning model then predicts the objective function for the remainder of the chemical space, and the next subset of molecules is picked for evaluation (for example, in order to optimise the objective or further explore the chemical space). By cycling through this procedure, the algorithm can iteratively make up for any lack of diversity in the initial training subset, and it has been found previously that the most promising compounds can be identified by evaluating only a fraction of the total chemical space.

Several studies have investigated the effects of choices such as machine learning

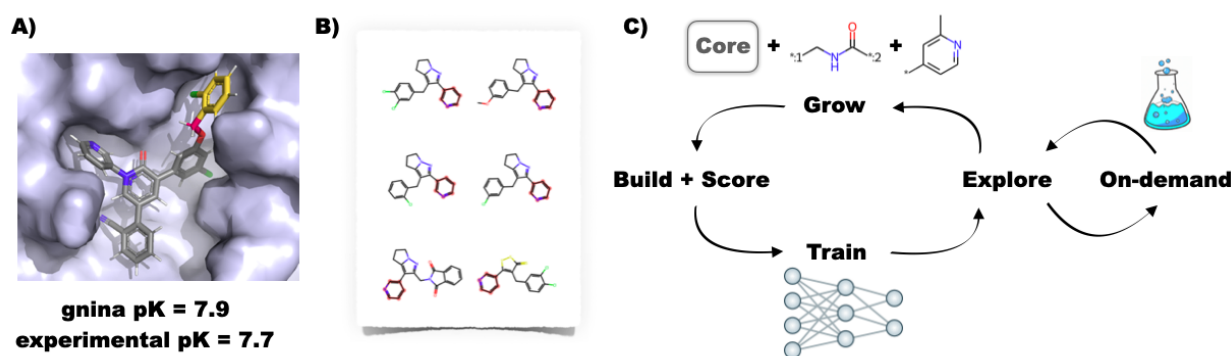


Figure 48: A) Example building and scoring of a SARS-CoV-2 inhibitor¹⁶ using the interactive FEgrow workflow¹⁷. The fixed core (grey) is extended using a user-defined, flexible linker (pink) and R-group (yellow), and scored using gnina⁶. B) Compound libraries with substructures that match the rigid core can now be automatically grown and scored, treating the rest of the molecule as fully flexible. C) Proposed active learning cycle. Compounds are grown, built in the binding pocket and scored with FEgrow. The outputs are used to train a machine learning model, which is used to select the next batch of compounds. Optionally, the chemical space can be seeded using compounds available from on-demand chemical libraries.

algorithm, sample selection protocol and total dataset size on active learning efficiency for experimental and computational affinity predictions^{360–364}. In general, active learning has been shown to increase enrichment of hits compared to either random or one-shot training of a machine learning model, at low additional cost, and to be relatively insensitive to choices of molecular representation, model hyperparameters and initial training subsets. Active learning has shown practical utility in prioritising compounds based on objective functions from docking^{365–367} or free energy calculations^{361,362,368,369}.

Here, we interface FEgrow with active learning to efficiently search the chemical space of linkers and R-groups from a user-defined vector. As well as using a docking score to guide optimisation, we also experiment with functions that combine other molecular properties, such as molecular weight, and 3D structural information, such as protein-ligand interaction profiles (PLIP)³⁷⁰. To address the issue of synthetic tractability of the compound designs, we combine the workflow with regular searches of the Enamine REAL database to ‘seed’ the chemical search space with promising purchasable compounds. After testing and optimising the hyperparameters of the active learning models, we apply the algorithm to the prospective design of inhibitors of the main protease (MPro) of SARS-CoV-2, the virus responsible for the COVID-19 pandemic. This target has undergone extensive study in recent years. The COVID Moonshot Consortium used open science crowd-sourced designs, in combination with high-throughput structural biology and assays, free energy calculations, and machine learning-driven synthetic route predictions, to generate a series of potent inhibitors²¹. Other notable approaches

that include biological confirmation of hits have employed, for example, structure-based design starting from a drug repurposing study¹⁶, virtual screening of a curated collection of commercially available compounds³⁷¹, a deep reinforcement learning model using pharmacophore and substructure matches with known inhibitors³⁷², and a deep generative framework using only target sequence information as input (along with prioritisation based on factors such as docking score and retrosynthetic feasibility)³⁷³. Here, we employ active learning to prioritise compounds for purchase and testing from the Enamine REAL database based only on early fragment hits. We suggest several novel designs that show activity in a fluorescence-based Mpro assay, as well as automatically generating several compounds that show high similarity to known Moonshot hits.

6.2 Methods

6.2.1 Workflow Design

In the first iteration of FEgrow, we used the gnina convolutional neural network (CNN), which has been jointly trained on binding pose and affinity prediction^{6,308,309}. We showed that the gnina ‘CNNAffinity’ scores (predicted pK) correlated reasonably well with experiment for ten series of congeneric inhibitors built using FEgrow¹⁷. Here, we add further options for scoring molecules based on protein-ligand interaction profile (PLIP)³⁷⁰, molecular properties, or a combination thereof. For construction of the PLIP score, interactions formed in the available protein-fragment complex crystal structures were one-hot encoded to form a reference vector of desired interactions (here, hydrophobic, hydrogen-bonding, π -stacking, and salt bridge were all identified). A similar vector was constructed for the designed de novo compound, and its Tanimoto similarity to the reference vector used as the objective for optimisation. It has been argued that combining information from various properties can also be advantageous³⁵¹, for example by using pharmacophore constraints in combination with docking scores, and we make use here of a simple, combined score (CS):

$$CS = \left(\frac{pK}{MW} \right) \times \left(\frac{PLIP}{0.3} \right) \times 100 \quad (54)$$

which aims to maximise the predicted gnina affinity (pK) and the protein-ligand interaction profile (PLIP) similarity to reference structures, while keeping the molecular weight (MW) low.

6.2.2 Database Search

A challenge for automated growing of linkers and R-groups, and for de novo design in general, is the synthetic tractability of the designed compounds. Approaches to address

this limitation could include a synthetic accessibility score in the objective function³¹² or the expert curation of libraries with known synthetic routes³⁶⁸. However, we wished to fully automate the design process, and be confident of acquiring compounds for rapid design-make-test-analyse cycles. We therefore make use of the rapidly-growing make-on-demand compound libraries as a surrogate measure of synthetic accessibility. Ideally, we might use the entire catalogue as a chemical space in which to perform the active learning. Although such an approach has been used as a one-off screen³⁷⁴, evaluating the regression models used here soon becomes prohibitively expensive in an active learning cycle. On the other hand, highly efficient methods have been developed for similarity and substructure searches of these libraries³⁵⁹. We therefore make use of these searches to seed the chemical space with compounds that are similar to the predicted actives at each step of the active learning cycle (Figure 48). In this way, at the subsequent acquisition step, we enable the algorithm to pick compounds for growing and scoring that are likely to be scored highly (due to similarity with other highly scoring compounds) and available for purchase or synthesis (due to presence in on-demand libraries).

In detail, the Enamine REAL database of 4.5 B compounds was searched for similarity to designed molecules through the public interface to SmallWorld <https://sw.docking.org>, using a graph-edit-distance space search³⁵⁹. At each cycle, 100 new, top-scoring compounds were searched, and up to 100 of the most similar compounds from the REAL database were extracted per search query (using a maximum distance of 5 steps). This 10 K compound set was filtered for substructure match with the core using RDKit⁵, and those compounds that passed were added to the active learning search space. Active learning then selects compounds for scoring following Enamine enrichment, as usual, but there is no explicit bias to select compounds from the on-demand catalogue.

6.2.3 Computational Details

Protein input structures were taken from the set of noncovalent complexes crystallised early during the COVID-19 pandemic³⁴⁷. In particular, the input PDB: 5R83 was used as the receptor structure for active learning design, and Chimera was used to add hydrogen atoms²⁸⁸. The ligand was truncated to include only the pyridyl moiety, as this appeared in other available crystallised fragments in a consistent binding mode (PDB: 5RE4, 5REH, 5R84, 5RF3³⁴⁷) and with a suitable vector for growth into the binding pocket. The full set of 23 non-covalent complexes (that had ligands bound in areas of the pocket accessible by a growth vector) was additionally used for construction of the reference PLIP³⁷⁰ interactions.

For testing of the active learning protocols, the chemical space was assembled by combining the pyridyl moiety with 508 R-groups³⁰¹ and 100 of the most common linkers³⁵⁷ from the FEGrow library. A total of 47710 unique molecules were successfully grown into

the binding pocket and scored using the gnina CNN scoring function⁶. A further 1656 molecules were assigned a penalty score of $pK = 0$ as they could not be embedded due to steric clash with the protein. In cases where rare errors occurred, such as a failure to assign force field parameters, the molecules were discarded completely.

The previously tested FEgrow molecule building protocol was applied throughout¹⁷. The ETKDG algorithm²⁹⁹ was used to generate 50 conformers, using a 0.5 Å root-mean-square similarity threshold. Any conformers with an atom closer than 1 Å to any atom in the protein was discarded. Energy minimisation was applied using a hybrid machine learning / molecular mechanics energy function in a mechanical embedding scheme¹⁷. The ANI-2x potential²²⁹ was used for the ligand, in cases where all elements in the molecule are covered by the model, or the Open Force Field Sage³⁷⁵ potential otherwise. The lowest energy conformer was retained for scoring.

An active learning library based on scikit³⁷⁶ and modAL³⁷⁷ python packages was adopted from another study³⁶². A set of molecules to initialise the active learning cycle can be selected via RDKit’s MaxMin picker⁵ from the chemical space, or picked at random. The processing was parallelised using the Python library Dask³⁷⁸, which supports a diverse set of technologies, including the Slurm Workload Manager that is deployed ubiquitously on high-performance computing clusters. Dask is used to secure resources (scheduling workers on Slurm), submitting work and retrieving results. The three major computationally-expensive components were parallelised: 1) building and scoring of the molecules, 2) computing the Morgan fingerprints, and 3) computing the Tanimoto similarity across the chemical space for the Gaussian Process modelling.

6.3 Results

6.3.1 Interfacing FEgrow with Active Learning Enables Efficient Search of Chemical Space

In order to investigate the performance of the active learning protocol, and the effect of machine learning hyperparameters, we built a labelled ‘oracle’ set of 47,000 compounds using standard FEgrow input settings (see Computational Details). This is a larger set of compounds than would be typically built and scored against a target, but knowing the affinities of the full chemical space enables us to assess the performance of the active learning approach. The common core was selected to be a pyridyl fragment common to several early crystal structures of the SARS-CoV-2 main protease³⁴⁷, located in the S1 pocket with a vector pointing into the enzyme active site (Figure 49(a)).

Figure 49(b) shows the distribution of predicted binding affinities, computed using the gnina convolutional neural network scoring function⁶ from FEgrow built structures. The scores are symmetrically distributed around $pK = 4.5$, with a maximum affinity of

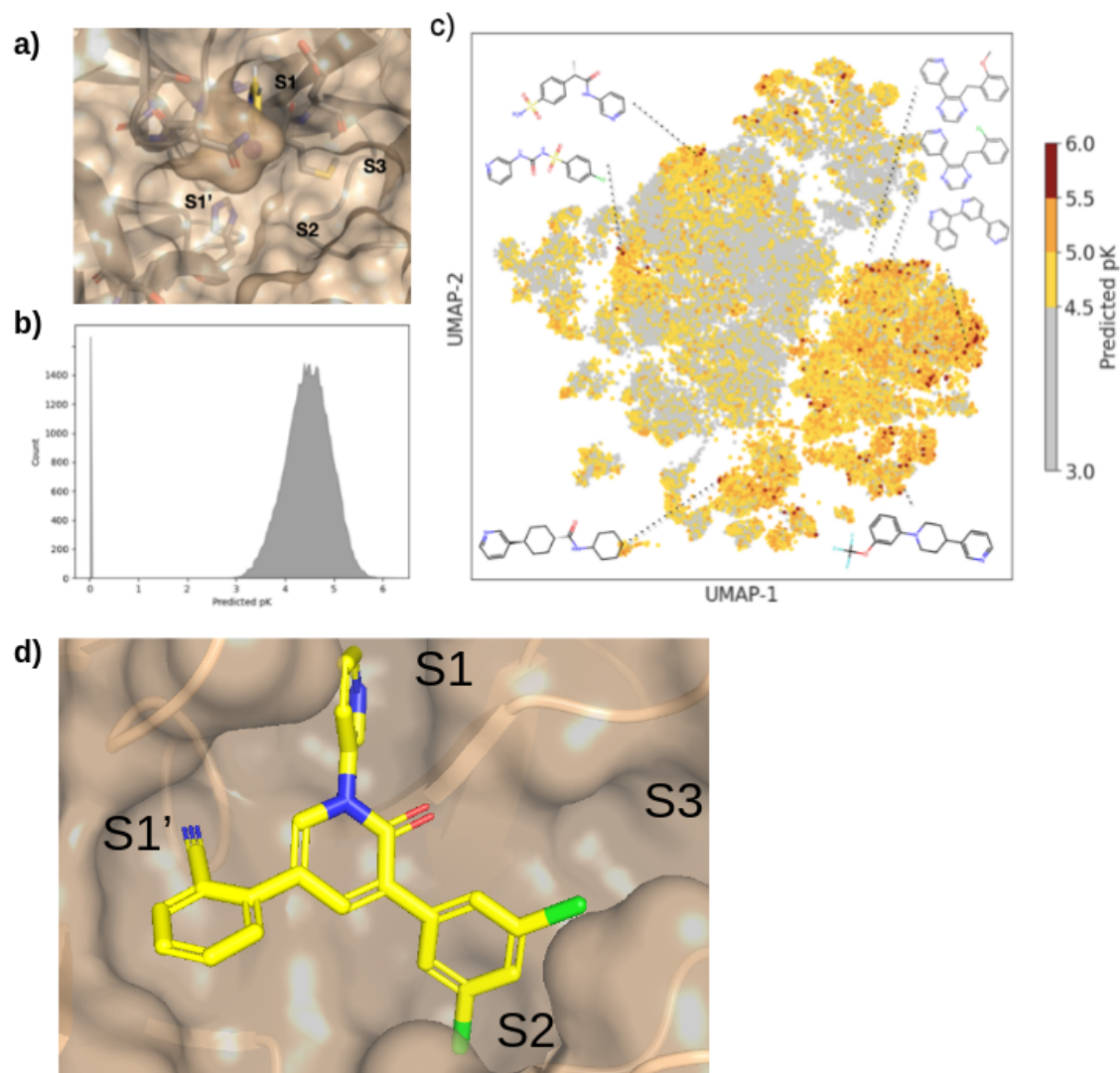


Figure 49: a) The position of the ligand core (in the M^{pro} active site) and definitions of binding pocket labels, the purple sphere is the hydrogen atom for replacement. b) Histogram of computed pK for the 47 K compound oracle dataset. c) UMAP of entire 47,000 oracle chemical space, coloured by computed pK (the activity limit of 4.5 was arbitrarily set). 2D structures of representative strong binders are included. d) A known, 4 μ M, uracil-based binder.¹⁸

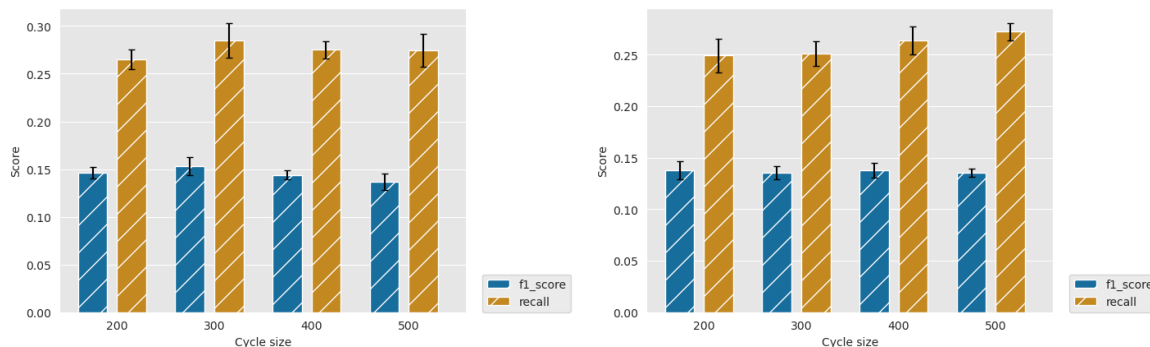


Figure 50: Recall and F1 score for diverse initial selection GBM (left) and GP (right) models, and greedy acquisition for identification of top 2 % scoring compounds for different cycle sizes. Error bars show standard errors over five runs.

around 6.0, which is indicative of a set of low molecular weight (range between 100 and 350 Da, **Figure S1**), unoptimised compounds at the start of a hit finding effort. Indeed, it is at this stage where the options for expansion are vast, and strategies to suggest exploration of hits are particularly valuable. Note that compounds that could not be built (for example, due to steric clashes with the protein) are arbitrarily assigned a pK of zero, so that this information can be included in the active learning model.

Figure 49(c) further shows the UMAP projection of the chemical space, coloured by gina predicted pK. The visualisation shows a diverse composition of linkers and functional groups, with well-spread clusters of the highest affinity binders, potentially providing a challenging search space for active learning. Figure 49(b) also shows locations in the chemical space of example linker and R-groups, attached to the pyridyl core, that make up the stronger predicted binders. Favourable predicted linkers include amides, sulfonylurea and various 6-membered ring heterocycles, and relatively bulky R-groups are feasible, which is generally expected given the size and shape of the binding pocket^{21,347}. (Note that at this stage no consideration is given to synthetic accessibility or stability of the compound designs).

We next sought to use active learning to accelerate the search through this chemical space, using the oracle to assess the performance of model hyperparameters, and using the predicted binding affinity as the optimisation target. In particular, we have investigated the effects of initial compound selection (random or diverse), number of compounds picked per cycle (in the range 200–500), machine learning model (GBM or GP) and acquisition method (greedy or UCB). As discussed, the dependence of active learning efficiency on the choice of model parameters is well documented, and so we do not devote much space to it here.

By way of example, Figure 50 shows the effect of the number of compounds picked per cycle on model recall and precision (F1 score) for the two machine learning models (GBM

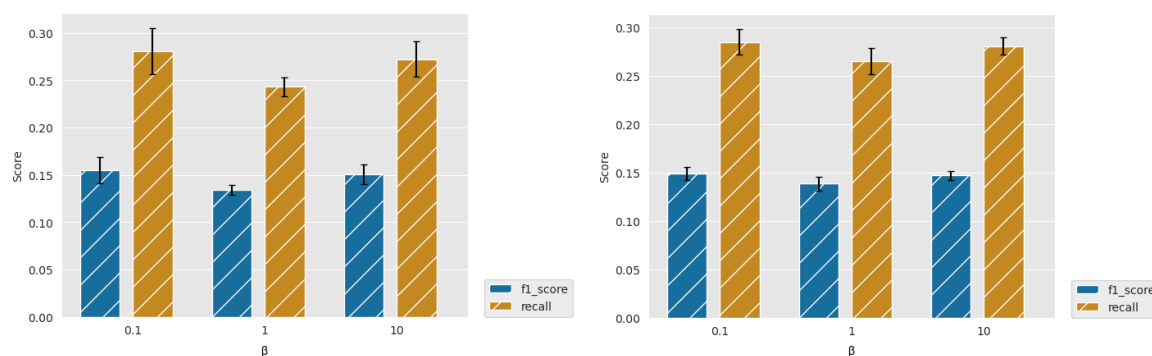


Figure 51: Recall and F1 score for diverse initial selection using GP and UCB acquisition (repeating the same protocol three times with different β values) with cycle sizes of 200 (left) and 400 (right) for identification of top 2 % scoring compounds. Error bars show standard errors over five runs.

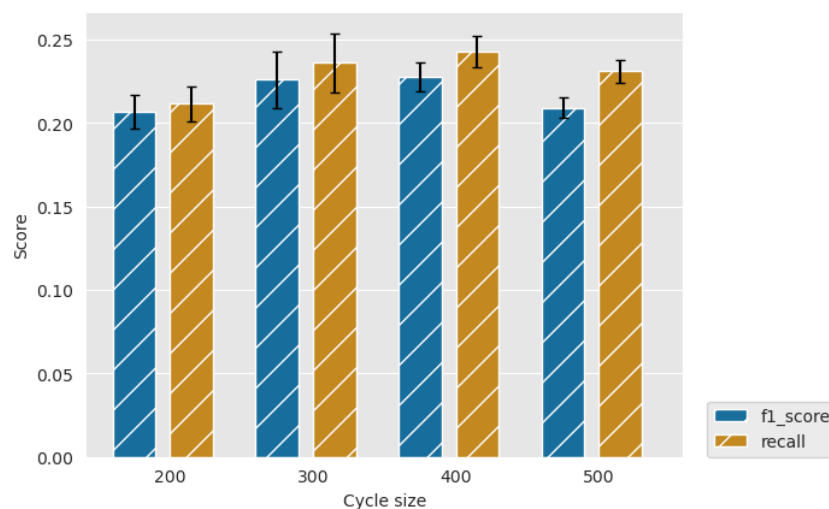


Figure 52: F1/recall for Experiment: Random initial molecule selection, GP regression model and greedy acquisition at 5 % as a function of different cycle sizes.

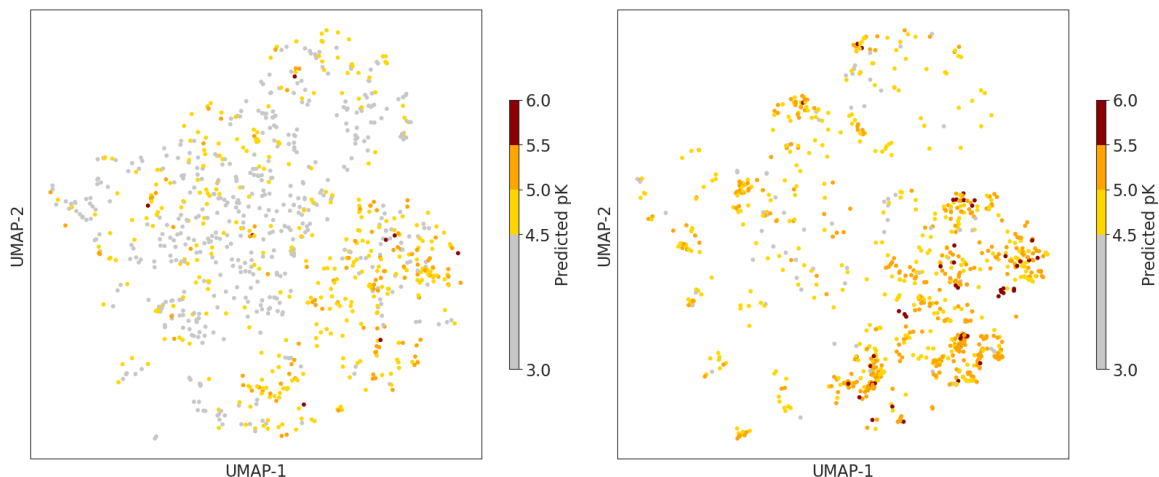


Figure 53: Difference in selection for first (left) and final (right) active learning cycles, for a GP model with UCB acquisition function ($\beta = 10$), a cycle size of 200 and a diverse set of starting compounds showing a narrowing into areas predicted to be potent and avoiding unpromising areas.

and GP). For a fixed total number of compounds selected (here, 2500), one might expect the model to improve at small sample sizes (hence, more active learning cycles), but we find that the efficiency is already well converged when picking 500 per cycle. Similarly, the choice of machine learning model has little effect, with slightly higher metrics for the GBM model, but both recall and precision comparisons are within the error bars. Figure 51 further shows the effect of using the UCB uncertainty-based acquisition function, instead of greedy selection, in conjunction with the GP machine learning model. There is some small improvement in recall over greedy selection, but no significant change in the metrics used either as a function of cycle size or the β parameter in eq 52.

Note that for the current dataset, random selection would give a recall of 0.05 and F1 score of 0.03 for identification of the top 2% of compounds. Therefore, with recall of around 0.25–0.30 for most of our experiments, we see efficiency improvements with active learning of around a factor of 5x compared to random selection. For reference, the growth and scoring of this compound set in FEGrow requires around 1000 cpuhrs, which is not prohibitive, but automated acceleration at no cost is clearly worthwhile.

In the next section, we choose to use a GP model with UCB acquisition function, with a cycle size of 200 and a diverse set of starting compounds. The overall accuracy of the chosen regression model (using $\beta = 10$), following training on 5 % of the dataset, is 0.97 pK units (**Figure S2**), which is competitive with typical models used in active learning with fingerprint-based representations³⁶³. Figure 53 shows a similar UMAP projection as in Figure 49, but now only showing compounds acquired by our chosen active learning model in the first (left) and final (right) cycles. We observe both a wide exploration of the chemical space, which is important to increase diversity in the final

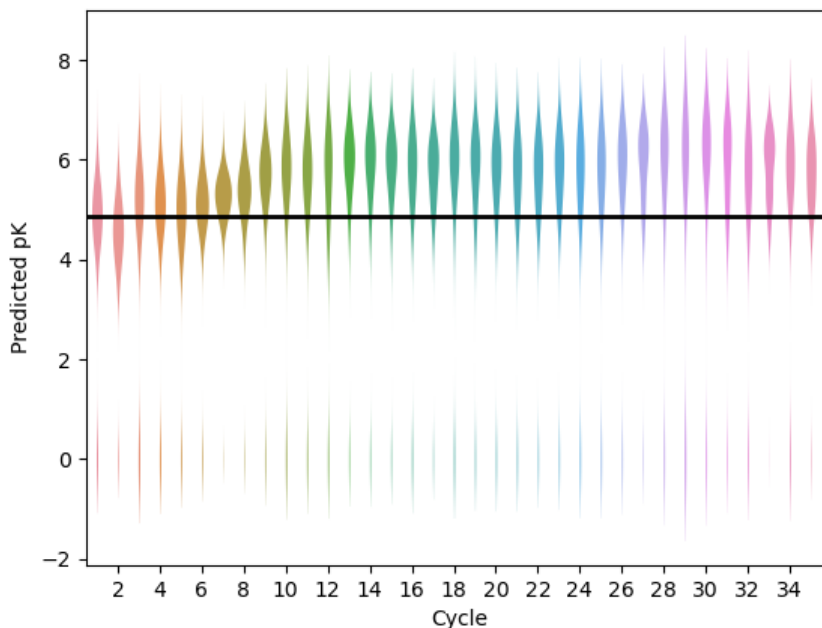


Figure 54: Active learning drives improvements in predicted binding affinity. A GP model is used, with UCB acquisition function ($\beta = 0.1$), a cycle size of 200 and a diverse set of starting compounds. The solid horizontal line shows the average score for 377 compounds randomly selected from the REAL database that were built with FEgrow.

set, and a focusing of the explored regions in the final cycle to compounds with a higher predicted binding affinity, which is important for the use of the model to identify strong binders.

6.3.2 Active Learning Driven Fragment Expansion Identifies Potential SARS-CoV-2 MPro Inhibitors

Having established that the active learning protocols tested here are able to improve the efficiency of chemical space searches with FEgrow, we turn now to prospective design of potential noncovalent SARS-CoV-2 MPro inhibitors. A wealth of computational and experimental data has been generated for this target in recent years, but here we limit ourselves to structural information that was available in the early months of the COVID-19 pandemic. In particular, as in the previous section, we consider expansion of the pyridyl fragment (PDB: 5R83) along a vector into the binding pocket containing the catalytic cysteine (Cys145)³⁴⁷. We now expand the size of the chemical space to an initial 250,000 molecules, built from the combination of supplied libraries of 500 linkers and 500 R-groups, such that full building and scoring of the space is prohibitively expensive for routine study. To address the issue of synthetic feasibility of the output designs, we add an additional step in the active learning cycle (Figure 48), whereby the chemical

space is periodically seeded with compounds from the REAL database that are similar to the highest scoring compounds (see Methods). **Figure S3** demonstrates successful incorporation of the Enamine compounds into the active learning cycles, with a significant fraction of the built and scored compounds originating from this source.

Figure 54 shows an example design run, optimising the compounds for predicted pK using the gina scoring function (further examples are given in the **Supporting Information**). The distribution of predicted affinity increases over the first 10 active learning cycles then starts to saturate with a mean gina predicted pK close to 6 (micromolar affinity). Over the full run, 95% of the compounds were successfully built (assigned $pK > 0$) and 15% had a predicted $pK > 6$. For comparison, we also extracted 1000 molecules at random that contained the pyridyl substructure from the REAL database used to seed the active learning cycles. For this set, 377 molecules (38%) could be successfully built, with an average predicted $pK = 4.9$ and only two compounds with predicted $pK > 6.0$ (0.2%).

Figure 55a) shows the highest scoring compound from this run, with a predicted affinity of 88 nM. The compound extends hydrophobic contacts into the S3 and S1' pockets, for example with Met165 and Thr25, but despite this does not form any specific polar interactions (other than the original core interaction with His163). Since an early fragment screen had provided valuable information about the nature of potential protein–ligand interactions in this binding pocket, we sought to reduce the reliance on the gina scoring function and drive the active learning towards compounds that recovered known crystallographic information (see Methods). Figure 55b) shows the top-scoring compound, as defined by the Tanimoto similarity to the vector of reference interactions. In this case, the grown molecule forms additional hydrogen bonding interactions with Asn142, Gly143, Ser144, Cys145 and Glu166, and hydrophobic interactions with Thr25 and Glu166. The majority of these interactions are recapitulated by, for example, fragments PDB: 5RGI and 5RF7 (Figure 55d)).

Finally, we sought to combine the strengths of both docking scores and crystallographic information to optimise a combined scoring function. Figure 55c) shows the top-scoring compound as defined by eq 54 after 33 cycles of active learning. Although this compound is scored much lower by the gina scoring function (predicted affinity 2 μ M), it extends into the S3 and S1' pockets and retains many of the interactions observed in Figure 55b) (e.g. hydrogen bonding interactions with Asn142, Gly143, Ser144, Cys145 and Glu166).

6.3.3 Analysis of Hit Compounds

The top 500 compounds from each of four active learning runs (two optimising gina predicted pK , one optimising protein–ligand interactions, and one optimising the combined scoring function) were checked for availability from the Enamine store. Interestingly, very

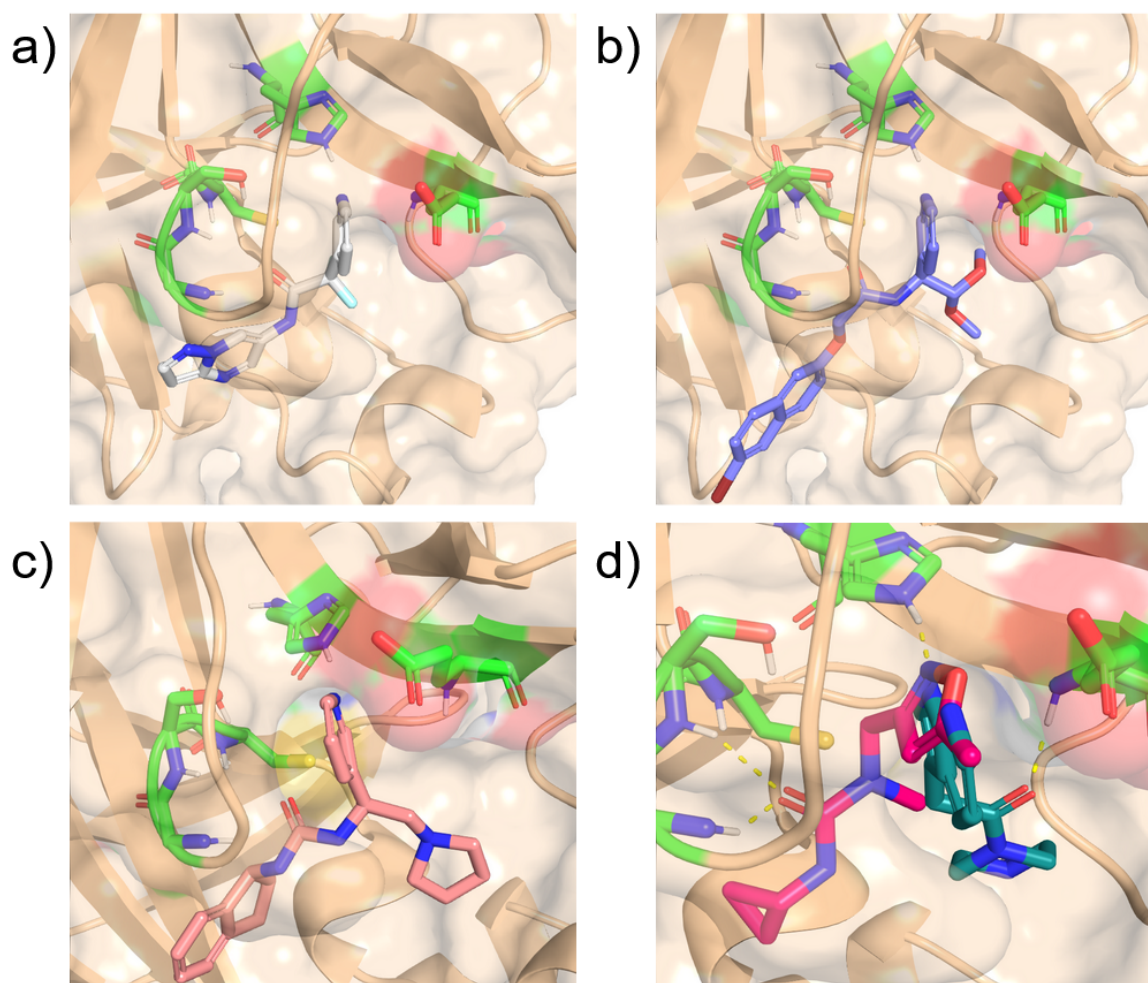


Figure 55: Top-scoring compounds from active learning FEgrow runs of the main protease of SARS-CoV2 (PDB: 5R83) using different scoring functions: a) gnina predicted pK (grey), b) protein-ligand interaction profile (blue), c) combined scoring function (pink) and d) Fragment 5RGI (red and teal) (H-bond donation by Gly143, Ser144, Cys145 and His163), and 5RF7 in green (hydrophobic and H-bond donation with Glu166).

few of the top scored by gina predicted pK were available (four in total). This is likely due to an important unavailable building block(s), and could be mitigated in future by increasing diversity and/or including direct store queries in the search process. In any case, we focussed here on outputs from the remaining two runs, and submitted the top 10 protein-ligand interaction and top 25 combination scoring compounds for costing. Finally, a total of 19 designed compounds were purchased (of which 15 had been optimised used the combination score) based on quoted price and excluding similar compounds (based on visual inspection). Two control compounds were also included; one known binder from a crystallographic fragment screen (Enamine ID: Z44592329; PDB: 5R83)³⁴⁷ and one elaborated compound from the COVID Moonshot study (Enamine ID: Z4943052515 (literature IC₅₀ 0.288 μ M))²¹. The twenty one purchased compounds (**Figure S11**) were evaluated in a fluorescence-based Mpro activity assay, performed by Siddique Amin, at 1000, 500, 10 μ M (**Figure S12**). Compounds **5** and **6** were excluded from the study due to solubility issues at 1000 μ M in the assay conditions. Five compounds (**8**, **10**, **12**, **14** and **21** (the positive control²¹)) showed reduction of Mpro activity $\leq 50\%$ at 1000 μ M. The IC₅₀ values of these compounds, except **8** which displayed background autofluorescence, were further determined (Figure 56). Compounds **10**, **12** and **14** showed a concentration-dependent inhibition of Mpro activity (measured pIC₅₀ 2.10, 3.01, 2.80 respectively). Nirmatrelvir, an orally bioavailable antiviral drug targeting Mpro, showed inhibition (pIC₅₀ 6.01), which was slightly higher than the reported IC₅₀ (0.022 μ M³⁷⁹), likely due to the limit of the assay (the enzyme concentration was at 0.2 μ M). Figure 57 shows the predicted structures of compounds **12** and **14** from the active learning design runs. Both compounds form hydrogen bonding interactions with the backbone of Glu166, as well as hydrophobic interactions in the S1' pocket.

Finally, to investigate whether the relatively low affinity of designed compounds is due to insufficient exploration of chemical space or the empirical objective functions used to optimise molecules, we performed a retrospective analysis of the designed compound space against known binders resulting from the COVID Moonshot crowd-sourced discovery campaign²¹.

In particular, Figure 58 shows the three most similar compounds from the active learning runs (as defined by Tanimoto similarity search between RDKit Morgan fingerprints with a radius of 3 and size of 2048) to a curated set of 292 hit compounds. Considering that our FEgrow runs took as input only a single PDB receptor structure and pyridyl fragment core, it is clear that this fragment growing and on-demand library screening approach holds promise for suggesting biologically active compounds early in hit discovery campaigns. However, further work is needed to ensure that the most promising compounds are located at the top of ranked lists for synthetic prioritisation and testing.

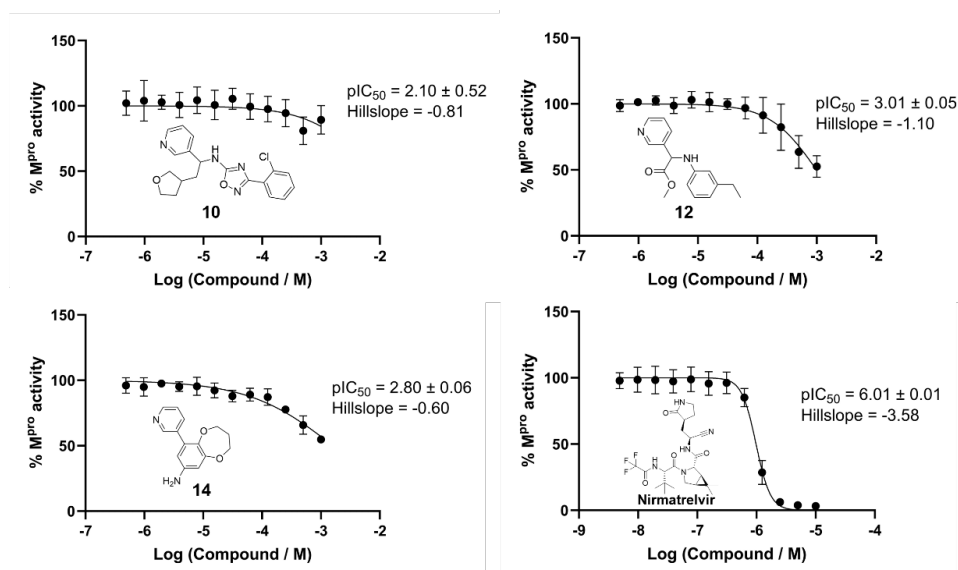


Figure 56: IC₅₀ determination of selected compounds with Mpro. Compounds **10**, **12** and **14** were tested at a top concentration of 1000 μ M. Nirmatrelvir was tested at a top concentration of 10 μ M as a positive control. Datapoints presented as mean \pm SD; pIC₅₀ presented as mean \pm SEM; two biological repeats consisting of three technical replicates. **10** consists of one biological repeat with three technical replicates. Conditions: Mpro (0.2 μ M), 12-hour pre-incubation with compounds, 20 μ M fluorescent substrate, 50 mM Tris-HCl (pH 7.3), 1 mM EDTA and temperature 25°C.

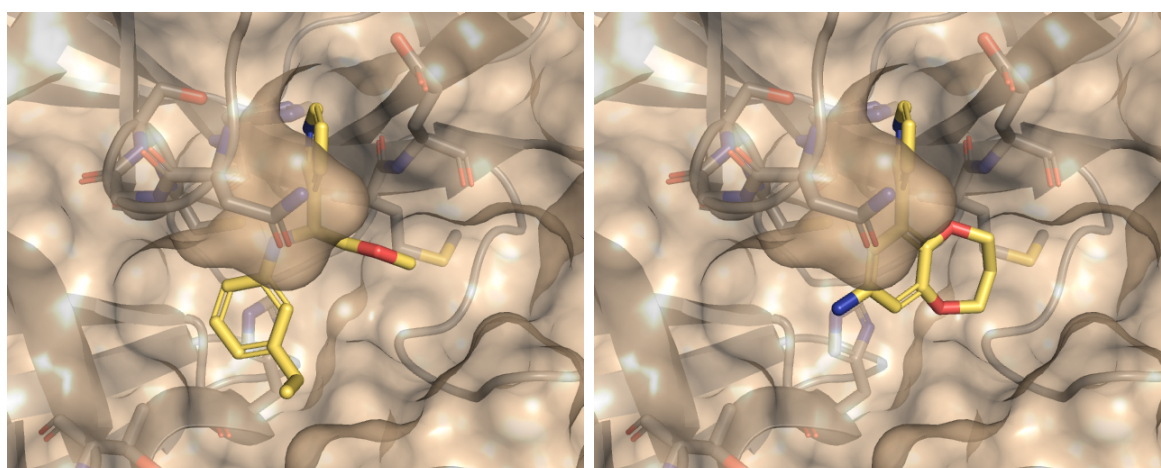


Figure 57: Predicted bound structures docked via gnina, of compounds **12** (Z1470573089) and **14** (Z8969017446).

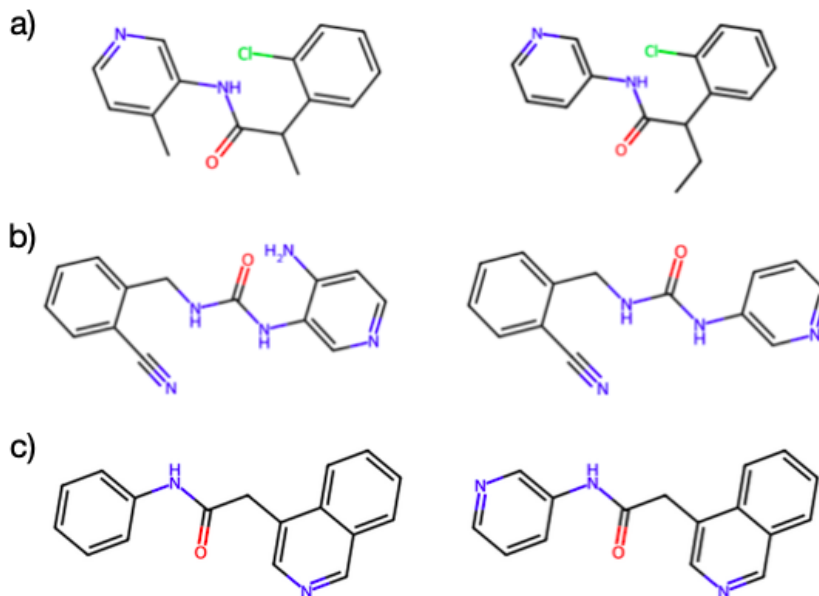


Figure 58: a) Experimental Moonshot compound (literature IC_{50} 17 μM)¹⁹ and most similar compound from this study, from active learning optimisation of predicted gnina pK ($\beta=10$), b) Experimental Moonshot compound (literature IC_{50} 54 μM)¹⁹ and most similar compound from this study, from active learning optimisation of predicted pK ($\beta=10$), c) Experimental Moonshot compound (literature IC_{50} 57 μM)¹⁹ and most similar compound from this study, from active learning optimisation of combination scoring function.

6.4 Discussion and Conclusions

In this study, we have combined the FEgrow software, an open modular workflow for building and scoring ligands in protein binding pockets, with active learning to guide and automate chemical space searches for promising binders. In agreement with numerous other studies³⁶³, we have shown that search efficiency is not too dependent on the hyperparameters of the active learning model, which include the choice of regression model, the acquisition function and number of compounds picked per cycle. For this particular study, we find efficiency improvements of a factor of around 5x over random selection, which will aid throughput of future prospective design efforts.

With the design of FEgrow, we hope to overcome some of the current limitations of de novo drug design discussed in the Introduction. Some of these limitations are addressed in the current study, and some will be addressed in future aided by ongoing advances in molecular modelling and machine learning. For example, we tackle the question of binding pose optimisation by using a fast and accurate machine learning potential (ANI-2x²²⁹) to describe the ligand energetics in a mechanical embedding scheme. However, with the flexibility of the FEgrow interface with OpenMM²²⁷, new models could be substituted in, and these are now approaching sufficient speed and accuracy (including for long-ranged

interactions) such that the entire protein-ligand complex could be described using a single, consistent machine learning potential^{3,380}. In this study, we made the approximation that the protein binding pocket is rigid and used a single receptor structure for design. However, now that ligand building and scoring is fully automated, future studies could use, for example, ensembles of receptor structures, which may be beneficial in cases where the pocket is more flexible.

A limitation of this and other similar studies is the choice of objective function in the active learning cycles. To demonstrate the flexibility of the FEgrow package, we demonstrated four design cycles here, two optimising for predicted affinity using the gnina CNN scoring function and two including a more direct optimisation of protein-ligand contacts extracted from crystallographic fragment screens. While we do not have enough data to assess the relative merits of these scoring functions, we expect the latter to be useful where experimental structural data exists, at least as part of a multi-objective optimisation in future³⁸¹. As a flexible alternative to PLIP scores trained on system-dependent crystal structures, it has also been shown that transferable neural networks can be trained on the PDBbind structural database to recognise favourable protein-ligand interactions³⁸².

As shown in Figure 48, to address the issue of synthetic tractability of the de novo built compounds, we inserted regular queries of the Enamine REAL database into the active learning cycles. In this way, we can use the initial chemical space to train the active learning regression models, and then over time seed the chemical space with compounds that are both similar to predicted actives and purchasable. In this way, we were able to test the predictions of the active learning workflow with a turn around time of a few weeks from order to biological testing. Of the 19 designed compounds that were purchased here, three showed measurable activity, but none approached the desired levels for further progression. Nevertheless, a similarity search showed the presence of effective inhibitors in the built chemical space, and so further investigation will focus on ranking compound designs ahead of purchase, perhaps via an extra stage of physics-based free energy calculations³⁶².

7 Conclusion

This thesis centres on the development and implementation of an automated workflow for computer-aided drug design (CADD). The concepts and methodologies that form the foundation of CADD (including cheminformatics, molecular dynamics, and binding free energy estimation techniques) were discussed and later exhibited in the context of FEgrow, a new open-source molecular builder. FEgrow integrates machine learning (ML) and cheminformatics toolkits to design and screen compounds across multiple targets, for fragment-based hit-to-lead optimisation. The aim of this workflow is to enhance the accuracy and efficiency of *in silico de novo* design by leveraging the modern computational chemistry software ecosystem and high-performance computing (HPC). It includes step-by-step guides and a simple API, enabling its use in projects by users with little to no coding background. Modern databases are becoming increasingly large and are now sized in the billions.³⁸³ Because of this, there is a need to be able to not only accurately test compounds, but also to navigate the vast chemical spaces these libraries offer. To this end, the FEgrow workflow was augmented utilising the Enamine REAL database to seed the chemical space of buildable molecules, with active learning. In this way, designs that are easily purchasable can be rapidly tested. Using FEgrow, compounds have been designed and experimentally validated for multiple targets, leading to half a dozen hit compounds in the low micromolar range, one of which had no previously known binders.

7.1 Future of Fragment-based Drug Design

Computer-aided drug design (CADD) has seen prolific progress over the last decade and is becoming more deeply integrated with drug design in a useful and pragmatic way. Computational techniques are seen as a trusted tool in the arsenal of drug hunting. Machine learning (ML) and AI have come to the fore, speeding up existing pipelines and expanding the applicability of others.⁷⁷ The term ‘undruggable’ is receding at pace³⁸⁴ and increasingly difficult targets are seeing new inhibitors being developed. Quantum algorithms are yet to make an appearance but are on the horizon, with the first nascent attempts being demonstrated and the first QPU (Quantum Processing Units) arriving — but these techniques are far from practical utility.³⁸⁵

This thesis has discussed a significant area of CADD, structure based drug design (SBDD), where the three-dimensional structure of a target, determined through techniques like X-ray crystallography, Cryo-EM and NMR, is used to guide the design of potential inhibitors. SBDD heavily relies on the availability of high quality structural data, and improvements in structural resolution and computational modelling are essential for its future success. There has also been a particular focus on fragment based drug design (FBDD) throughout the thesis. FBDD begins with small chemical

fragments that bind to a target and then uses computational methods, such as a free energy perturbation (FEP) and fragment growing algorithms to iteratively expand and optimise these fragments into potent drug candidates.

This work addressed the need for semi-automatic preparation of structures for FEP calculations for targets that have X-ray crystal structures available. Automating the preparation of high-quality starting models can significantly improve the efficiency of FEP workflows. The FEgrow workflow has shown success when used in real-world drug discovery campaigns, yielding promising molecules for further development. More broadly, FEP has proven to be useful in drug discovery efforts.

The accessibility of quantity and quality of structures can improve the efficacy of FBDD, especially with respect to training machine learning models. New X-ray structure techniques like XFELs that produce extremely powerful X-rays that are highly intense, coherent and short X-ray pulses, deliver in femtoseconds a similar quantity of photons that a 3rd generation synchrotron delivers in one second via electromagnetic undulation of electron beams.³⁸⁶ In XFELs, crystals are not needed and diffraction patterns can be detected off single proteins due the intensity, obviating the need for crystallisation. This can happen at room temperature and outpaces the radiative damage of the proteins since the diffraction pattern is generated before the protein is damaged. These factors mean that inherently flexible or unstable structures, such as G-protein-coupled receptors (GPCRs) can be analysed, contributing to the expanding efficacy of CADD.

Structural dynamics and reaction pathways can also be resolved (which is typically adulterated in standard X-ray conditions) and the structures of room-temperature allosteric inhibitors which are usually unavailable can be determined. This extra dynamic information under biologically relevant conditions similar to *in vivo* can be used to build more sophisticated models of protein structures, useful in fragment-based drug discovery.

Scoring remains a bottleneck in structure based computer-aided drug design (SB-CADD), particularly for *de novo* design, where the success heavily depends on the scoring functions employed. Conventional scoring methods often yield poor performance during screening, characterised by low hit rates and a high prevalence of false positives.³⁸⁷ The advent of machine learning (ML) scoring functions is not a panacea, however and, while promising, are constrained by the breadth of their training data, making them unreliable for novel targets with limited known binders. Using docking/scoring functions to predict accurate binding modes is essential for accurate binding free energy prediction e.g. FEP. Although checks exist to assure basic validity,³⁸⁸ generating poses completely *in silico* is an active area of research and would significantly advance fragment-based drug discovery.

7.2 Future of FEgrow

The half-life of knowledge in computational science is low. A testament to this fact is that the field is unrecognisable from even a decade ago, and the next decade is sure to similarly present seismic changes. Nowhere is this more apparent than the maintenance of the software; all code if left unmaintained will decay to the point of non-functionality, and quickly, due to theoretical improvements in algorithms, the release of new hardware, and the evolution of programming languages themselves.³⁸⁹

It is critical to keep in mind the context in which FEgrow was developed when considering its future, including its sustainability. In academic contexts, code is generally not of production grade and is often written for a singular and possibly mutable purpose. The exact implementation details can remain obscure (even to the authors) until relatively late in the project's life, and code often has no sense in which it is 'complete' with new applications and techniques being devised and applied to the existing code-base. This is in contrast to production code that has a clearly defined scope which exists within an environment that assures both the ability and incentive to maintain it.³⁹⁰

Software projects that commence with the hiring of postdoctoral researchers and/or PhDs still need to be maintained long after they are no longer associated with the project, and this is a common source of abandoned codebases. The assurance of longevity for projects can be achieved in various ways, for instance by integrating the project into the ecosystem of a larger organisation that possesses the resources to maintain it. Examples include Open Force Field (OpenFF), Open Free Energy within the Open Molecular Software Foundation ecosystem, and SOMD within OpenBioSim.^{391,392}

In an ideal world, each package would represent a part of a whole ecosystem built from reusable libraries, which consequently reduces development time spent on what is essentially higher order boilerplate code (large language models have also shown promise in their ability to achieve this sort of task³⁹³). Packages should have minimal dependencies; be open-sourced whilst following standard software engineering practices, and — crucially — do a single thing well in an interoperable fashion. The benefit of such a system is lower overhead for anyone wishing to develop software, more numerous tools and ultimately higher quality research that is more easily reproducible.

A focus on modular and well designed open-source software is of supreme importance, and is the key to a successful tool. That is, a tool that in the context of drug discovery is used by enough people over a long enough period of time to have contributed to the discovery of drug-like compounds. The ability for FEgrow's code to be updated and maintained is core to the nature of the project and its design facilitates replacing individual components as and when new approaches and techniques appear.

For example, incorporating newer NNPs to generate better binding poses such as MACE-OFF23 in conjunction with OpenMM or utilising ML methods for faster charge

generation — replacing semi-empirical calculations that currently take minutes per molecule with much faster neural network charges that have typical root mean square errors relative to their underlying training data of less than 0.02e.¹⁹⁸

Chemical beauty is notoriously difficult to quantify and ADMET/QED prediction is always improving. Currently implemented using standard RDKit functionality in FEgrow, it is an area that is ripe for improvement, and recently,³⁹⁴ human-in-the-loop models have been trained exploiting the expertise of humans to evaluate molecules in an Elo style ranking system, which outperform standard metrics in deprioritising undesirable compounds.

Synthetic accessibility is another issue for the FEgrow workflow, and integrating FEgrow with automated methods for suggesting libraries of synthetically tractable mutations is a promising avenue (e.g. REINVENT)³⁹⁵ ML packages like these can be implemented into FEgrow as and when they appear.

One of the most important components of FEgrow is the scoring function used to predict the binding affinity of a ligand in complexation with a protein. While the current CNN scoring function used is reasonably accurate, predictions for affinity using gnina (see 5.3) tended to have a low dynamic range and failed to accurately predict experimental values, but were useful for ranking compounds relative affinity. It is also important to note that gnina is not the only possible choice, nor will it remain state-of-the-art for very long, again emphasising the importance of a modular approach. The use of different scoring functions is another future direction, such as FRAME, a SE(3) equivariant neural network, which has appeared recently.³⁹⁶ Remarkably, despite only being trained on around 4000 examples of protein-ligand structures extracted from the PDBbind database (supplemented by negative decoy structures), FRAME is able to learn to recognise molecular interactions, recapitulating, for example, the experimental geometry of hydrogen bonding and π - π stacking interactions in protein-ligand complexes.

Incorporation of experimental assay data to iteratively fine-tune the scoring function on a per target basis directly (as opposed to using AL based regression models) is another feature that would be desirable. Currently, the scoring function used is static and valuable information ascertained from experiment is unable to be used except in an *ad hoc* manner.

In the current version of FEgrow, the protein binding site is static and as such is unable to model induced fit effects that have been shown to be integral to a ligand's affinity for its target.³⁹⁷ There are various ML methods for predicting protein dynamics,³⁹⁸ along with different experimental techniques that can directly measure them e.g. X-ray free electron lasers (XFELs)³⁹⁹. Protein conformational change is an obvious omission to the current workflow, and the inclusion of binding pocket residue adjustment between e.g. apo and holo structures might allow more accurate affinity prediction, especially for larger ligands.

The role of water in drug design is not to be understated, and predicting the effect

ligand binding has on water networks can improve the accuracy of *de novo* design. Water models were incorporated in version 1.0.0 of FEgrow in a basic fashion, but not integrated into the core workflow. This could be done, for example, by adding functionality for water network prediction, allowing the prediction of non-labile waters that are likely to act as conduits for polar interactions. At present, the workflow requires an experimental binding mode as a starting point, typically relying on X-ray crystal structures of protein-ligand complexes, which are often limited in availability. Further, FEgrow implicitly assumes that as the core fragment is expanded, the binding mode remains unchanged. However, this assumption is generally unreliable and its validity diminishes the more the molecule grows beyond the initial core. The ability to predict binding modes or assess whether modifications, such as added groups, are likely to alter the binding interaction is a critical factor for *de novo* design.

Similarity search needs to be addressed as *de novo* designs are only as good as what is available to test experimentally. There are newer metrics that are more sophisticated than simple fingerprint searches using 3D shape or electrostatic similarity.⁴⁰⁰

In summary, FEgrow is a modular workflow that aims to be easy to maintain and use. It contains tutorials and a simple API so that it can easily be applied to any drug discovery project, even by those that have limited coding experience. The workflow has potential to be further iterated to ensure its sustainability, longevity and utility, for example by embedding into the ecosystem of a larger organisation; making use of alternative scoring functions and ML methods for predicting protein dynamics, and integrating water network prediction functionality into the workflow.

S8 Appendix 1: FEgrow Supplementary Information

S8.1 Case Study II: SARS-CoV-2 Main Protease

Calculations in this section were performed by Dr Mateusz Bieniek, but included here for completeness.

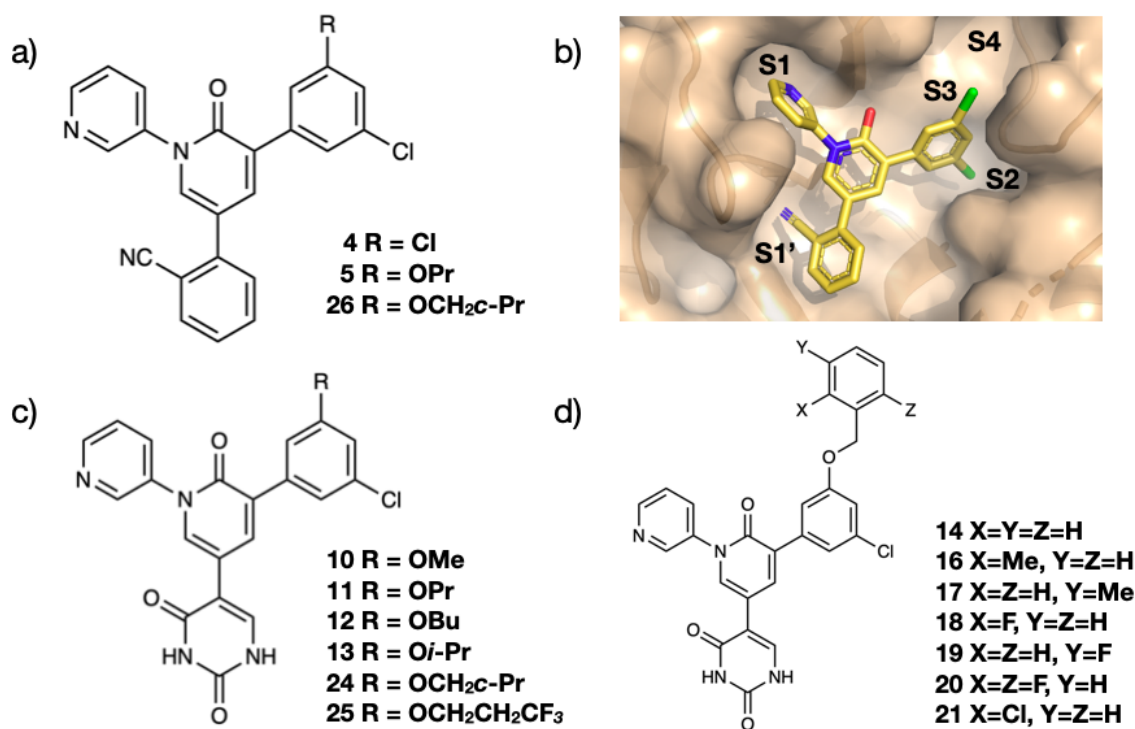


Figure S1: a) Cyanophenyl-based M^{pro} inhibitors. b) X-ray crystal structure of **4** in complex with the protease, with discussed binding pockets labelled. c,d) Uracil-based M^{pro} inhibitors.

The main protease (M^{pro}) of SARS-CoV-2, the virus responsible for the COVID-19 pandemic, is an attractive target for the development of antiviral agents⁴⁰¹. The Jorgensen lab has focused on the development of drug-like, non-covalent inhibitors of the protease through lead optimisation of virtual screening hits¹⁶. In particular, starting from the anti-epileptic drug, perampanel, researchers combined model building with the BOMB software, with free energy calculations, to rapidly yield potent antiviral compounds. Figure S1 shows the structures of the two main series of cyanophenyl- and uracil-based compounds investigated. A high-resolution x-ray crystal structure of **4** with M^{pro} confirmed binding to the S1, S1' and S2 pockets, with space to grow into the S3–S4 region¹⁶.

In what follows, we employ FEgrow to retrospectively build and score the listed analogs (Figure S1) to demonstrate the potential utility of the workflow in guiding future design efforts. Starting from the crystal structure of **4** (PDBID: 7L10), we begin by replacing

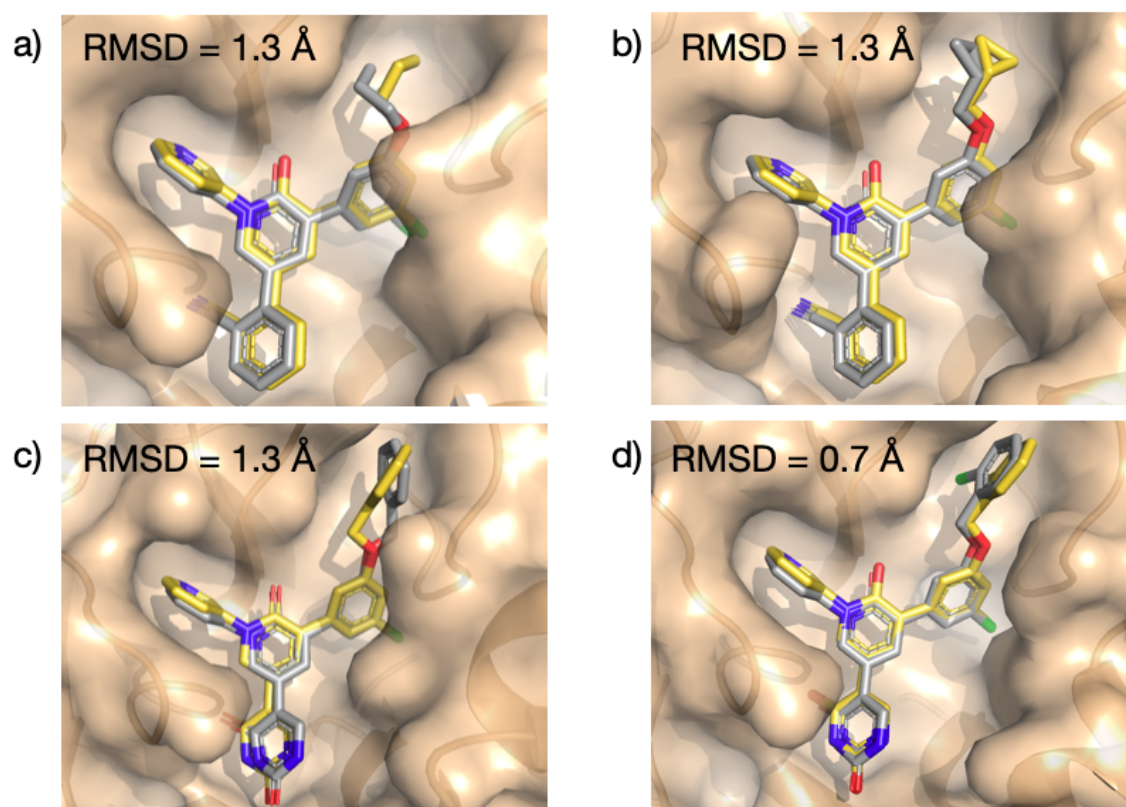


Figure S2: Overlay of (a) **5** and PDBID: 7L11, (b) **26** and 7L14, (c) **14** and 7L12, (d) **21** and 7L13. Crystal structures are coloured in yellow, and modelled binding poses in grey. Root-mean-square distances (RMSD) between predicted and experimental coordinates of atoms in the built R-groups were calculated using RDKit⁵.

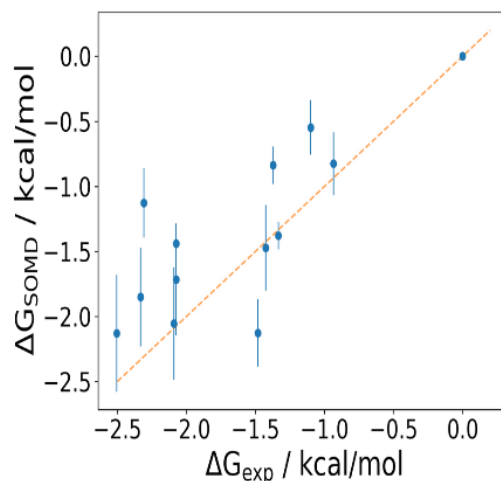


Figure S3: Comparison between free energy calculations and experiment. Binding free energies of 13 analogs of the uracil-based M^{pro} inhibitors, relative to compound **10**. The error bars indicate one standard error based on least square fitting²⁰.

one of the *meta* chlorine atoms by propoxy to form **5**. The modelled structure agrees well with the corresponding high-resolution crystal structure (Figure S2(a)). In particular, the propoxy OCCC dihedral angle in the lowest energy structure (53°) matches the experimental *gauche* conformation (47°), which allows hydrophobic contact with Met165 and Leu167. Similarly, good agreement is obtained for the cyclopropyl analogue **26** with the corresponding experimental crystal structure (Figure S2(b)).

Turning attention to the uracil series, the core molecule was again built from the crystal structure of **4**, by removing the cyanophenyl group. The added uracil group has three low energy conformations, and in this case we retained the second lowest energy structure, which forms key hydrogen bonding interactions with the backbone of Thr26 and the catalytic Cys145. In agreement with the original modelling, performed using the BOMB software¹⁶, we find that again a range of substituents are permitted in the S3/S4 pocket, including substituted benzyloxy side chains (Figure S1). Figure S2(c) shows that the modelled uracil group in the S1' pocket is in good agreement with the corresponding crystal structure (7L12). However, the predicted conformation of the unsubstituted benzyloxy side chain is at odds with the crystal structure (7L12). The correct conformer is output as an alternative low energy conformer and, interestingly, the majority of the modelled larger, substituted benzyloxy groups adopt the crystal conformation. This is exemplified by **21** in Figure S2(d), which also correctly orients the *ortho*-Cl down into the S4 pocket.

The uracil series comprises a set of 13 analogs, spanning around 2.5 kcal/mol in binding free energy, and as such provides a useful benchmark for demonstrating the next stage of the workflow. Although the gnina CNN affinities for these compounds are reasonably

well correlated with experimental IC₅₀ measurements in a kinetic assay (**Figure S5**)¹⁶, it is desirable to investigate whether more rigorous free energy methods can be used to improve accuracy. Hence, relative binding free energies were computed using the SOMD software²⁹⁶, starting from the structures output by the FEgrow workflow in complex with the receptor (see **Computational Methods**). Note that we have used the lowest energy structures as input to the free energy calculations (using instead the structure of e.g. **14** that corresponds most closely to the crystal structure can introduce differences of up to 0.4 kcal/mol in free energies in our tests, but this information would not be available for prospective studies). Figure S5 shows the agreement between experiment and simulations (MUE = 0.45 kcal/mol, $R^2 = 0.53$), and the raw data is provided in **Table S5**. Here, we can see that even though we have only used information from a single crystal structure of **4** bound to the protease, the combination of structure building and optimisation with the FEgrow workflow and free energy calculations with SOMD allows the (retrospective) prioritisation of compounds, such as compounds **20** and **21** for synthesis and testing.

Molecular Property Filters. Here, we provide further information on the simple molecular property filters that are included in FEGrow.

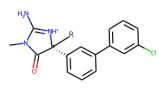
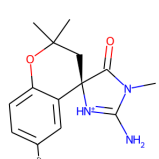
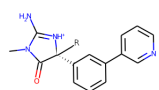
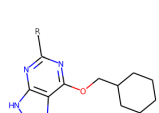
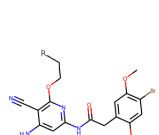
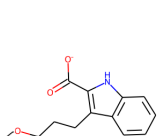
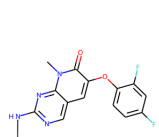
Pan Assay Interference Compounds (or PAINS)³¹³ are molecules that bind non-specifically to multiple protein targets. This can lead to unwanted side effects and increased risk of toxicity, as well as often causing false positive results in high throughput screening. Common PAINS include toxoflavin, isothiazolones, hydroxyphenyl hydrazones, curcumin, phenol-sulfonamides, rhodanines, enones, quinones, and catechols. It is worth noting that there are many instances of approved drugs containing PAINS, so they should be considered with caution.⁴⁰²

Brenk and coworkers proposed a list of unwanted substructures with undesirable pharmacokinetics or toxicity that they made use of in assembling screening libraries for neglected diseases.³¹⁶ This list of features includes sulfates and phosphates (likely resulting in unfavorable pharmacokinetic properties), nitro groups (mutagenic), 2-halopyridines and thiols (reactive).

The NIH filter (based on the work by Jadhav *et al.*³¹⁴ and Doveston *et al.*³¹⁵) defines a list of unwanted functional groups. These are split into two groups: reactive functionalities and medicinal chemistry exclusions. The reactive functionalities include Michael acceptors, aldehydes, epoxides, alkyl halides, metals, 2-halo pyridines, phosphorus nitrogen bonds, α -chloroketones and β -lactams. The medicinal chemistry exclusions include groups such as oximes, crown ethers, hydrazines, flavanoids, polyphenols, primary halide sulfates and multiple nitro groups.

Finally, we include a synthetic accessibility (SA) score³¹². This function returns a score of 1 for easy to synthesise and 10 for more challenging compounds. The score is based on a combination of fragment contributions derived through analysis of one million compounds from Pubchem, and a complexity penalty that accounts for the presence of large rings, non-standard ring fusions, stereocomplexity and size.

Table S3: PDB ID, number of R-groups grown, net ligand charge, and 2D common core structure for each target. Attachment vectors are labelled by “-R”.

Target	PDB ID	Number of R-groups	Charge	Common Core
BACE	4DJW	16	+1	
BACE(Hunt)	4JPC	31	+1	
BACE(P2)	3IN4	12	+1	
CDK2	1H1Q	16	0	
JNK1	2GMX	10	0	
MCL1	4HW2	22	-1	
P38	3FLY	14	0	

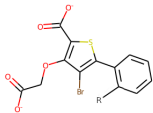
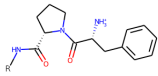
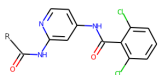
Target	PDB ID	Number of R-groups	Charge	Common Core
PTP1B	2QBS	18	-2	
Thrombin	2ZFF	11	+1	
TYK2	4GIH	12	0	

Table S4: (Continued) PDB ID, number of R-groups grown, net ligand charge, and 2D common core structure for each target. Attachment vectors are labelled by “-R”.

Target	RMSE / kcal/mol	R ²
BACE	0.94	0.00
BACE(Hunt)	1.23	0.03
BACE(P2)	0.89	0.00
CDK2	1.01	0.08
Jnk1	1.72	0.23
MCL1	1.19	0.27
P38	1.20	0.28
PTP1B	0.95	0.55
Thrombin	0.93	0.68
TYK2	1.03	0.20

Table S5: Root mean square error (RMSE) and correlation coefficient (R²) between gnina CNN affinities (converted to free energies) and experimental binding free energy, calculated as $RT \times \ln(IC_{50})$.

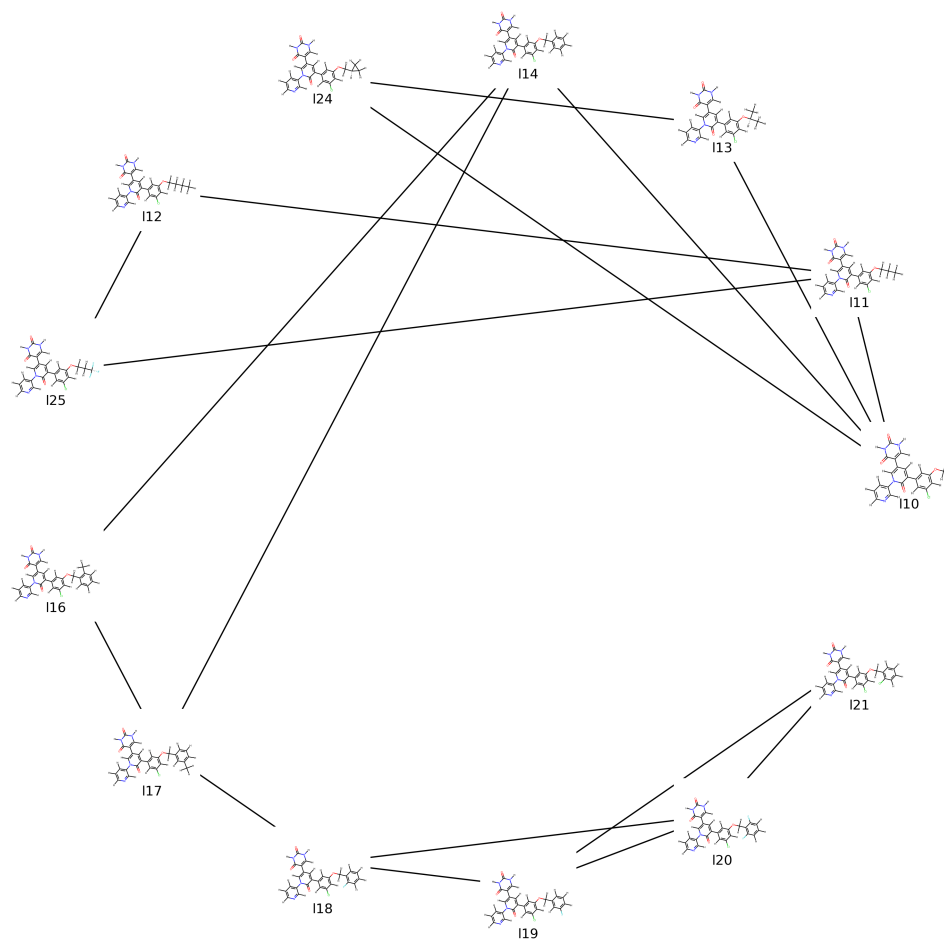


Figure S4: Network of alchemical transformations used for calculation of relative binding free energies of 13 analogs of the uracil-based M^{pro} inhibitors.

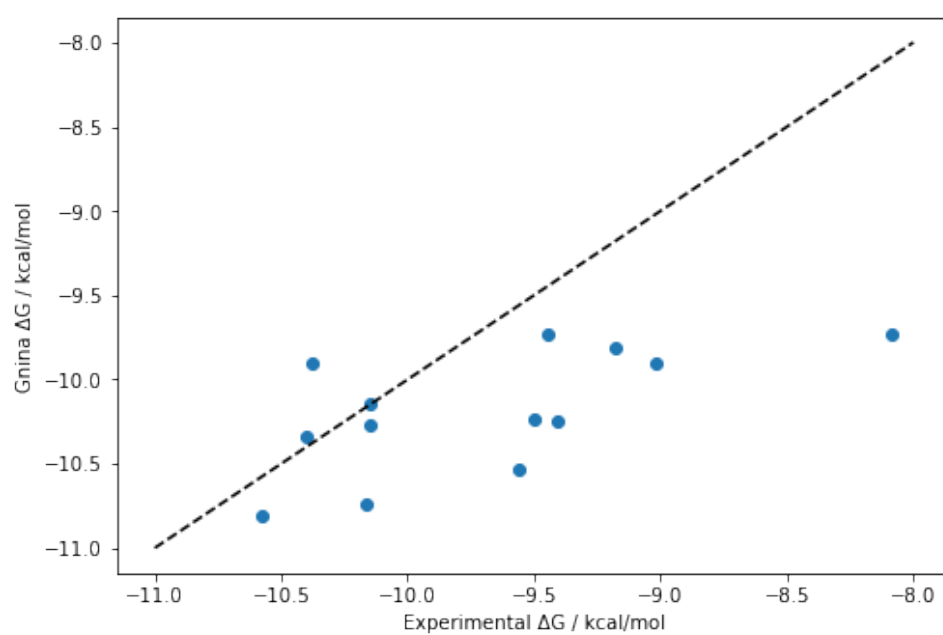


Figure S5: Comparison between gnina and experiment. Absolute binding free energies of 13 analogs of the uracil-based M^{pro} inhibitors using the gnina CNN affinity.

Compound	$\Delta G(\text{EXP})$ / kcal/mol	$\Delta G(\text{SOMD})$ / kcal/mol
10	0	0
11	-1.37	-0.84
12	-0.94	-0.82
13	-1.10	-0.55
14	-1.33	-1.38
16	-1.48	-2.13
17	-1.42	-1.47
18	-2.33	-1.85
19	-2.07	-1.72
20	-2.09	-2.05
21	-2.5	-2.13
24	-2.07	-1.44
25	-2.31	-1.13

Table S6: Comparison between free energy calculations and experiment. Binding free energies of 13 analogs of the uracil-based M^{pro} inhibitors, relative to compound **10**.

Cycle	Cycle Closure Error (kcal/mol)
24-13-10	-0.98
14-16-17	-0.14
12-11-25	-0.46
18-20-19	0.01
19-20-21	0.43

Table S7: Cycle closure errors for the network of M^{pro} inhibitors (Figure S3). Errors are calculated from the raw free energy data from SOMD, averaged over duplicate runs and forward/backward transitions.

S9 Appendix 2: Active Learning Supplementary Information

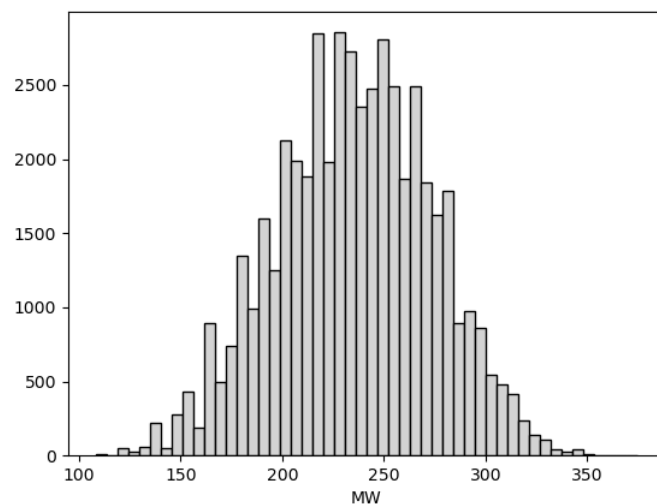


Figure S1: Distribution of molecular weights (MW, Da) for the 47 K compound oracle dataset.

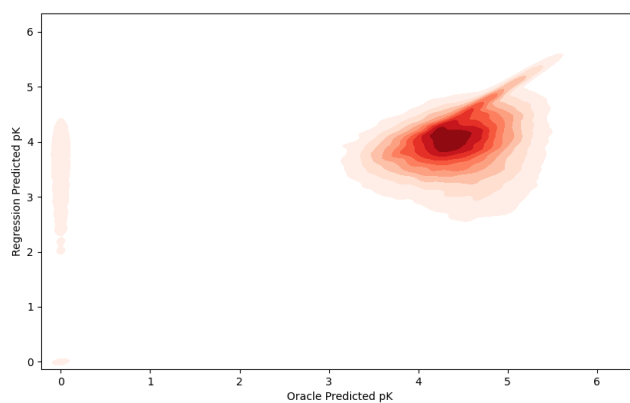


Figure S2: Correlation between the predicted pK using a Gaussian process regression model and the oracle predictions. Overall RMSE between the predictions is 0.97 pK units. (Cycle size = 200, diverse initial selection of molecules, UCB acquisition function, $\beta = 10$).

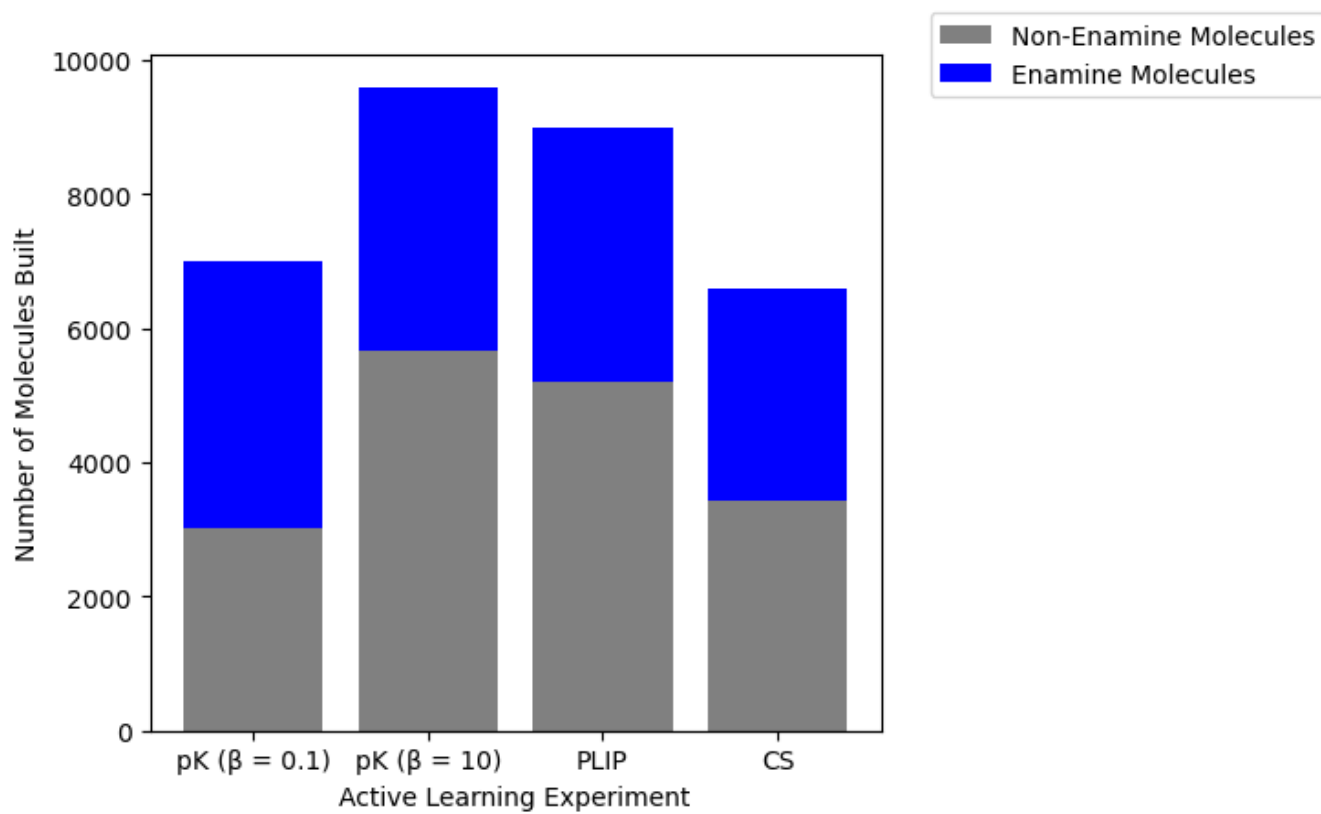


Figure S3: Histogram of the number of Enamine molecules added for each experiment, as a fraction of the total number of molecules built.

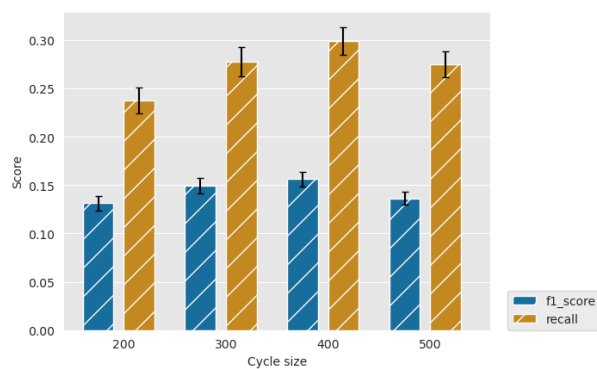


Figure S4: Top 2% activity

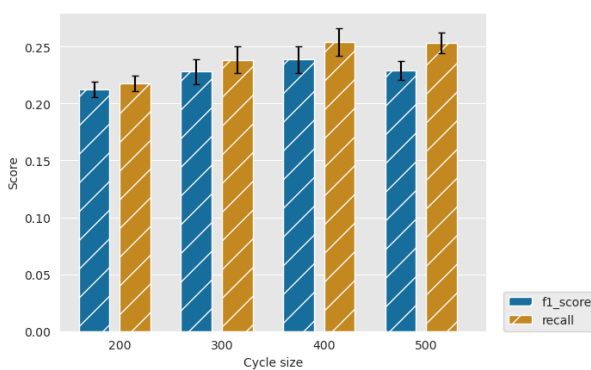


Figure S5: Top 5% activity

Figure S6: F1/recall for Experiment: Random initial molecule selection, GBM regression model and greedy acquisition at 2 and 5 % as a function of different cycle sizes.

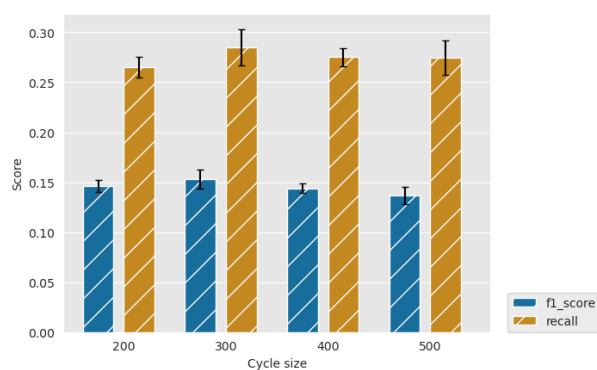


Figure S7: Top 2% activity

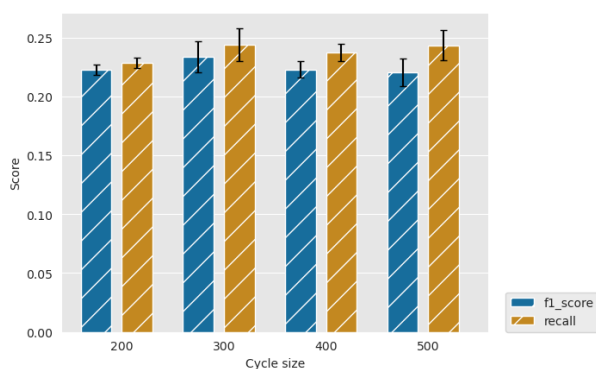


Figure S8: Top 5% activity

Figure S9: F1/recall for Experiment: Diverse (MaxMin) initial molecule selection, GBM regression model and greedy acquisition at 2 and 5 % as a function of different cycle sizes.

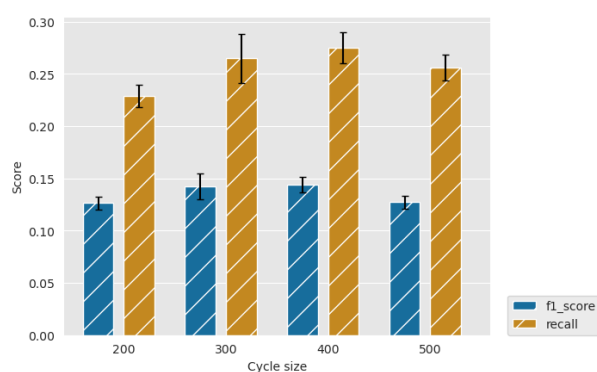


Figure S10: Top 2% activity

Figure S11: F1/recall for Experiment: Random initial molecule selection, GP regression model and greedy acquisition at 2 % as a function of different cycle sizes.

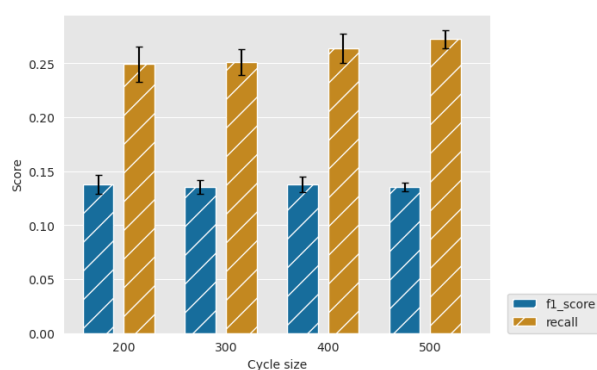


Figure S12: Top 2%

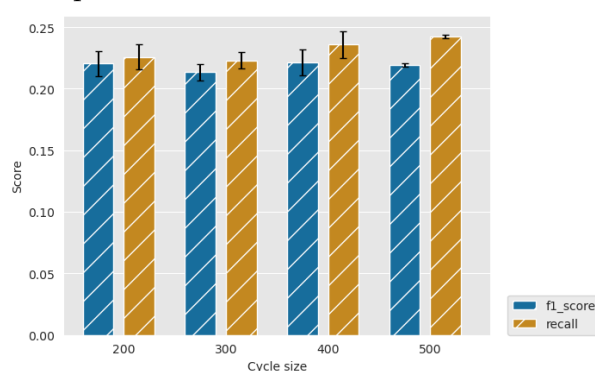


Figure S13: Top 5%

Figure S14: F1/recall for Experiment: Diverse (MaxMin) initial molecule selection, GP regression model and greedy acquisition at 2 and 5 % as a function of different cycle sizes.

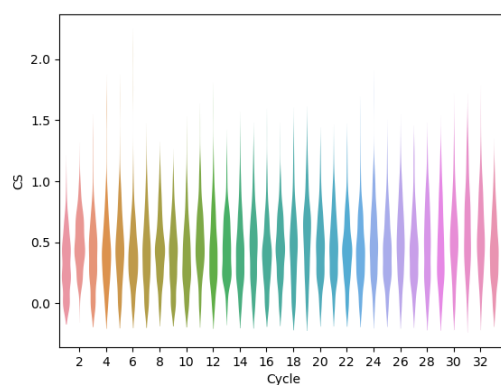


Figure S15: Active learning drives improvements in predicted CS scoring function. A GP model is used, with UCB acquisition function ($\beta = 0.1$), a cycle size of 200 and a diverse set of starting compounds.

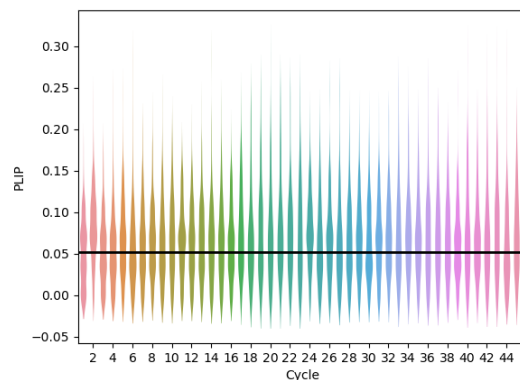


Figure S16: Active learning drives improvements in predicted PLIP scoring function. A GP model is used, with UCB acquisition function ($\beta = 0.1$), a cycle size of 200 and a diverse set of starting compounds.

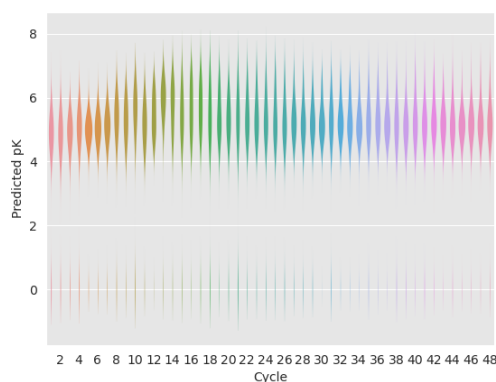


Figure S17: Active learning drives improvements in predicted binding affinity scoring function. A GP model is used, with UCB acquisition function ($\beta = 10$), a cycle size of 200 and a diverse set of starting compounds.

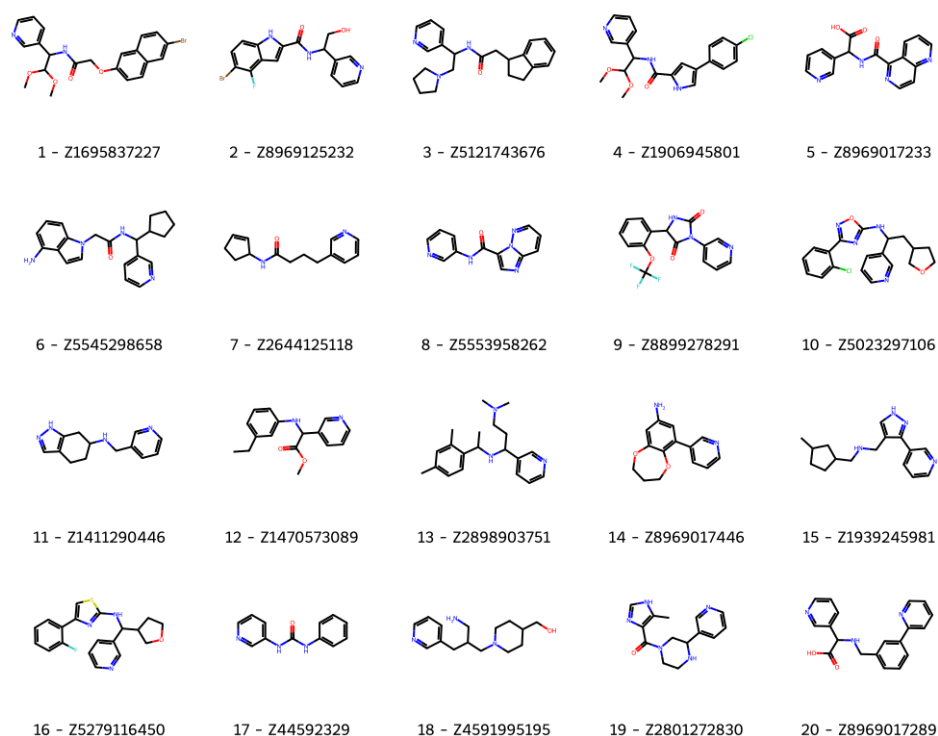


Figure S18: 2D structures of the Enamine compounds ordered, along with their compound number and Enamine IDs. Note that compound **17** is a control compound taken from a previous study²¹.

Experimental

Protein Production and Purification

Recombinant His-tagged Mpro was produced as described³⁴⁷ in *E. coli* BL21 (DE3) containing pGEX-6P-1 Mpro plasmid. A 50 mL starter culture was inoculated and grown with carbenicillin (100 μ g/mL) in LB broth for 8 hours at 37 °C with shaking (200 RPM). The expression media (1 L Formedium LB autoinduction media + 10 mL glycerol in a 2.5 L baffled Erlenmeyer flask) was inoculated with 10 mL of the starter culture and grown at 37 °C with carbenicillin (100 μ g/mL) at 200 RPM until reaching an OD₆₀₀ of 0.6. Cells were further incubated for 16 h at 18°C, harvested by centrifugation (8000 g, 10 min) and pellets were frozen at -80 °C. The frozen pellet (12 g) was re-suspended in lysis buffer [50 mM Tris, 300 mM NaCl, pH 8] before being lysed by sonication for 6 mins (5s on, 20s off cycle, 40 % AMP, Sonics VCX-500) twice before subsequent centrifugation (50,000 g, 30 min). The cell lysate was filtered in a 0.45 μ m syringe filter and added to a 3 mL bed volume Nickel Sepharose gravity flow column pre-equilibrated with lysis buffer. The column was washed with 72 mL wash buffer [50 mM Tris, 300 mM NaCl, 25 mM imidazole, pH 8] and eluted with 12 mL elution buffer [50 mM Tris, 300 mM NaCl, 500 mM imidazole, pH 8] and collected as 3 mL fractions. Protein-containing fractions

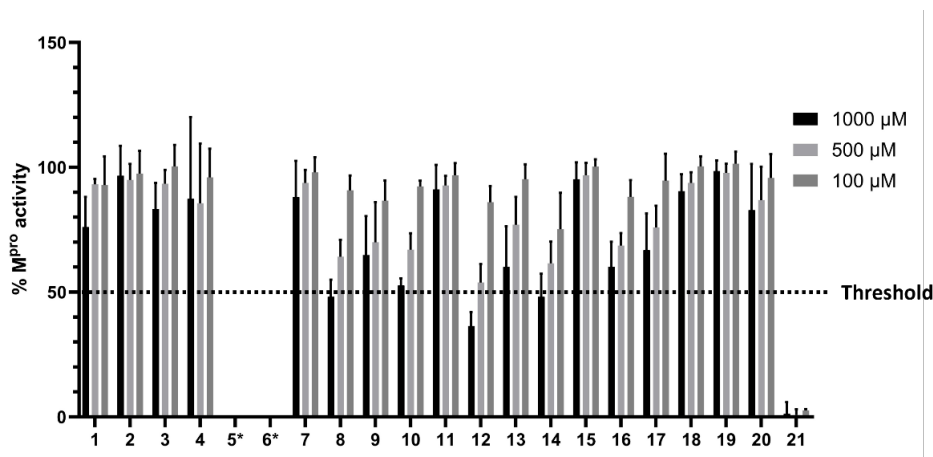


Figure S19: Initial compound screening for inhibition of Mpro enzyme activity. Compounds were tested for inhibition of Mpro catalytic activity at concentrations of 1000 μ M, 500 μ M and 100 μ M. Compounds **17** and **21** were included as controls. Compounds **12** and **14** reduced the Mpro activity below the threshold (≤ 50 % Mpro activity) at 1000 μ M and were selected for subsequent IC₅₀ analysis. **8** was not chosen for further analysis due to background auto-fluorescent activity. Data represented as mean \pm SD; 2 biological repeats consisting of 3 technical replicates. 10 consists of 1 biological repeat with 3 technical replicates. Conditions: Mpro (0.2 μ M) 12-hour pre-incubation with compounds, 20 μ M fluorescent substrate, 50 mM Tris-HCl (pH 7.3), 1 mM EDTA and temp: 25°C. Compounds **5** and **6** were excluded from the analysis due to poor solubility in assay conditions.

were combined and concentrated using a 10 kDa molecular weight cut off concentrator (Amicon Ultra), and subsequently purified by size-exclusion chromatography using a gel filtration column (HiLoad 16/600 Superdex 75pg) on AKTA Pure system in SEC buffer [50 mM Tris, 300 mM NaCl, pH 8]. His-tagged Mpro-containing fractions (>90 % purity by 10 % SDS PAGE gel) were concentrated (478.7 μ M), aliquoted and stored at -80 °C.

Fluorescent Activity Assay

Mpro fluorescent substrate peptide (MCA-AVLQSGFR-Lys(Dnp)-Lys-NH₂) was purchased from GL Biochem. All assays were performed as described⁴⁰³ in black 384-well microplates (Greiner Bio-One). Concentrations reported as used in the final assay volume of 30 μ L. Compound dilutions were prepared in assay buffer [50 mM Tris-HCl, 1mM EDTA, pH 7.3] with 3 % DMSO (1 % DMSO final). 10 μ L compound was incubated with 10 μ L Mpro (0.2 μ M final) for 30 min at 25 °C before addition of 10 μ L of the fluorescent substrate peptide (20 μ M final). Fluorescence (330 nm excitation / 390 nm emission) after 20.5 min (at linear range) at 25 °C (BMG Pherastar FSX) was used to calculate the IC₅₀ values. Background fluorescence of compounds **4**, **7**, **8**, **11**, **16** and **21** was subtracted from the raw datapoints. Datapoints were normalised to DMSO control (maximum Mpro activity) and no enzyme control and presented as ‘% Mpro activity’.

Analysis was performed on GraphPad Prism V10 fitted with the model 'log(inhibitor) vs. normalized response, variable slope'. All assays were performed twice in technical triplicates unless stated.

References

- [1] A. F. Ángyán, B. Szappanos, A. Perczel and Z. Gáspári, *BMC Structural Biology*, 2010, **10**, 39.
- [2] S. Wang, J. Witek, G. A. Landrum and S. Riniker, *Journal of Chemical Information and Modeling*, 2020, **60**, 2044–2058.
- [3] D. P. Kovács, J. H. Moore, N. J. Browning, I. Batatia, J. T. Horton, V. Kapil, W. C. Witt, I.-B. Magdau, D. J. Cole and G. Csányi, *MACE-OFF23: Transferable Machine Learning Force Fields for Organic Molecules*, 2023.
- [4] S.-L. J. Lahey, T. N. Thien Phuc and C. N. Rowley, *Journal of Chemical Information and Modeling*, 2020, **60**, 6258–6268.
- [5] G. Landrum, *RDKit: Open-source cheminformatics*, <http://www.rdkit.org/>.
- [6] A. T. McNutt, P. Francoeur, R. Aggarwal, T. Masuda, R. Meli, M. Ragoza, J. Sunseri and D. R. Koes, *J. Cheminf.*, 2021, **13**, 1–20.
- [7] J. Liang, V. Tsui, A. Van Abbema, L. Bao, K. Barrett, M. Beresini, L. Berezhkovskiy, W. S. Blair, C. Chang, J. Driscoll, C. Eigenbrot, N. Ghilardi, P. Gibbons, J. Halladay, A. Johnson, P. B. Kohli, Y. Lai, M. Liimatta, P. Mantik, K. Menghrajani, J. Murray, A. Sambrone, Y. Xiao, S. Shia, Y. Shin, J. Smith, S. Sohn, M. Stanley, M. Ultsch, B. Zhang, L. C. Wu and S. Magnuson, *European Journal of Medicinal Chemistry*, 2013, **67**, 175–187.
- [8] B. Baum, M. Mohamed, M. Zayed, C. Gerlach, A. Heine, D. Hangauer and G. Klebe, *Journal of Molecular Biology*, 2009, **390**, 56–69.
- [9] D. M. Goldstein, M. Soth, T. Gabriel, N. Dewdney, A. Kuglstatter, H. Arzeno, J. Chen, W. Bingenheimer, S. A. Dalrymple, J. Dunn, R. Farrell, S. Frauchiger, J. La Fargue, M. Ghate, B. Graves, R. J. Hill, F. Li, R. Litman, B. Loe, J. McIntosh, D. McWeeney, E. Papp, J. Park, H. F. Reese, R. T. Roberts, D. Rotstein, B. San Pablo, K. Sarma, M. Stahl, M.-L. Sung, R. T. Suttman, E. B. Sjogren, Y. Tan, A. Trejo, M. Welch, P. Weller, B. R. Wong and H. Zecic, *Journal of Medicinal Chemistry*, 2011, **54**, 2255–2265.
- [10] D. P. Wilson, Z.-K. Wan, W.-X. Xu, S. J. Kirincich, B. C. Follows, D. Joseph-McCarthy, K. Foreman, A. Moretto, J. Wu, M. Zhu, E. Binnun, Y.-L. Zhang, M. Tam, D. V. Erbe, J. Tobin, X. Xu, L. Leung, A. Shilling, S. Y. Tam, T. S. Mansour and J. Lee, *Journal of Medicinal Chemistry*, 2007, **50**, 4681–4698.

- [11] K. W. Hunt, A. W. Cook, R. J. Watts, C. T. Clark, G. Vigers, D. Smith, A. T. Metcalf, I. W. Gunawardana, M. Burkard, A. A. Cox, M. K. Geck Do, D. Dutcher, A. A. Thomas, S. Rana, N. C. Kallan, R. K. DeLisle, J. P. Rizzi, K. Regal, D. Sammond, R. Groneberg, M. Siu, H. Purkey, J. P. Lyssikatos, A. Marlow, X. Liu and T. P. Tang, *Journal of Medicinal Chemistry*, 2013, **56**, 3379–3403.
- [12] *WHAT IS CACHE | CACHE*, <https://cache-challenge.org/what-cache>.
- [13] S. L. Grimes and M. R. Denison, *Virus Research*, 2024, **346**, 199401.
- [14] J. A. Newman, A. Douangamath, S. Yadzani, Y. Yosaatmadja, A. Aimon, J. Brandão-Neto, L. Dunnett, T. Gorrie-stone, R. Skyner, D. Fearon, M. Schapira, F. von Delft and O. Gileadi, *Nature Communications*, 2021, **12**, 4848.
- [15] J. A. Newman, A. Douangamath, S. Yadzani, Y. Yosaatmadja, A. Aimon, J. Brandão-Neto, L. Dunnett, T. Gorrie-stone, R. Skyner, D. Fearon, M. Schapira, F. von Delft and O. Gileadi, *Nature Communications*, 2021, **12**, 4848.
- [16] C.-H. Zhang, E. A. Stone, M. Deshmukh, J. A. Ippolito, M. M. Ghahremanpour, J. Tirado-Rives, K. A. Spasov, S. Zhang, Y. Takeo, S. N. Kudalkar, Z. Liang, F. Isaacs, B. Lindenbach, S. J. Miller, K. S. Anderson and W. L. Jorgensen, *ACS Central Science*, 2021, **7**, 467–475.
- [17] M. Bieniek, B. Cree, R. Pirie, J. Horton, N. Tatum and D. Cole, *Commun. Chem.*, 2022, **5**, 136.
- [18] C.-H. Zhang, E. A. Stone, M. Deshmukh, J. A. Ippolito, M. M. Ghahremanpour, J. Tirado-Rives, K. A. Spasov, S. Zhang, Y. Takeo, S. N. Kudalkar, Z. Liang, F. Isaacs, B. Lindenbach, S. J. Miller, K. S. Anderson and W. L. Jorgensen, *ACS Central Science*, 2021, **7**, 467–475.
- [19] M. L. Bobby, D. Fearon, M. Ferla, M. Filep, L. Koekemoer, M. C. Robinson, The COVID Moonshot Consortium†, J. D. Chodera, A. A. Lee, N. London, A. Von Delft, F. Von Delft, H. Achdout, A. Aimon, D. S. Alonzi, R. Arbon, J. C. Aschenbrenner, B. H. Balcomb, E. Bar-David, H. Barr, A. Ben-Shmuel, J. Bennett, V. A. Bilenko, B. Borden, P. Boulet, G. R. Bowman, L. Brewitz, J. Brun, S. Bvnbs, M. Calmiano, A. Carbery, D. W. Carney, E. Cattermole, E. Chang, E. Chernyshenko, A. Clyde, J. E. Coffland, G. Cohen, J. C. Cole, A. Contini, L. Cox, T. I. Croll, M. Cvitkovic, S. De Jonghe, A. Dias, K. Donckers, D. L. Dotson, A. Douangamath, S. Duberstein, T. Dudgeon, L. E. Dunnett, P. Eastman, N. Erez, C. J. Eyermann, M. Fairhead, G. Fate, O. Fedorov, R. S. Fernandes, L. Ferrins, R. Foster, H. Foster, L. Fraisse, R. Gabizon, A. García-Sastre, V. O. Gawriljuk, P. Gehrtz, C. Gileadi, C. Giroud,

- W. G. Glass, R. C. Glen, I. Glinert, A. S. Godoy, M. Gorichko, T. Gorrie-Stone, E. J. Griffen, A. Haneef, S. Hassell Hart, J. Heer, M. Henry, M. Hill, S. Horrell, Q. Y. J. Huang, V. D. Huliak, M. F. D. Hurley, T. Israely, A. Jajack, J. Jansen, E. Jnoff, D. Jochmans, T. John, B. Kaminow, L. Kang, A. L. Kantsadi, P. W. Kenny, J. L. Kiappes, S. O. Kinakh, B. Kovar, T. Krojer, V. N. T. La, S. Laghnimi-Hahn, B. A. Lefker, H. Levy, R. M. Lithgo, I. G. Logvinenko, P. Lukacik, H. B. Macdonald, E. M. MacLean, L. L. Makower, T. R. Malla, P. G. Marples, T. Matviiuk, W. McCorkindale, B. L. McGovern, S. Melamed, K. P. Melnykov, O. Michurin, P. Miesen, H. Mikolajek, B. F. Milne, D. Minh, A. Morris, G. M. Morris, M. J. Morwitzer, D. Moustakas, C. E. Mowbray, A. M. Nakamura, J. B. Neto, J. Neyts, L. Nguyen, G. D. Noske, V. Oleinikovas, G. Oliva, G. J. Overheul, C. D. Owen, R. Pai, J. Pan, N. Paran, A. M. Payne, B. Perry, M. Pingle, J. Pinjari, B. Politi, A. Powell, V. Pšenák, I. Pulido, R. Puni, V. L. Rangel, R. N. Reddi, P. Rees, S. P. Reid, L. Reid, E. Resnick, E. G. Ripka, R. P. Robinson, J. Rodriguez-Guerra, R. Rosales, D. A. Rufa, K. Saar, K. S. Saikatendu, E. Salah, D. Schaller, J. Scheen, C. A. Schiffer, C. J. Schofield, M. Shafeev, A. Shaikh, A. M. Shaqra, J. Shi, K. Shurrush, S. Singh, A. Sittner, P. Sjö, R. Skyner, A. Smalley, B. Smeets, M. D. Smilova, L. J. Solmesky, J. Spencer, C. Strain-Damerell, V. Swamy, H. Tamir, J. C. Taylor, R. E. Tennant, W. Thompson, A. Thompson, S. Tomásio, C. W. E. Tomlinson, I. S. Tsurupa, A. Tumber, I. Vakonakis, R. P. Van Rij, L. Vangeel, F. S. Varghese, M. Vaschetto, E. B. Vitner, V. Voelz, A. Volkamer, M. A. Walsh, W. Ward, C. Weatherall, S. Weiss, K. M. White, C. F. Wild, K. D. Witt, M. Wittmann, N. Wright, Y. Yahalom-Ronen, N. K. Yilmaz, D. Zaidmann, I. Zhang, H. Zidane, N. Zitzmann and S. N. Zvornicanin, *Science*, 2023, **382**, eabo7201.
- [20] Q. Yang, W. Burchett, G. S. Steeno, S. Liu, M. Yang, D. L. Mobley and X. Hou, *J. Comput. Chem.*, 2020, **41**, 247–257.
- [21] M. L. Bobby, D. Fearon, M. Ferla, M. Filep, L. Koekemoer, M. C. Robinson, T. C. M. Consortium†, J. D. Chodera, A. A. Lee, N. London, A. von Delft, F. von Delft, H. Achdout, A. Aimon, D. S. Alonzi, R. Arbon, J. C. Aschenbrenner, B. H. Balcomb, E. Bar-David, H. Barr, A. Ben-Shmuel, J. Bennett, V. A. Bilenko, B. Borden, P. Boulet, G. R. Bowman, L. Brewitz, J. Brun, S. BVNBS, M. Calmiano, A. Carbery, D. W. Carney, E. Cattermole, E. Chang, E. Chernyshenko, A. Clyde, J. E. Coffland, G. Cohen, J. C. Cole, A. Contini, L. Cox, T. I. Croll, M. Cvitkovic, S. D. Jonghe, A. Dias, K. Donckers, D. L. Dotson, A. Douangamath, S. Duberstein, T. Dudgeon, L. E. Dunnett, P. Eastman, N. Erez, C. J. Eyermann, M. Fairhead, G. Fate, O. Fedorov, R. S. Fernandes, L. Ferrins, R. Foster, H. Foster, L. Fraisse, R. Gabizon, A. García-Sastre, V. O. Gawriljuk, P. Gehrtz, C. Gileadi, C. Giroud,

- W. G. Glass, R. C. Glen, I. Glinert, A. S. Godoy, M. Gorichko, T. Gorrie-Stone, E. J. Griffen, A. Haneef, S. H. Hart, J. Heer, M. Henry, M. Hill, S. Horrell, Q. Y. J. Huang, V. D. Huliak, M. F. D. Hurley, T. Israely, A. Jajack, J. Jansen, E. Jnoff, D. Jochmans, T. John, B. Kaminow, L. Kang, A. L. Kantsadi, P. W. Kenny, J. L. Kiappes, S. O. Kinakh, B. Kovar, T. Krojer, V. N. T. La, S. Laghnimi-Hahn, B. A. Lefker, H. Levy, R. M. Lithgo, I. G. Logvinenko, P. Lukacik, H. B. Macdonald, E. M. MacLean, L. L. Makower, T. R. Malla, P. G. Marples, T. Matviuk, W. McCorkindale, B. L. McGovern, S. Melamed, K. P. Melnykov, O. Michurin, P. Miesen, H. Mikolajek, B. F. Milne, D. Minh, A. Morris, G. M. Morris, M. J. Morwitzer, D. Moustakas, C. E. Mowbray, A. M. Nakamura, J. B. Neto, J. Neyts, L. Nguyen, G. D. Noske, V. Oleinikovas, G. Oliva, G. J. Overheul, C. D. Owen, R. Pai, J. Pan, N. Paran, A. M. Payne, B. Perry, M. Pingle, J. Pinjari, B. Politi, A. Powell, V. Pšenák, I. Pulido, R. Puni, V. L. Rangel, R. N. Reddi, P. Rees, S. P. Reid, L. Reid, E. Resnick, E. G. Ripka, R. P. Robinson, J. Rodriguez-Guerra, R. Rosales, D. A. Rufa, K. Saar, K. S. Saikatendu, E. Salah, D. Schaller, J. Scheen, C. A. Schiffer, C. J. Schofield, M. Shafeev, A. Shaikh, A. M. Shaqra, J. Shi, K. Shurrush, S. Singh, A. Sittner, P. Sjö, R. Skyner, A. Smalley, B. Smeets, M. D. Smilova, L. J. Solmesky, J. Spencer, C. Strain-Damerell, V. Swamy, H. Tamir, J. C. Taylor, R. E. Tennant, W. Thompson, A. Thompson, S. Tomásio, C. W. E. Tomlinson, I. S. Tsurupa, A. Tumber, I. Vakonakis, R. P. van Rij, L. Vangeel, F. S. Varghese, M. Vaschetto, E. B. Vitner, V. Voelz, A. Volkamer, M. A. Walsh, W. Ward, C. Weatherall, S. Weiss, K. M. White, C. F. Wild, K. D. Witt, M. Wittmann, N. Wright, Y. Yahalom-Ronen, N. K. Yilmaz, D. Zaidmann, I. Zhang, H. Zidane, N. Zitzmann and S. N. Zvornicanin, *Science*, 2023, **382**, eabo7201.
- [22] R. Nussinov and C.-J. Tsai, *Current Pharmaceutical Design*, **18**, 1311–1316.
- [23] A. V. Sadybekov and V. Katritch, *Nature*, 2023, **616**, 673–685.
- [24] J.-L. Reymond, L. Ruddigkeit, L. Blum and R. van Deursen, *WIREs Computational Molecular Science*, 2012, **2**, 717–733.
- [25] M. Bon, A. Bilsland, J. Bower and K. McAulay, *Molecular Oncology*, 2022, **16**, 3761–3777.
- [26] J. Müller, R. Klein, O. Tarkhanova, A. Gryniukova, P. Borysko, S. Merkl, M. Ruf, A. Neumann, M. Gastreich, Y. S. Moroz, G. Klebe and S. Glinca, *Journal of Medicinal Chemistry*, 2022, **65**, 15663–15678.
- [27] D. C. Swinney and J. Anthony, *Nature Reviews Drug Discovery*, 2011, **10**, 507–519.

- [28] J. G. Moffat, F. Vincent, J. A. Lee, J. Eder and M. Prunotto, *Nature Reviews Drug Discovery*, 2017, **16**, 531–543.
- [29] J. G. Moffat, F. Vincent, J. A. Lee, J. Eder and M. Prunotto, *Nat Rev Drug Discov*, 2017, **16**, 531–543.
- [30] F. Vincent, A. Nueda, J. Lee, M. Schenone, M. Prunotto and M. Mercola, *Nat Rev Drug Discov*, 2022, **21**, 899–914.
- [31] S. Morgan, P. Grootendorst, J. Lexchin, C. Cunningham and D. Greyson, *Health Policy*, 2011, **100**, 4–17.
- [32] I. J. d. S. Nascimento, T. M. de Aquino and E. F. da Silva-Júnior, *Letters in Drug Design & Discovery*, 2022, **19**, 951–955.
- [33] O. M. H. Salo-Ahen, I. Alanko, R. Bhadane, A. M. J. J. Bonvin, R. V. Honorato, S. Hossain, A. H. Juffer, A. Kabedev, M. Lahtela-Kakkonen, A. S. Larsen, E. Lescrinier, P. Marimuthu, M. U. Mirza, G. Mustafa, A. Nunes-Alves, T. Pantsar, A. Saadabadi, K. Singaravelu and M. Vanmeert, *Processes*, 2021, **9**, 71.
- [34] A. Gryniukova, F. Kaiser, I. Myziuk, D. Aliexsieieva, C. Leberecht, P. P. Heym, O. O. Tarkhanova, Y. S. Moroz, P. Borysko and V. J. Haupt, *Journal of Medicinal Chemistry*, 2023, **66**, 10241–10251.
- [35] A. D. Wade, A. P. Bhati, S. Wan and P. V. Coveney, *Journal of Chemical Theory and Computation*, 2022, **18**, 3972–3987.
- [36] M. M. Ghahremanpour, A. Saar, J. Tirado-Rives and W. L. Jorgensen, *Journal of Chemical Information and Modeling*, 2023, **63**, 5309–5318.
- [37] A. Hospital, J. R. Goñi, M. Orozco and J. L. Gelpí, *Advances and Applications in Bioinformatics and Chemistry*, 2015, **8**, 37–47.
- [38] B. J. Alder and T. E. Wainwright, *The Journal of Chemical Physics*, 1957, **27**, 1208–1209.
- [39] J. A. Stevens, F. Grünewald, P. A. M. van Tilburg, M. König, B. R. Gilbert, T. A. Brier, Z. R. Thornburg, Z. Luthey-Schulten and S. J. Marrink, *Frontiers in Chemistry*, 2023, **11**, year.
- [40] Anton 3 | *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, <https://dl.acm.org/doi/abs/10.1145/3458817.3487397>.

- [41] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti and G. Csányi, *Chem. Rev.*, 2021, **121**, 10073–10141.
- [42] P. Karande, B. Gallagher and T. Y.-J. Han, *Chemistry of Materials*, 2022, **34**, 7650–7665.
- [43] R. A. Shenvi, *ACS Cent. Sci.*, 2024, **10**, 519–528.
- [44] N. Fay, C. Kouklovsky and A. de la Torre, *ACS Organic & Inorganic Au*, 2023, **3**, 350–363.
- [45] J. G. Mahdi, A. J. Mahdi, A. J. Mahdi and I. D. Bowen, *Cell Proliferation*, 2006, **39**, 147–155.
- [46] J. W. Bennett and K.-T. Chung, in *Advances in Applied Microbiology*, Academic Press, 2001, vol. 49, pp. 163–184.
- [47] P. W. Kenny, *J. Med. Chem.*, 2022, **65**, 14261–14275.
- [48] C. Laurence and M. Berthelot.
- [49] M. J. Waring, *Expert Opinion on Drug Discovery*, 2010, **5**, 235–248.
- [50] T. P. Silverstein, *ChemTexts*, 2020, **6**, 26.
- [51] A. M. Birch, P. W. Kenny, I. Simpson and P. R. O. Whittamore, *Bioorganic & Medicinal Chemistry Letters*, 2009, **19**, 850–853.
- [52] J. F. Darby, A. P. Hopkins, S. Shimizu, S. M. Roberts, J. A. Brannigan, J. P. Turkenburg, G. H. Thomas, R. E. Hubbard and M. Fischer, *J. Am. Chem. Soc.*, 2019, **141**, 15818–15826.
- [53] F. Spyrakis, M. H. Ahmed, A. S. Bayden, P. Cozzini, A. Mozzarelli and G. E. Kellogg, *J. Med. Chem.*, 2017, **60**, 6781–6827.
- [54] H. S. Frank and M. W. Evans, *The Journal of Chemical Physics*, 1945, **13**, 507–532.
- [55] B. Kronberg, *Current Opinion in Colloid & Interface Science*, 2016, **22**, 14–22.
- [56] A. E. Modell, S. L. Blosser and P. S. Arora, *Trends in pharmacological sciences*, 2016, **37**, 702–713.
- [57] A. T. García-Sosa, R. L. Mancera and P. M. Dean, *J Mol Model*, 2003, **9**, 172–182.
- [58] M. Schauperl, M. Podewitz, T. S. Ortner, F. Waibl, A. Thoeny, T. Loerting and K. R. Liedl, *Scientific Reports*, 2017, **7**, 11901.

- [59] J. W. Pitera, M. Falta and W. F. van Gunsteren, *Biophysical Journal*, 2001, **80**, 2546–2555.
- [60] H. Meyer, 1899.
- [61] C. A. S. Bergström and P. Larsson, *International Journal of Pharmaceutics*, 2018, **540**, 185–193.
- [62] J. Gasteiger and M. Marsili, *Tetrahedron*, 1980, **36**, 3219–3228.
- [63] H. Moriwaki, Y.-S. Tian, N. Kawashita and T. Takagi, *Journal of Cheminformatics*, 2018, **10**, 4.
- [64] C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Advanced Drug Delivery Reviews*, 1997, **23**, 3–25.
- [65] R. Roskoski, *Pharmacological Research*, 2023, **191**, 106774.
- [66] J. E. Ladbury, G. Klebe and E. Freire, *Nature Reviews. Drug Discovery*, 2010, **9**, 23–27.
- [67] M.-Q. Zhang and B. Wilkinson, *Current Opinion in Biotechnology*, 2007, **18**, 478–488.
- [68] S. Evoli, D. L. Mobley, R. Guzzi and B. Rizzuti, *Physical Chemistry Chemical Physics*, 2016, **18**, 32358–32368.
- [69] A. A. Seyhan, *Translational Medicine Communications*, 2019, **4**, 18.
- [70] Y. Yang, O. Engkvist, A. Llinàs and H. Chen, *Journal of Medicinal Chemistry*, 2012, **55**, 3667–3677.
- [71] U. Norinder and C. A. S. Bergström, **1**, 920–937.
- [72] I. Aliagas, A. Gobbi, M.-L. Lee and B. D. Sellers, *Journal of Computer-Aided Molecular Design*, 2022, **36**, 253–262.
- [73] A. F. Stepan, D. P. Walker, J. Bauman, D. A. Price, T. A. Baillie, A. S. Kalgutkar and M. D. Aleo, *Chemical Research in Toxicology*, 2011, **24**, 1345–1410.
- [74] J. B. Baell and J. W. M. Nissink, *ACS Chemical Biology*, 2018, **13**, 36–44.
- [75] S. J. Capuzzi, E. N. Muratov and A. Tropsha, *Journal of Chemical Information and Modeling*, 2017, **57**, 417–427.

- [76] L. R. Vidler, I. A. Watson, B. J. Margolis, D. J. Cummins and M. Brunavs, *ACS Medicinal Chemistry Letters*, 2018, **9**, 792–796.
- [77] D. Boldini, L. Friedrich, D. Kuhn and S. A. Sieber, *ACS Cent. Sci.*, 2024, **10**, 823–832.
- [78] S. Jasial, E. Gilberg, T. Blaschke and J. Bajorath, *Journal of Medicinal Chemistry*, 2018, **61**, 10255–10264.
- [79] E. C. Harrington *et al.*, *Industrial quality control*, 1965, **21**, 494–498.
- [80] G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan and A. L. Hopkins, *Nature Chemistry*, 2012, **4**, 90–98.
- [81] M. Alnammi, S. Liu, S. S. Ericksen, G. E. Ananiev, A. F. Voter, S. Guo, J. L. Keck, F. M. Hoffmann, S. A. Wildman and A. Gitter, *Evaluating Scalable Supervised Learning for Synthesize-on-Demand Chemical Libraries*, 2023.
- [82] G. A. Landrum and S. Riniker, *J. Chem. Inf. Model.*, 2024, **64**, 1560–1567.
- [83] J. C. Baber and M. Feher, *Mini Rev Med Chem*, 2004, **4**, 681–692.
- [84] P. Ertl and A. Schuffenhauer, *Journal of Cheminformatics*, 2009, **1**, 8.
- [85] E. J. Corey and W. T. Wipke, *Science*, 1969, **166**, 178–192.
- [86] S. Genheden, A. Thakkar, V. Chadimová, J.-L. Reymond, O. Engkvist and E. Bjerrum, *J Cheminform*, 2020, **12**, 70.
- [87] C. Bilodeau, W. Jin, T. Jaakkola, R. Barzilay and K. F. Jensen, *WIREs Computational Molecular Science*, 2022, **12**, e1608.
- [88] J. H. Van Drie and L. Tong, *Bioorganic & Medicinal Chemistry Letters*, 2020, **30**, 127524.
- [89] J. S. Fraser, H. van den Bedem, A. J. Samelson, P. T. Lang, J. M. Holton, N. Echols and T. Alber, *Proceedings of the National Academy of Sciences*, 2011, **108**, 16247–16252.
- [90] R. F. de Freitas and M. Schapira, *MedChemComm*, 2017, **8**, 1970–1981.
- [91] R. P. D. Bank, *PDB Statistics*, <https://www.rcsb.org/stats>.
- [92] L. Maveyraud and L. Mourey, *Molecules*, 2020, **25**, 1030.

- [93] A. Srivastava, T. Nagai, A. Srivastava, O. Miyashita and F. Tama, *IJMS*, 2018, **19**, 3401.
- [94] J. W. Vant, D. Sarkar, J. Nguyen, A. T. Baker, J. V. Vermaas and A. Singharoy, *Biochemical Society Transactions*, 2022, **50**, 569–581.
- [95] B. S. Drown, K. Jooß, R. D. Melani, C. Lloyd-Jones, J. M. Camarillo and N. L. Kelleher, *Journal of Proteome Research*, 2022, **21**, 1299–1310.
- [96] I. Bludau, M. Frank, C. Dörig, Y. Cai, M. Heusel, G. Rosenberger, P. Picotti, B. C. Collins, H. Röst and R. Aebersold, *Nature Communications*, 2021, **12**, 3810.
- [97] C. H. Arrowsmith, J. E. Audia, C. Austin, J. Baell, J. Bennett, J. Blagg, C. Bountra, P. E. Brennan, P. J. Brown, M. E. Bunnage, C. Buser-Doepner, R. M. Campbell, A. J. Carter, P. Cohen, R. A. Copeland, B. Cravatt, J. L. Dahlin, D. Dhanak, A. M. Edwards, M. Frederiksen, S. V. Frye, N. Gray, C. E. Grimshaw, D. Hepworth, T. Howe, K. V. M. Huber, J. Jin, S. Knapp, J. D. Kotz, R. G. Kruger, D. Lowe, M. M. Mader, B. Marsden, A. Mueller-Fahrnow, S. Müller, R. C. O'Hagan, J. P. Overington, D. R. Owen, S. H. Rosenberg, R. Ross, B. Roth, M. Schapira, S. L. Schreiber, B. Shoichet, M. Sundström, G. Superti-Furga, J. Taunton, L. Toledo-Sherman, C. Walpole, M. A. Walters, T. M. Willson, P. Workman, R. N. Young and W. J. Zuercher, *Nature Chemical Biology*, 2015, **11**, 536–541.
- [98] L. M. Smith, J. N. Agar, J. Chamot-Rooke, P. O. Danis, Y. Ge, J. A. Loo, L. Pašatolić, Y. O. Tsybin, N. L. Kelleher and THE CONSORTIUM FOR TOP-DOWN PROTEOMICS, *Science Advances*, 2021, **7**, eabk0734.
- [99] Y. J. Edwards, A. E. Lobley, M. M. Pentony and D. T. Jones, *Genome Biology*, 2009, **10**, R50.
- [100] N. Perdigão, A. C. Rosa and S. I. O'Donoghue, *BioData Mining*, 2017, **10**, 24.
- [101] A. Kim and M. S. Cohen, *Expert Opinion on Drug Discovery*, 2016, **11**, 907–916.
- [102] C. A. Shepherd, A. L. Hopkins and I. Navratilova, *Progress in Biophysics and Molecular Biology*, 2014, **116**, 113–123.
- [103] D. J. Wood, J. D. Lopez-Fernandez, L. E. Knight, I. Al-Khawaldeh, C. Gai, S. Lin, M. P. Martin, D. C. Miller, C. Cano, J. A. Endicott, I. R. Hardcastle, M. E. M. Noble and M. J. Waring, *J. Med. Chem.*, 2019, **62**, 3741–3752.
- [104] J. D. Bauman, J. J. E. K. Harrison and E. Arnold, *IUCrJ*, 2016, **3**, 51–60.
- [105] R. Wang, Y. Gao and L. Lai, *Molecular modeling annual*, 2000, **6**, 498–516.

- [106] G. Schneider and U. Fechner, *Nature Reviews Drug Discovery*, 2005, **4**, 649–663.
- [107] N. Chéron, N. Jasty and E. I. Shakhnovich, *Journal of Medicinal Chemistry*, 2016, **59**, 4171–4188.
- [108] A. V. Ishchenko and E. I. Shakhnovich, *Journal of Medicinal Chemistry*, 2002, **45**, 2770–2780.
- [109] C. Perez, D. Soler, R. Soliva and V. Guallar, *Journal of Chemical Information and Modeling*, 2020, **60**, 1728–1736.
- [110] L. Hefke, K. Hiesinger, W. F. Zhu, J. S. Kramer and E. Proschak, *ACS Medicinal Chemistry Letters*, 2020, **11**, 1244–1249.
- [111] U. Garscha, E. Romp, S. Pace, A. Rossi, V. Temml, D. Schuster, S. König, J. Gerstmeier, S. Liening, M. Werner, H. Atze, S. Wittmann, C. Weinigel, S. Rummler, G. K. Scriba, L. Sautebin and O. Werz, *Scientific Reports*, 2017, **7**, 9398.
- [112] R. E. Carhart, D. H. Smith and R. Venkataraghavan, *Journal of Chemical Information and Computer Sciences*, 1985, **25**, 64–73.
- [113] Z. Deng, C. Chuaqui and J. Singh, *Journal of Medicinal Chemistry*, 2004, **47**, 337–344.
- [114] M. Nazaré, H. Matter, D. W. Will, M. Wagner, M. Urmann, J. Czech, H. Schreuder, A. Bauer, K. Ritter and V. Wehner, *Angewandte Chemie International Edition*, 2012, **51**, 905–911.
- [115] O. Ichihara, J. Barker, R. J. Law and M. Whittaker, *Molecular Informatics*, 2011, **30**, 298–306.
- [116] L. E. Burgess, B. J. Newhouse, P. Ibrahim, J. Rizzi, M. A. Kashem, A. Hartman, B. J. Brandhuber, C. D. Wright, D. S. Thomson, G. P. A. Vigers and K. Koch, *Proceedings of the National Academy of Sciences*, 1999, **96**, 8348–8352.
- [117] D. C. Thompson, R. Aldrin Denny, R. Nilakantan, C. Humblet, D. Joseph-McCarthy and E. Feyfant, *Journal of Computer-Aided Molecular Design*, 2008, **22**, 761.
- [118] P. Pfeffer, T. Fober, E. Hüllermeier and G. Klebe, *Journal of Chemical Information and Modeling*, 2010, **50**, 1644–1659.
- [119] L. Batiste, A. Unzue, A. Dolbois, F. Hassler, X. Wang, N. Deearain, J. Zhu, D. Spiliotopoulos, C. Nevado and A. Caffisch, *ACS Central Science*, 2018, **4**, 180–188.

- [120] S. Ruiz-Carmona, D. Alvarez-Garcia, N. Foloppe, A. B. Garmendia-Doval, S. Juhos, P. Schmidtke, X. Barril, R. E. Hubbard and S. D. Morley, *PLOS Computational Biology*, 2014, **10**, e1003571.
- [121] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, *Journal of Computational Chemistry*, 1983, **4**, 187–217.
- [122] A. D. MacKerell, M. Feig and C. L. Brooks, *Journal of the American Chemical Society*, 2004, **126**, 698–699.
- [123] F. Imrie, A. R. Bradley, M. van der Schaar and C. M. Deane, *Journal of Chemical Information and Modeling*, 2020, **60**, 1983–1995.
- [124] F. Imrie, T. E. Hadfield, A. R. Bradley and C. M. Deane, *Chemical Science*, 2021, **12**, 14577–14589.
- [125] Y. Yang, S. Zheng, S. Su, C. Zhao, J. Xu and H. Chen, *Chemical Science*, 2020, **11**, 8312–8322.
- [126] S. Urbaczek, A. Kolodzik, J. R. Fischer, T. Lippert, S. Heuser, I. Groth, T. Schulz-Gasch and M. Rarey, *Journal of Chemical Information and Modeling*, 2011, **51**, 3199–3207.
- [127] N.-O. Friedrich, C. De Bruyn Kops, F. Flachsenberg, K. Sommer, M. Rarey and J. Kirchmair, *Journal of Chemical Information and Modeling*, 2017, **57**, 2719–2728.
- [128] N. Brown, M. Fiscato, M. H. Segler and A. C. Vaucher, *Journal of Chemical Information and Modeling*, 2019, **59**, 1096–1108.
- [129] G. Corso, H. Stärk, B. Jing, R. Barzilay and T. Jaakkola, *DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking*, 2023, <http://arxiv.org/abs/2210.01776>, arXiv:2210.01776 [q-bio].
- [130] A. Tropsha, *Molecular Informatics*, 2010, **29**, 476–488.
- [131] H. Kubinyi, *Quantitative Structure-Activity Relationships*, 1988, **7**, 121–133.
- [132] H. Chen, L. Carlsson, M. Eriksson, P. Varkonyi, U. Norinder and I. Nilsson, *Journal of Chemical Information and Modeling*, 2013, **53**, 1324–1336.
- [133] Y. C. Martin, *WIREs Computational Molecular Science*, 2012, **2**, 435–442.
- [134] T. A. Soares, A. Nunes-Alves, A. Mazzolari, F. Ruggiu, G.-W. Wei and K. Merz, *Journal of Chemical Information and Modeling*, 2022, **62**, 5317–5320.

- [135] K. A. Scott, N. Ropek, B. Melillo, S. L. Schreiber, B. F. Cravatt and E. V. Vinogradova, *Current Research in Chemical Biology*, 2022, **2**, 100028.
- [136] G. M. Maggiora, *J. Chem. Inf. Model.*, 2006, **46**, 1535–1535.
- [137] M. V. Sabando, I. Ponzoni, E. E. Milios and A. J. Soto, *Briefings in Bioinformatics*, 2022, **23**, bbab365.
- [138] M. Cruz-Monteagudo, J. L. Medina-Franco, Y. Pérez-Castillo, O. Nicolotti, M. N. D. S. Cordeiro and F. Borges, *Drug Discovery Today*, 2014, **19**, 1069–1080.
- [139] J. Delaney, *Drug Discovery Today*, 2009, **14**, 198–207.
- [140] A. H. Cheng, A. Cai, S. Miret, G. Malkomes, M. Phielipp and A. Aspuru-Guzik, *Digital Discovery*, 2023, **2**, 748–758.
- [141] A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland and J. Laufer, *Journal of Chemical Information and Computer Sciences*, 1992, **32**, 244–255.
- [142] D. Rogers and M. Hahn, *Journal of Chemical Information and Modeling*, 2010, **50**, 742–754.
- [143] M. R. Keyvanpour and M. B. Shirzad, *Current Drug Discovery Technologies*, 2021, **18**, 17–30.
- [144] A. Lai, J. Schaub, C. Steinbeck and E. L. Schymanski, *Journal of Cheminformatics*, 2022, **14**, 85.
- [145] R. Guha and J. H. Van Drie, *Journal of Chemical Information and Modeling*, 2008, **48**, 646–658.
- [146] D. D. Nguyen, Z. Cang and G.-W. Wei, *Physical Chemistry Chemical Physics*, 2020, **22**, 4343–4367.
- [147] T. S. Rush, J. A. Grant, L. Mosyak and A. Nicholls, *Journal of Medicinal Chemistry*, 2005, **48**, 1489–1495.
- [148] N. J. Moerke, *Current Protocols in Chemical Biology*, 2009, **1**, 1–15.
- [149] D. M. Jameson and G. Mocz, in *Protein-Ligand Interactions: Methods and Applications*, ed. G. Ulrich Nienhaus, Humana Press, Totowa, NJ, 2005, pp. 301–322.

- [150] M. D. Hall, A. Yasgar, T. Peryea, J. C. Braisted, A. Jadhav, A. Simeonov and N. P. Coussens, *Methods and Applications in Fluorescence*, 2016, **4**, 022001.
- [151] S. P. Yadav, S. Bergqvist, M. L. Doyle, T. A. Neubert and A. P. Yamniuk, *Journal of Biomolecular Techniques : JBT*, 2012, **23**, 94–100.
- [152] A. Olaru, C. Bala, N. Jaffrezic-Renault and H. Y. Aboul-Enein, *Critical Reviews in Analytical Chemistry*, 2015, **45**, 97–105.
- [153] P. Singh, *Sensors and Actuators B: Chemical*, 2016, **229**, 110–130.
- [154] C. A. Wartchow, F. Podlaski, S. Li, K. Rowan, X. Zhang, D. Mark and K.-S. Huang, *Journal of Computer-Aided Molecular Design*, 2011, **25**, 669–676.
- [155] G. Carleo and M. Troyer, *Science*, 2017, **355**, 602–606.
- [156] A. Albouy, H. E. Cabral and A. A. Santos, *Celestial Mechanics and Dynamical Astronomy*, 2012, **113**, 369–375.
- [157] G. F. von Rudorff, *The Journal of Chemical Physics*, 2021, **155**, 224103.
- [158] S. F. Sousa, A. J. M. Ribeiro, R. P. P. Neves, N. F. Brás, N. M. F. S. A. Cerqueira, P. A. Fernandes and M. J. Ramos, *WIREs Computational Molecular Science*, 2017, **7**, e1281.
- [159] L. L. Foldy, *Journal of Mathematical Physics*, 1962, **3**, 531–539.
- [160] J. C. Slater, *Physical Review*, 1951, **81**, 385–390.
- [161] N. Argaman and G. Makov, *American Journal of Physics*, 2000, **68**, 69–79.
- [162] E. S. Kryachko and E. V. Ludeña, *Physics Reports*, 2014, **544**, 123–239.
- [163] J. T. Horton, A. E. A. Allen, L. S. Dodda and D. J. Cole, *Journal of Chemical Information and Modeling*, 2019, **59**, 1366–1381.
- [164] M. Bursch, J.-M. Mewes, A. Hansen and S. Grimme, *Angewandte Chemie International Edition*, 2022, **61**, e202205735.
- [165] P. K. Behara, H. Jang, J. T. Horton, T. Gokey, D. L. Dotson, S. Boothroyd, C. I. Bayly, D. J. Cole, L.-P. Wang and D. L. Mobley, *The Journal of Physical Chemistry B*, 2024, **128**, 7888–7902.
- [166] Z. N. Gerek and J. R. Elliott, *Industrial & Engineering Chemistry Research*, 2010, **49**, 3411–3423.

- [167] P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L.-P. Wang, D. Shukla, T. Tye, M. Houston, T. Stich, C. Klein, M. R. Shirts and V. S. Pande, *Journal of Chemical Theory and Computation*, 2013, **9**, 461–469.
- [168] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark and H. J. C. Berendsen, *Journal of Computational Chemistry*, 2005, **26**, 1701–1718.
- [169] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dulak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng and K. W. Jacobsen, *Journal of Physics: Condensed Matter*, 2017, **29**, 273002.
- [170] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks and V. S. Pande, *PLOS Computational Biology*, 2017, **13**, e1005659.
- [171] M. De Vivo, M. Masetti, G. Bottegoni and A. Cavalli, *Journal of Medicinal Chemistry*, 2016, **59**, 4035–4061.
- [172] N. Bou-Rabee, *Entropy*, 2014, **16**, 138–162.
- [173] L. Verlet, *Physical Review*, 1967, **159**, 98–103.
- [174] B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, *Journal of Computational Chemistry*, 1997, **18**, 1463–1472.
- [175] V. Kräutler, W. F. van Gunsteren and P. H. Hünenberger, *Journal of Computational Chemistry*, 2001, **22**, 501–508.
- [176] L. W. Votapka, B. R. Jagger, A. L. Heyneman and R. E. Amaro, *The Journal of Physical Chemistry B*, 2017, **121**, 3597–3606.
- [177] S. C. Gill, N. M. Lim, P. B. Grinaway, A. S. Rustenburg, J. Fass, G. A. Ross, J. D. Chodera and D. L. Mobley, *The Journal of Physical Chemistry B*, 2018, **122**, 5579–5598.
- [178] E. Braun, J. Gilmer, H. B. Mayes, D. L. Mobley, J. I. Monroe, S. Prasad and D. M. Zuckerman, *Living Journal of Computational Molecular Science*, 2019, **1**, 5957–5957.

- [179] S.-W. Chiu, M. Clark, S. Subramaniam and E. Jakobsson, *Journal of Computational Chemistry*, 2000, **21**, 121–131.
- [180] Q. Ke, X. Gong, S. Liao, C. Duan and L. Li, *Journal of Molecular Liquids*, 2022, **365**, 120116.
- [181] E. Braun, S. M. Moosavi and B. Smit, *Journal of Chemical Theory and Computation*, 2018, **14**, 5262–5272.
- [182] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, *The Journal of Chemical Physics*, 1984, **81**, 3684–3690.
- [183] J. E. Basconi and M. R. Shirts, *Journal of Chemical Theory and Computation*, 2013, **9**, 2887–2899.
- [184] L. Wang, P. K. Behara, M. W. Thompson, T. Gokey, Y. Wang, J. R. Wagner, D. J. Cole, M. K. Gilson, M. R. Shirts and D. L. Mobley, *The Journal of Physical Chemistry B*, 2024, **128**, 7043–7067.
- [185] D. L. Mobley, C. C. Bannan, A. Rizzi, C. I. Bayly, J. D. Chodera, V. T. Lim, N. M. Lim, K. A. Beauchamp, D. R. Slochow, M. R. Shirts, M. K. Gilson and P. K. Eastman, *J. Chem. Theory Comput.*, 2018, **14**, 6076–6092.
- [186] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, *Journal of the American Chemical Society*, 1995, **117**, 5179–5197.
- [187] D. L. Veenstra, D. M. Ferguson and P. A. Kollman, *Journal of Computational Chemistry*, 1992, **13**, 971–978.
- [188] Y. Qiu, D. G. A. Smith, S. Boothroyd, H. Jang, D. F. Hahn, J. Wagner, C. C. Bannan, T. Gokey, V. T. Lim, C. D. Stern, A. Rizzi, B. Tjanaka, G. Tresadern, X. Lucas, M. R. Shirts, M. K. Gilson, J. D. Chodera, C. I. Bayly, D. L. Mobley and L.-P. Wang, *Journal of Chemical Theory and Computation*, 2021, **17**, 6262–6280.
- [189] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *Journal of Computational Chemistry*, 2004, **25**, 1157–1174.
- [190] S. Boothroyd, L.-P. Wang, D. L. Mobley, J. D. Chodera and M. R. Shirts, *Journal of Chemical Theory and Computation*, 2022, **18**, 3566–3576.
- [191] S. J. Rukmani, G. Kupgan, D. M. Anstine and C. M. Colina, *Molecular Simulation*, 2019, **45**, 310–321.

- [192] K. Vanommeslaeghe, O. Guvench and J. Alexander D. MacKerell, *Current pharmaceutical design*, 2014, **20**, 3281.
- [193] R. S. Paton and J. M. Goodman, *Journal of Chemical Information and Modeling*, 2009, **49**, 944–955.
- [194] S. Riniker, *Journal of Chemical Information and Modeling*, 2018, **58**, 565–578.
- [195] J. T. Horton, S. Boothroyd, P. K. Behara, D. L. Mobley and D. J. Cole, *Digital Discovery*, 2023, **2**, 1178–1187.
- [196] A. Jakalian, D. B. Jack and C. I. Bayly, *Journal of Computational Chemistry*, 2002, **23**, 1623–1641.
- [197] T. Husch, A. C. Vaucher and M. Reiher, *International Journal of Quantum Chemistry*, 2018, **118**, e25799.
- [198] Y. Wang, I. Pulido, K. Takaba, B. Kaminow, J. Scheen, L. Wang and J. D. Chodera, *The Journal of Physical Chemistry A*, 2024, **128**, 4160–4167.
- [199] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *The Journal of Chemical Physics*, 1983, **79**, 926–935.
- [200] L. Wickstrom, A. Okur and C. Simmerling, *Biophysical Journal*, 2009, **97**, 853–856.
- [201] C. Tian, K. Kasavajhala, K. A. A. Belfon, L. Raguetta, H. Huang, A. N. Migués, J. Bickel, Y. Wang, J. Pincay, Q. Wu and C. Simmerling, *Journal of Chemical Theory and Computation*, 2020, **16**, 528–552.
- [202] X. Liu, D. Shi, S. Zhou, H. Liu, H. Liu and X. Yao, *Expert Opinion on Drug Discovery*, 2018, **13**, 23–37.
- [203] A. Gupta and H.-X. Zhou, *Journal of Chemical Information and Modeling*, 2021, **61**, 4236–4244.
- [204] C. R. Groom and F. H. Allen, *Angewandte Chemie International Edition*, 2014, **53**, 662–671.
- [205] S. Riniker and G. A. Landrum, *Journal of Chemical Information and Modeling*, 2015, **55**, 2562–2574.
- [206] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. I. Goddard and W. M. Skiff, *Journal of the American Chemical Society*, 1992, **114**, 10024–10035.

- [207] J.-P. Ebejer, G. M. Morris and C. M. Deane, *Journal of Chemical Information and Modeling*, 2012, **52**, 1146–1158.
- [208] A. Krizhevsky, I. Sutskever and G. E. Hinton, *Commun. ACM*, 2017, **60**, 84–90.
- [209] S. Shanmuganathan, in *Artificial Neural Network Modelling*, ed. S. Shanmuganathan and S. Samarasinghe, Springer International Publishing, Cham, 2016, pp. 1–14.
- [210] *Pattern Recognition and Machine Learning*.
- [211] S. Sidana, M. Trofimov, O. Horodnytskyi, C. Laclau, Y. Maximov and M.-R. Amini, *Data Mining and Knowledge Discovery*, 2021, **35**, 568–592.
- [212] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang and J. Tang, *AI Open*, 2023.
- [213] P. Xu, X. Ji, M. Li and W. Lu, *npj Computational Materials*, 2023, **9**, 1–15.
- [214] A. De Simone and T. Jacques, *The European Physical Journal C*, 2019, **79**, 289.
- [215] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro and Y. Zhang, *Sparks of Artificial General Intelligence: Early Experiments with GPT-4*, 2023.
- [216] H. Abdi, *Journal of Biological Systems*, 1994, **02**, 247–281.
- [217] H. Ma, A. Narayanaswamy, P. Riley and L. Li, *Science Advances*, 2022, **8**, eabq0279.
- [218] T. C. Terwilliger, D. Liebschner, T. I. Croll, C. J. Williams, A. J. McCoy, B. K. Poon, P. V. Afonine, R. D. Oeffner, J. S. Richardson, R. J. Read and P. D. Adams, *Nat Methods*, 2024, **21**, 110–116.
- [219] T. T. Duignan, *ACS Physical Chemistry Au*, 2024, **4**, 232–241.
- [220] T. B. Blank, S. D. Brown, A. W. Calhoun and D. J. Doren, *The Journal of Chemical Physics*, 1995, **103**, 4129–4137.
- [221] Y. Wang, K. Takaba, M. S. Chen, M. Wieder, Y. Xu, T. Zhu, J. Z. H. Zhang, A. Nagle, K. Yu, X. Wang, D. J. Cole, J. A. Rackers, K. Cho, J. G. Greener, P. Eastman, S. Martiniani and M. E. Tuckerman, *On the Design Space between Molecular Mechanics and Machine Learning Force Fields*, 2024.
- [222] M. I. Jordan and T. M. Mitchell, *Science*, 2015, **349**, 255–260.
- [223] S. Stocker, J. Gasteiger, F. Becker, S. Günnemann and J. T. Margraf, *Machine Learning: Science and Technology*, 2022, **3**, 045010.

- [224] J. S. Smith, O. Isayev and A. E. Roitberg, *Scientific Data*, 2017, **4**, 170193.
- [225] A. S. Christensen, L. A. Bratholm, F. A. Faber and O. Anatole von Lilienfeld, *The Journal of Chemical Physics*, 2020, **152**, 044107.
- [226] C. Middleton, B. F. E. Curchod and T. J. Penfold, *Physical Chemistry Chemical Physics*, 2024, **26**, 24477–24487.
- [227] P. Eastman, R. Galvelis, R. P. Peláez, C. R. A. Abreu, S. E. Farr, E. Gallicchio, A. Gorenko, M. M. Henry, F. Hu, J. Huang, A. Krämer, J. Michel, J. A. Mitchell, V. S. Pande, J. P. Rodrigues, J. Rodriguez-Guerra, A. C. Simmonett, S. Singh, J. Swails, P. Turner, Y. Wang, I. Zhang, J. D. Chodera, G. De Fabritiis and T. E. Markland, *The Journal of Physical Chemistry B*, 2024, **128**, 109–116.
- [228] A. Hofstetter, L. Bösel and S. Riniker, *Physical Chemistry Chemical Physics*, 2022, **24**, 22497–22512.
- [229] C. Devereux, J. S. Smith, K. K. Huddleston, K. Barros, R. Zubatyuk, O. Isayev and A. E. Roitberg, *J. Chem. Theory Comput.*, 2020, **16**, 4192–4202.
- [230] M. Goldey, A. Dutoi and M. Head-Gordon, *Physical Chemistry Chemical Physics*, 2013, **15**, 15869–15875.
- [231] T. A. Halgren, *Journal of Computational Chemistry*, 1996, **17**, 490–519.
- [232] J. J. P. Stewart, *Journal of Molecular Modeling*, 2007, **13**, 1173–1213.
- [233] J. Behler and M. Parrinello, *Physical Review Letters*, 2007, **98**, 146401.
- [234] P. Eastman, P. K. Behara, D. L. Dotson, R. Galvelis, J. E. Herr, J. T. Horton, Y. Mao, J. D. Chodera, B. P. Pritchard, Y. Wang, G. De Fabritiis and T. E. Markland, **10**, 11.
- [235] J. H. Moore, D. J. Cole and G. Csanyi, *Computing Hydration Free Energies of Small Molecules with First Principles Accuracy*, <https://arxiv.org/abs/2405.18171v2>, 2024.
- [236] C. Sutton, M. Boley, L. M. Ghiringhelli, M. Rupp, J. Vreeken and M. Scheffler, *Nature Communications*, 2020, **11**, 4428.
- [237] A. S. J. S. Mey, B. K. Allen, H. E. B. McDonald, J. D. Chodera, D. F. Hahn, M. Kuhn, J. Michel, D. L. Mobley, L. N. Naden, S. Prasad, A. Rizzi, J. Scheen, M. R. Shirts, G. Tresadern and H. Xu, *Living Journal of Computational Molecular Science*, 2020, **2**, 18378–18378.

- [238] S. Liu, Y. Wu, T. Lin, R. Abel, J. P. Redmann, C. M. Summa, V. R. Jaber, N. M. Lim and D. L. Mobley, *Journal of computer-aided molecular design*, 2013, **27**, 10.1007/s10822-013-9678-y.
- [239] S. Zhang, T. J. Giese, T.-S. Lee and D. M. York, *Journal of Chemical Theory and Computation*, 2024, **20**, 3935–3953.
- [240] N. Ferruz and G. De Fabritiis, *Molecular Informatics*, 2016, **35**, 216–226.
- [241] C. D. Christ, A. E. Mark and W. F. van Gunsteren, *Journal of Computational Chemistry*, 2010, **31**, 1569–1582.
- [242] C. H. Bennett, *Journal of Computational Physics*, 1976, **22**, 245–268.
- [243] M. R. Shirts and J. D. Chodera, *The Journal of Chemical Physics*, 2008, **129**, 124105.
- [244] G. A. Ross, C. Lu, G. Scarabelli, S. K. Albanese, E. Houang, R. Abel, E. D. Harder and L. Wang, *Communications Chemistry*, 2023, **6**, 1–12.
- [245] M. Bissaro, M. Sturlese and S. Moro, *Drug Discovery Today*, 2020, **25**, 1693–1701.
- [246] E. B. Lenselink, J. Louvel, A. F. Forti, J. P. D. van Veldhoven, H. de Vries, T. Mulder-Krieger, F. M. McRobb, A. Negri, J. Goose, R. Abel, H. W. T. van Vlijmen, L. Wang, E. Harder, W. Sherman, A. P. IJzerman and T. Beuming, *ACS Omega*, 2016, **1**, 293–304.
- [247] P. Procacci, *Journal of Molecular Graphics and Modelling*, 2017, **71**, 233–241.
- [248] J. H. Moore, C. Margreitter, J. P. Janet, O. Engkvist, B. L. de Groot and V. Gapsys, *Communications Chemistry*, 2023, **6**, 1–12.
- [249] D. P. Kovács, J. H. Moore, N. J. Browning, I. Batatia, J. T. Horton, V. Kapil, W. C. Witt, I.-B. Magdău, D. J. Cole and G. Csányi, *MACE-OFF23: Transferable Machine Learning Force Fields for Organic Molecules*, 2023.
- [250] J. Thompson, W. P. Walters, J. A. Feng, N. A. Pabon, H. Xu, M. Maser, B. B. Goldman, D. Moustakas, M. Schmidt and F. York, *Artificial Intelligence in the Life Sciences*, 2022, **2**, 100050.
- [251] E. Yuriev, M. Agostino and P. A. Ramsland, *Journal of Molecular Recognition*, 2011, **24**, 149–164.
- [252] V. Zoete, A. Grosdidier and O. Michielin, *Journal of Cellular and Molecular Medicine*, 2009, **13**, 238–248.

- [253] M. P. Repasky, M. Shelley and R. A. Friesner, *Current Protocols in Bioinformatics*, 2007, **18**, 8.12.1–8.12.36.
- [254] C. M. Venkatachalam, X. Jiang, T. Oldfield and M. Waldman, *Journal of Molecular Graphics and Modelling*, 2003, **21**, 289–307.
- [255] R. N. dos Santos, L. G. Ferreira and A. D. Andricopulo, in *Computational Drug Discovery and Design*, ed. M. Gore and U. B. Jagtap, Springer, New York, NY, 2018, pp. 31–50.
- [256] Corwin. Hansch and Toshio. Fujita, *Journal of the American Chemical Society*, 1964, **86**, 1616–1626.
- [257] C. Yang and Y. Zhang, *Journal of Chemical Information and Modeling*, 2021, **61**, 4630–4644.
- [258] J. Li, A. Fu and L. Zhang, *Interdisciplinary Sciences: Computational Life Sciences*, 2019, **11**, 320–328.
- [259] M. Wójcikowski, P. J. Ballester and P. Siedlecki, *Scientific Reports*, 2017, **7**, 46710.
- [260] M. Su, G. Feng, Z. Liu, Y. Li and R. Wang, *Journal of Chemical Information and Modeling*, 2020, **60**, 1122–1136.
- [261] H. Li, K.-H. Sze, G. Lu and P. J. Ballester, *WIREs Computational Molecular Science*, 2020, **10**, e1465.
- [262] J. Bao, X. He and J. Z. H. Zhang, *Journal of Chemical Information and Modeling*, 2021, **61**, 2231–2240.
- [263] P. G. Francoeur, T. Masuda, J. Sunseri, A. Jia, R. B. Iovanisci, I. Snyder and D. R. Koes, *Journal of chemical information and modeling*, 2020, **60**, 4200–4215.
- [264] M. Lovrić, T. Đuričić, H. T. N. Tran, H. Hussain, E. Lacić, M. A. Rasmussen and R. Kern, *Pharmaceuticals*, 2021, **14**, 758.
- [265] J. Yu, X. Li and M. Zheng, *Artificial Intelligence in the Life Sciences*, 2021, **1**, 100023.
- [266] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti and G. Csányi, *Chemical Reviews*, 2021, **121**, 10073–10141.
- [267] J. Wang, *Computing in Science & Engineering*, 2023, **25**, 4–11.

- [268] A. Wang, H. Liang, A. McDannald, I. Takeuchi and A. G. Kusne, *Oxford Open Materials Science*, 2022, **2**, itac006.
- [269] B. J. Bender, S. Gahbauer, A. Lutten, J. Lyu, C. M. Webb, R. M. Stein, E. A. Fink, T. E. Balius, J. Carlsson, J. J. Irwin *et al.*, *Nature protocols*, 2021, **16**, 4799–4832.
- [270] G. Schneider and U. Fechner, *Nat. Rev. Drug Discovery*, 2005, **4**, 649–663.
- [271] N. Chéron, N. Jasty and E. I. Shakhnovich, *J. Med. Chem.*, 2016, **59**, 4171–4188.
- [272] J. D. Durrant, R. E. Amaro and J. A. McCammon, *Chem. Biol. Drug Des.*, 2009, **73**, 168–178.
- [273] Y. Yuan, J. Pei and L. Lai, *Journal of Chemical Information and Modeling*, 2011, **51**, 1083–1091.
- [274] T. Sousa, J. Correia, V. Pereira and M. Rocha, *Journal of Chemical Information and Modeling*, 2021, **61**, 5343–5361.
- [275] G. Schneider and D. E. Clark, *Angew. Chem. Int. Ed.*, 2019, **58**, 10792–10803.
- [276] B. C. Pearce, D. R. Langley, J. Kang, H. Huang and A. Kulkarni, *J. Chem. Inf. Model.*, 2009, **49**, 1797–1809.
- [277] S. Cross and G. Cruciani, *J. Chem. Inf. Model.*, 2022, **2022**, year.
- [278] P. J. Goodford, *J. Med. Chem.*, 1985, **28**, 849–857.
- [279] H. Green, D. R. Koes and J. D. Durrant, *Chem. Sci.*, 2021, **12**, 8036–8047.
- [280] F. Imrie, T. E. Hadfield, A. R. Bradley and C. M. Deane, *bioRxiv*, 2021.
- [281] W. L. Jorgensen, *Accounts of Chemical Research*, 2009, **42**, 724–733.
- [282] Z. Cournia, B. Allen and W. Sherman, *Journal of Chemical Information and Modeling*, 2017, **57**, 2911–2937.
- [283] Z. Cournia, B. K. Allen, T. Beuming, D. A. Pearlman, B. K. Radak and W. Sherman, *Journal of Chemical Information and Modeling*, 2020, **60**, 4153–4169.
- [284] A. S. J. S. Mey, B. K. Allen, H. E. Bruce McDonald, J. D. Chodera, D. F. Hahn, M. Kuhn, J. Michel, D. L. Mobley, L. N. Naden, S. Prasad, A. Rizzi, J. Scheen, M. R. Shirts, G. Tresadern and H. Xu, *Living Journal of Computational Molecular Science*, 2020, **2**, 18378.
- [285] D. L. Mobley and M. K. Gilson, *Annual Review of Biophysics*, 2017, **46**, 531–558.

- [286] V. Gapsys, D. F. Hahn, G. Tresadern, D. L. Mobley, M. Rampp and B. L. de Groot, *Journal of Chemical Information and Modeling*, 2022, **62**, 1172–1177.
- [287] *Citations | Schrödinger*, 2022, <https://www.schrodinger.com/citations#Maestro>, [Online; accessed 4. Mar. 2022].
- [288] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng and T. E. Ferrin, *J. Comput. Chem.*, 2004, **25**, 1605–1612.
- [289] W. L. Jorgensen, J. Ruiz-Caro, J. Tirado-Rives, A. Basavapathruni, K. S. Anderson and A. D. Hamilton, *Bioorg. Med. Chem. Lett.*, 2006, **16**, 663–667.
- [290] W.-G. Lee, R. Gallardo-Macias, K. M. Frey, K. A. Spasov, M. Bollini, K. S. Anderson and W. L. Jorgensen, *Journal of the American Chemical Society*, 2013, **135**, 16705–16713.
- [291] P. Dziedzic, J. A. Cisneros, M. J. Robertson, A. A. Hare, N. E. Danford, R. H. G. Baxter and W. L. Jorgensen, *J. Am. Chem. Soc.*, 2015, **137**, 2996–3003.
- [292] The COVID Moonshot Consortium, *COVID Moonshot: Open Science Discovery of SARS-CoV-2 Main Protease Inhibitors by Combining Crowdsourcing, High-Throughput Experiments, Computational Simulations, and Machine Learning*, 2020, [Online; accessed 4. Mar. 2022].
- [293] P. C. D. Hawkins, A. G. Skillman, G. L. Warren, B. A. Ellingson and M. T. Stahl, *Journal of Chemical Information and Modeling*, 2010, **50**, 572–584.
- [294] choderalab, *perses*, 2022, <https://github.com/choderalab/perses>, [Online; accessed 4. Mar. 2022].
- [295] J. S. Smith, O. Isayev and A. E. Roitberg, *Chem. Sci.*, 2017, **8**, 3192–3203.
- [296] *Sire Molecular Simulation Framework*, 2021, <http://siremol.org>.
- [297] N. Rego and D. Koes, *Bioinformatics*, 2014, **31**, 1322–1324.
- [298] N. M. O’Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *Journal of Cheminformatics*, 2011, **3**, 33.
- [299] S. Riniker and G. A. Landrum, *Journal of Chemical Information and Modeling*, 2015, **55**, 2562–2574.
- [300] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard and W. M. Skiff, *J. Am. Chem. Soc.*, 1992, **114**, 10024–10035.

- [301] K. Takeuchi, R. Kunimoto and J. Bajorath, *Future Science OA*, 2021, **7**, FSO742.
- [302] C. Bouysset, *mols2grid - Interactive molecule viewer for 2D structures*, <https://github.com/cbouy/mols2grid>.
- [303] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks and V. S. Pande, *PLoS Comput. Biol.*, 2017, **13**, e1005659.
- [304] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, *J. Chem. Theory Comput.*, 2015, **11**, 3696–3713.
- [305] Y. Qiu, D. G. A. Smith, S. Boothroyd, H. Jang, D. F. Hahn, J. Wagner, C. C. Bannan, T. Gokey, V. T. Lim, C. D. Stern, A. Rizzi, B. Tjanaka, G. Tresadern, X. Lucas, M. R. Shirts, M. K. Gilson, J. D. Chodera, C. I. Bayly, D. L. Mobley and L.-P. Wang, *J. Chem. Theory Comput.*, 2021, **17**, 6262–6280.
- [306] D. J. Cole, L. Mones and G. Csányi, *Faraday Discuss.*, 2020, **224**, 247–264.
- [307] S.-L. J. Lahey and C. N. Rowley, *Chem. Sci.*, 2020, **11**, 2362–2368.
- [308] M. Ragoza, J. Hochuli, E. Idrobo, J. Sunseri and D. R. Koes, *Journal of Chemical Information and Modeling*, 2017, **57**, 942–957.
- [309] P. G. Francoeur, T. Masuda, J. Sunseri, A. Jia, R. B. Iovanisci, I. Snyder and D. R. Koes, *J. Chem. Inf. Model.*, 2020, **60**, 4200–4215.
- [310] D. R. Koes, M. P. Baumgartner and C. J. Camacho, *Journal of Chemical Information and Modeling*, 2013, **53**, 1893–1904.
- [311] O. Trott and A. J. Olson, *Journal of Computational Chemistry*, 2010, **31**, 455–461.
- [312] P. Ertl and A. Schuffenhauer, *Journal of Cheminformatics*, 2009, **1**, 8.
- [313] J. B. Baell and G. A. Holloway, *Journal of Medicinal Chemistry*, 2010, **53**, 2719–2740.
- [314] A. Jadhav, R. S. Ferreira, C. Klumpp, B. T. Mott, C. P. Austin, J. Inglese, C. J. Thomas, D. J. Maloney, B. K. Shoichet and A. Simeonov, *Journal of Medicinal Chemistry*, 2010, **53**, 37–51.
- [315] R. G. Doveston, P. Tosatti, M. Dow, D. J. Foley, H. Y. Li, A. J. Campbell, D. House, I. Churcher, S. P. Marsden and A. Nelson, *Organic & Biomolecular Chemistry*, 2014, **13**, 859–865.

- [316] R. Brenk, A. Schipani, D. James, A. Krasowski, I. H. Gilbert, J. Frearson and P. G. Wyatt, *Chemmedchem*, 2008, **3**, 435–444.
- [317] D. Sydow, A. Morger, M. Driller and A. Volkamer, *Journal of Cheminformatics*, 2019, **11**, 29.
- [318] W. L. Jorgensen, *Science*, 2004, **303**, 1813–1818.
- [319] D. F. Hahn, C. I. Bayly, H. E. B. Macdonald, J. D. Chodera, V. Gapsys, A. S. J. S. Mey, D. L. Mobley, L. P. Benito, C. E. M. Schindler, G. Tresadern and G. L. Warren, *arXiv*, 2021.
- [320] C. Bannwarth, S. Ehlert and S. Grimme, *Journal of Chemical Theory and Computation*, 2019, **15**, 1652–1671.
- [321] M. L. Samways, H. E. Bruce Macdonald and J. W. Essex, *J. Chem. Inf. Model.*, 2020, **60**, 4436–4441.
- [322] R. Abel, T. Young, R. Farid, B. J. Berne and R. A. Friesner, *J. Am. Chem. Soc.*, 2008, **130**, 2817–2831.
- [323] Y. Ge, D. C. Wych, M. L. Samways, M. E. Wall, J. W. Essex and D. L. Mobley, *J. Chem. Theory Comput.*, 2022, **18**, 1359–1381.
- [324] B. Webb and A. Sali, *Current Protocols in Bioinformatics*, 2016, **54**, 5.6.1–5.6.37.
- [325] L. Hedges, A. Mey, C. Laughton, F. Gervasio, A. Mulholland, C. Woods and J. Michel, *Journal of Open Source Software*, 2019, **4**, 1831.
- [326] L. Nelson, S. Bariami, C. Ringrose, J. T. Horton, V. Kurdekar, A. S. J. S. Mey, J. Michel and D. J. Cole, *Journal of Chemical Information and Modeling*, 2021, **61**, 2124–2130.
- [327] M. Kuhn, S. Firth-Clark, P. Tosco, A. S. J. S. Mey, M. Mackey and J. Michel, *J. Chem. Inf. Model.*, 2020, **60**, 3120–3130.
- [328] A. S. Mey, J. J. Jiménez and J. Michel, *J. Comput. Aided Mol. Des.*, 2018, **32**, 199–210.
- [329] J. F.-W. Chan, K.-H. Kok, Z. Zhu, H. Chu, K. K.-W. To, S. Yuan and K.-Y. Yuen, *Emerging Microbes & Infections*, 2020, **9**, 221–236.
- [330] Z. Deng, K. C. Lehmann, X. Li, C. Feng, G. Wang, Q. Zhang, X. Qi, L. Yu, X. Zhang, W. Feng, W. Wu, P. Gong, Y. Tao, C. C. Posthuma, E. J. Snijder, A. E. Gorbalenya and Z. Chen, *Nucleic Acids Research*, 2014, **42**, 3464–3477.

- [331] M. Kuzikov, J. Reinshagen, K. Wycisk, A. Corona, F. Esposito, P. Malune, C. Manelfi, D. Iaconis, A. Beccari, E. Tramontano, M. Nowotny, B. Windshügel, P. Gribbon and A. Zaliani, *Virus Research*, 2024, **343**, 199356.
- [332] B. Cosar, Z. Y. Karagulleoglu, S. Unal, A. T. Ince, D. B. Uncuoglu, G. Tuncer, B. R. Kilinc, Y. E. Ozkan, H. C. Ozkoc, I. N. Demir, A. Eker, F. Karagoz, S. Y. Simsek, B. Yasar, M. Pala, A. Demir, I. N. Atak, A. H. Mendi, V. U. Bengi, G. Cengiz Seval, E. Gunes Altuntas, P. Kilic and D. Demir-Dora, *Cytokine & Growth Factor Reviews*, 2022, **63**, 10–22.
- [333] M. A. White, W. Lin and X. Cheng, *The Journal of Physical Chemistry Letters*, 2020, **11**, 9144–9151.
- [334] The UniProt Consortium, *Nucleic Acids Research*, 2023, **51**, D523–D531.
- [335] T. A. Halgren, *Journal of Chemical Information and Modeling*, 2009, **49**, 377–389.
- [336] A. O. Adedeji, K. Singh, N. E. Calcaterra, M. L. DeDiego, L. Enjuanes, S. Weiss and S. G. Sarafianos, *Antimicrobial Agents and Chemotherapy*, 2012, **56**, 4718–4728.
- [337] J. A. Sommers, L. N. Loftus, M. P. Jones, R. A. Lee, C. E. Haren, A. J. Dumm and R. M. Brosh, *Journal of Biological Chemistry*, 2023, **299**, year.
- [338] Y. Otsuka, L. Zhang, H. Mou, J. Shumate, C. E. Kitzmiller, L. Scampavia, T. D. Bannister, M. Farzan, H. Choe and T. P. Spicer, *SLAS Discovery*, 2024, **29**, year.
- [339] Y. Otsuka, E. Kim, A. Krueger, J. Shumate, C. Wang, B. Bdiri, S. Ullah, H. Park, L. Scampavia, T. D. Bannister, D. Chung and T. P. Spicer, *SLAS Discovery*, 2024, 100180.
- [340] A. Nicholls, G. B. McGaughey, R. P. Sheridan, A. C. Good, G. Warren, M. Mathieu, S. W. Muchmore, S. P. Brown, J. A. Grant, J. A. Haigh, N. Nevins, A. N. Jain and B. Kelley, *Journal of Medicinal Chemistry*, 2010, **53**, 3862–3886.
- [341] G. A. Holdgate, T. D. Meek and R. L. Grimley, *Nature Reviews Drug Discovery*, 2018, **17**, 115–132.
- [342] J. Chen, Q. Wang, B. Malone, E. Llewellyn, Y. Pechersky, K. Maruthi, E. T. Eng, J. K. Perry, E. A. Campbell, D. E. Shaw and S. A. Darst, *Nature Structural & Molecular Biology*, 2022, **29**, 250–260.
- [343] F. Clark, G. Robb, D. Cole and J. Michel, *Automated Adaptive Absolute Binding Free Energy Calculations*, 2024.

- [344] B. Cree, M. Bieniek, S. Amin, A. Kawamura and D. Cole, *Digital Discovery*, 2024.
- [345] A. Douangamath, A. Powell, D. Fearon, P. M. Collins, R. Talon, R. Krojer, T. and Skyner, J. Brandao-Neto, L. Dunnett, A. Dias, A. Aimon, N. M. Pearce, C. Wild, T. Gorrie-Stone and F. von Delft, *J. Vis. Exp.*, 2021, **171**, e62414.
- [346] D. J. Wood, J. D. Lopez-Fernandez, L. E. Knight, I. Al-Khawaldeh, C. Gai, S. Lin, M. P. Martin, D. C. Miller, C. Cano, J. A. Endicott, I. R. Hardcastle, M. E. M. Noble and M. J. Waring, *Journal of Medicinal Chemistry*, 2019, **62**, 3741–3752.
- [347] A. Douangamath, D. Fearon, P. Gehrtz, T. Krojer, P. Lukacik, C. D. Owen, E. Resnick, C. Strain-Damerell, A. Aimon, P. Abranyi-Balogh, J. Brandao-Neto, A. Carbery, G. Davison, A. Dias, T. D. Downes, L. Dunnett, M. Fairhead, J. D. Firth, S. P. Jones, A. Keeley, G. M. Keseru, H. F. Klein, M. P. Martin, M. E. M. Noble, P. O'Brien, A. Powell, R. N. Reddi, R. Skyner, M. Snee, M. J. Waring, C. Wild, N. London, F. von Delft and M. A. Walsh, *Nature Communications*, 2020, **11**, 5047.
- [348] S. R. Mackinnon, T. Krojer, W. R. Foster, L. Diaz-Saez, M. Tang, K. V. M. Huber, F. von Delft, K. Lai, P. E. Brennan, G. Arruda Bezerra and W. W. Yue, *ACS Chemical Biology*, 2021, **16**, 586–595.
- [349] O. O. Grygorenko, D. S. Radchenko, I. Dziuba, A. Chuprina, K. E. Gubina and Y. S. Moroz, *iScience*, 2020, **23**, 101681.
- [350] W. A. Warr, M. C. Nicklaus, C. A. Nicolaou and M. Rarey, *Journal of Chemical Information and Modeling*, 2022, **62**, 2021–2034.
- [351] J. Kuan, M. Radaeva, A. Avenido, A. Cherkasov and F. Gentile, *WIREs Computational Molecular Science*, 2023, **13**, e1678.
- [352] F. Imrie, T. E. Hadfield, A. R. Bradley and C. M. Deane, *Chem. Sci.*, 2021, **12**, 14577–14589.
- [353] N. T. Runcie and A. S. Mey, *Journal of Chemical Information and Modeling*, 2023, **63**, 5996–6005.
- [354] A. A. Sadybekov, A. V. Sadybekov, Y. Liu, C. Iliopoulos-Tsoutsouvas, X.-P. Huang, J. Pickett, B. Houser, N. Patel, N. K. Tran, F. Tong, N. Zvonok, M. K. Jain, O. Savych, D. S. Radchenko, S. P. Nikas, N. A. Petasis, Y. S. Moroz, B. L. Roth, A. Makriyannis and V. Katritch, *Nature*, 2022, **601**, 452–459.

- [355] S. Gahbauer, G. J. Correy, M. Schuller, M. P. Ferla, Y. U. Doruk, M. Rachman, T. Wu, M. Diolaiti, S. Wang, R. J. Neitz, D. Fearon, D. S. Radchenko, Y. S. Moroz, J. J. Irwin, A. R. Renslo, J. C. Taylor, J. E. Gestwicki, F. von Delft, A. Ashworth, I. Ahel, B. K. Shoichet and J. S. Fraser, *Proceedings of the National Academy of Sciences*, 2023, **120**, e2212931120.
- [356] M. Ferla, R. Sánchez-García, R. Skyner, S. Gahbauer, J. Taylor, F. von Delft, B. Marsden and C. Deane, *ChemRxiv*, 2024.
- [357] P. Ertl, E. Altmann and S. Racine, *Bioorganic & Medicinal Chemistry*, 2023, **81**, 117194.
- [358] V. Fialková, J. Zhao, K. Papadopoulos, O. Engkvist, E. J. Bjerrum, T. Kogej and A. Patronov, *Journal of Chemical Information and Modeling*, 2022, **62**, 2046–2063.
- [359] J. J. Irwin, K. G. Tang, J. Young, C. Dandarchuluun, B. R. Wong, M. Khurelbaatar, Y. S. Moroz, J. Mayfield and R. A. Sayle, *Journal of Chemical Information and Modeling*, 2020, **60**, 6065–6073.
- [360] D. E. Graff, E. I. Shakhnovich and C. W. Coley, *Chem. Sci.*, 2021, **12**, 7866–7881.
- [361] Y. Khalak, G. Tresadern, D. F. Hahn, B. L. de Groot and V. Gapsys, *Journal of Chemical Theory and Computation*, 2022, **18**, 6259–6270.
- [362] J. Thompson, W. P. Walters, J. A. Feng, N. A. Pabon, H. Xu, M. Maser, B. B. Goldman, D. Moustakas, M. Schmidt and F. York, *Artificial Intelligence in the Life Sciences*, 2022, **2**, 100050.
- [363] R. Gorantla, A. Kubincová, B. Suutari, B. P. Cossins and A. S. J. S. Mey, *Journal of Chemical Information and Modeling*, 2024, **64**, 1955–1965.
- [364] D. van Tilborg and F. Grisoni, *ChemRxiv*, 2024.
- [365] F. Gentile, V. Agrawal, M. Hsing, A.-T. Ton, F. Ban, U. Norinder, M. E. Gleave and A. Cherkasov, *ACS Central Science*, 2020, **6**, 939–949.
- [366] Y. Yang, K. Yao, M. P. Repasky, K. Leswing, R. Abel, B. K. Shoichet and S. V. Jerome, *Journal of Chemical Theory and Computation*, 2021, **17**, 7106–7119.
- [367] T. Sivula, L. Yetukuri, T. Kalliokoski, H. Käsnänen, A. Poso and I. Pöhner, *Journal of Chemical Information and Modeling*, 2023, **63**, 5773–5783.
- [368] K. D. Konze, P. H. Bos, M. K. Dahlgren, K. Leswing, I. Tubert-Brohman, A. Bortolato, B. Robbason, R. Abel and S. Bhat, *Journal of Chemical Information and Modeling*, 2019, **59**, 3782–3793.

- [369] F. Gusev, E. Gutkin, M. G. Kurnikova and O. Isayev, *Journal of Chemical Information and Modeling*, 2023, **63**, 583–594.
- [370] M. F. Adasme, K. L. Linnemann, S. N. Bolz, F. Kaiser, S. Salentin, V. J. Haupt and M. Schroeder, *Nucleic Acids Research*, 2021, **49**, W530–W534.
- [371] E. Glaab, G. B. Manoharan and D. Abankwa, *Journal of Chemical Information and Modeling*, 2021, **61**, 4082–4096.
- [372] J. Hazemann, T. Kimmerlin, R. Lange, A. M. Sweeney, G. Bourquin, D. Ritz and P. Czodrowski, *bioRxiv*, 2024.
- [373] V. Chenthamarakshan, S. C. Hoffman, C. D. Owen, P. Lukacik, C. Strain-Damerell, D. Fearon, T. R. Malla, A. Tumber, C. J. Schofield, H. M. Duyvesteyn, W. Dejnirattisai, L. Carrique, T. S. Walter, G. R. Screaton, T. Matviuk, A. Mojsilovic, J. Crain, M. A. Walsh, D. I. Stuart and P. Das, *Science Advances*, 2023, **9**, eadg7865.
- [374] M. Alnammi, S. Liu, S. S. Ericksen, G. E. Ananiev, A. F. Voter, S. Guo, J. L. Keck, F. M. Hoffmann, S. A. Wildman and A. Gitter, *Journal of Chemical Information and Modeling*, 2023, **63**, 5513–5528.
- [375] S. Boothroyd, P. K. Behara, O. C. Madin, D. F. Hahn, H. Jang, V. Gapsys, J. R. Wagner, J. T. Horton, D. L. Dotson, M. W. Thompson, J. Maat, T. Gokey, L.-P. Wang, D. J. Cole, M. K. Gilson, J. D. Chodera, C. I. Bayly, M. R. Shirts and D. L. Mobley, *Journal of Chemical Theory and Computation*, 2023, **19**, 3251–3275.
- [376] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *Journal of Machine Learning Research*, 2011, **12**, 2825–2830.
- [377] T. Danka and P. Horvath.
- [378] Dask Development Team, *Dask: Library for dynamic task scheduling*, 2016.
- [379] G. D. Noske¹, E. de Souza Silva¹, M. O. de Godoy¹, I. Dolci¹, R. S. Fernandes¹, R. V. C. Guido¹, P. Sjö, G. Oliva¹ and A. S. Godoy, *Journal of Biological Chemistry*, 2023, **299**, 103004.
- [380] D. Anstine, R. Zubatyuk and O. Isayev, *ChemRxiv*, 2024.
- [381] J. C. Fromer, D. E. Graff and C. W. Coley, *Digital Discovery*, 2024, **3**, 467–481.

- [382] A. S. Powers, H. H. Yu, P. Suriana, R. V. Koodli, T. Lu, J. M. Paggi and R. O. Dror, *ACS Central Science*, 2023, **9**, 2257–2267.
- [383] J. J. Irwin, K. G. Tang, J. Young, C. Dandarchuluun, B. R. Wong, M. Khurelbaatar, Y. S. Moroz, J. Mayfield and R. A. Sayle, *J. Chem. Inf. Model.*, 2020, **60**, 6065–6073.
- [384] G. Zhang, J. Zhang, Y. Gao, Y. Li and Y. Li, *Expert Opinion on Drug Discovery*, 2022, **17**, 55–69.
- [385] B. Lau, P. S. Emani, J. Chapman, L. Yao, T. Lam, P. Merrill, J. Warrell, M. B. Gerstein and H. Y. K. Lam, *Bioinformatics*, 2023, **39**, btac789.
- [386] J. B. Rosenzweig, N. Majernik, R. R. Robles, G. Andonian, O. Camacho, A. Fukasawa, A. Kogar, G. Lawler, J. Miao, P. Musumeci, B. Naranjo, Y. Sakai, R. Candler, B. Pound, C. Pellegrini, C. Emma, A. Halavanau, J. Hastings, Z. Li, M. Nasr, S. Tantawi, P. Anisimov, B. Carlsten, F. Krawczyk, E. Simakov, L. Faillace, M. Ferrario, B. Spataro, S. Karkare, J. Maxson, Y. Ma, J. Wurtele, A. Murokh, A. Zholents, A. Cianchi, D. Cocco and S. B. van der Geer, *New Journal of Physics*, 2020, **22**, 093067.
- [387] A. T. McNutt, P. Francoeur, R. Aggarwal, T. Masuda, R. Meli, M. Ragoza, J. Sunseri and D. R. Koes, *J. Cheminform*, 2021, **13**, 43.
- [388] M. Buttenschoen, G. M. Morris and C. M. Deane, *Chemical Science*, 2024, **15**, 3130–3139.
- [389] S. Lehtola, *The Journal of Chemical Physics*, 2023, **159**, 180901.
- [390] C. S. Adorf, V. Ramasubramani, J. A. Anderson and S. C. Glotzer, *Computing in Science & Engineering*, 2019, **21**, 66–79.
- [391] A. S. J. S. Mey, J. J. Jiménez and J. Michel, *Journal of Computer-Aided Molecular Design*, 2018, **32**, 199–210.
- [392] *OMSF*, <https://omsf.io/>.
- [393] L. Belzner, T. Gabor and M. Wirsing, *Bridging the Gap Between AI and Reality*, Cham, 2024, pp. 355–374.
- [394] O.-H. Choung, R. Vianello, M. Segler, N. Stiefl and J. Jiménez-Luna, *Nature Communications*, 2023, **14**, 6651.
- [395] T. Blaschke, J. Arús-Pous, H. Chen, C. Margreitter, C. Tyrchan, O. Engkvist, K. Papadopoulos and A. Patronov, *Journal of Chemical Information and Modeling*, 2020, **60**, 5918–5922.

- [396] W. Jin, S. Sarkizova, X. Chen, N. HaCohen and C. Uhler, *Advances in Neural Information Processing Systems*, 2023, **36**, 33514–33528.
- [397] X. Du, Y. Li, Y.-L. Xia, S.-M. Ai, J. Liang, P. Sang, X.-L. Ji and S.-Q. Liu, *International Journal of Molecular Sciences*, 2016, **17**, 144.
- [398] S. C. Musson and M. T. Degiacomi, *Journal of Open Source Software*, 2023, **8**, 5523.
- [399] H. Liu and W. Lee, *International Journal of Molecular Sciences*, 2019, **20**, 3421.
- [400] CHEESE, <https://cheese.deepmedchem.com/>.
- [401] L. Zhang, D. Lin, X. Sun, U. Curth, C. Drosten, L. Sauerhering, S. Becker, K. Rox and R. Hilgenfeld, *Science*, 2020, **368**, 409–412.
- [402] S. J. Capuzzi, E. N. Muratov and A. Tropsha, *Journal of Chemical Information and Modeling*, 2017, **57**, 417–427.
- [403] Z. Jin, X. Du, Y. Xu, Y. Deng, M. Liu, Y. Zhao, B. Zhang, X. Li, L. Zhang, C. Peng, Y. Duan, J. Yu, L. Wang, K. Yang, F. Liu, R. Jiang, X. Yang, T. You, X. Liu, X. Yang, F. Bai, H. Liu, X. Liu, L. W. Guddat, W. Xu, G. Xiao, C. Qin, Z. Shi, H. Jiang, Z. Rao and H. Yang, *Nature*, 2020, **582**, 289–293.