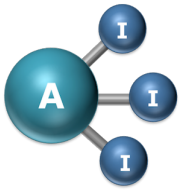



# Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery


## Royal Society of Chemistry: Chemical Information and Computer Applications Group

AI$^3$ Science Discovery Network+ & Royal Society of Chemistry: Chemical Information and Computer Applications Group: AI4Proteins Seminar Series 2021
14/04/2021, 05/05/2021, 26/05/2021 & 16-17/06/2021
AI$^3$ Science Discovery Network+
Online

Dr. Wendy A. Warr
Wendy Warr & Associates

11/11/2020

AI3SD-Event-Series:Report-23

AI$^3$ Science Discovery Network+ & Royal Society of Chemistry: Chemical Information and Computer Applications Group: AI4Proteins Seminar Series 2021
AI3SD-Event-Series:Report-23
11/11/2020
DOI: 10.5258/SOTON/AI3SD0176
Published by University of Southampton

**Network: Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery**

This Network+ is EPSRC Funded under Grant No: EP/S000356/1
Principal Investigator: *Professor Jeremy Frey*
Co-Investigator: *Professor Mahesan Niranjan*
Network+ Coordinator: *Dr Samantha Kanza*

**Royal Society of Chemistry: Chemical Information and Computer Applications Group**

The Chemical Information and Computer Applications Group (CICAG) is one of the RSC's many member-led Interest Groups, which exist to benefit RSC members and the wider chemical science community, and to meet the requirements of the RSC's strategy and charter.
Chair: *Dr Chris Swain*
Secretary: *Professor Jeremy Frey*
Treasurer: *Dr Diana Leitch*

# Contents

A report by Wendy A. Warr (wendy@warr.com) on a series of virtual meetings organized by the AI3SD Network (Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery) and the Royal Society of Chemistry Chemical Information and Computer Applications Group (RSC-CICAG), April 14, May 5, May 26, and June 16-17, 2021

# 1 Event details

| | |
|---|---|
| Title | AI4Proteins: Protein Structure Prediction |
| Organisers | AI$^3$ Science Discovery Network+ (AI$^3$SD),<br>Royal Society of Chemistry: Chemical Information and Computer Applications Group (RSC-CICAG) |
| Dates | 14/04/2021, 05/05/2021, 26/05/2021 & 16-17/06/2021 |
| Programme | Programme |
| No. Participants | 99 |
| Location | Online |
| Organisation Committee | Dr Samantha Kanza and Professor Jeremy Frey (AI$^3$SD),<br>Dr Chris Swain and Dr Nathan Brown, (RSC-CICAG) |
| Conference Chairs | Dr Samantha Kanza (University of Southampton),<br>Professor Jeremy Frey (University of Southampton),<br>Dr Melanie Vollmar (Diamond),<br>Dr Chris Swain (MedChemica Limited),<br>Dr Márton Vass (Benevolent AI Limited),<br>Dr Simone Fulle (Novo Nordisk A/S),<br>Professor Jonathan Goodman (University of Cambridge),<br>Dr Lucy Colwell (University of Cambridge) and<br>Dr Nathan Brown (Benevolent AI) |

# 2 Machine learning for biological sequence design

**Lucy Colwell, Centre for Molecular Informatics, Yusuf Hamied Department of Chemistry, University of Cambridge, United Kingdom**

The full video of Colwell's talk can be viewed here: https://youtu.be/_LNsLNnEx7E.

Proteins are made from sequences of 20 amino acids that fold into 3D structures and carry out complex functions. There is a wealth of sequence data and programs such as AlphaFold can predict 3D protein structure from sequences. Nature does countless experiments, by mutation and selection, to find good proteins. Prediction of protein functional properties from sequence is a challenge that would allow the discovery of new proteins with specific functionality.

Directed evolution[1] mimics the process of natural selection to steer proteins toward a user-defined goal. It consists of subjecting a gene to iterative rounds of mutagenesis (creating a library of variants), selection (expressing those variants and isolating members with the desired function) and amplification (generating a template for the next round).

Colwell's team has made progress on the challenge of adding machine learning (ML) to directed evolution. They learn sequence representations or embeddings from raw unaligned protein sequences, where proximity encodes functional activity; use the initial labeled data to calibrate a model of the relationship between sequence and phenotype; and optimize the model to select the most informative next batch of sequences.

PFam[2] is a database of proteins with the functional parts (domains) annotated. Full families are assembled using hidden Markov models (HMMs) fit on seed families. Seeds are mostly human-verified, with members chosen as "exemplars" of families. One-third of all protein-coding genes from bacterial genomes cannot be annotated with a function.[3] Colwell and her co-workers seek to address this issue. They have reported convolutional neural networks (CNNs) that are significantly more accurate and computationally efficient than BLAST[4] or HMMER,[5] and do not require sequence alignment, while learning sequence features such as structural disorder and transmembrane (TM) helices.[6] The model colocates sequences from unseen families in embedding space, allowing sequences from novel families to be annotated accurately. Such models can complement existing approaches and provide protein function prediction tools with broad coverage of the protein universe, enabling more distant sequences to be annotated. The European Molecular Biology Laboratory European Bioinformatics Institute (EMBL-EBI) has recently released a new file called Pfam-N which provides additional Pfam 34.0 matches identified by this technology.

Once a machine-learning model has been built, it has to be optimized to select the most informative next batch of sequences. Directed evolution is a form of randomized local search which is sample-inefficient and relies on greedy hill-climbing to the optimal sequences. Recent work has demonstrated that optimization guided by machine learning can find better sequences faster.

Reinforcement learning (RL) provides a flexible framework for black-box optimization that can harness deep generative sequence models. Colwell and her colleagues have proposed a method[7] for improving the sample efficiency of policy gradient methods such as proximal policy optimization (PPO)[8] by using surrogate models that are trained online to approximate an experimentally measured functional property.

They have provided a model-based RL algorithm, DyNA PPO, and demonstrated its effectiveness in performing sample-efficient, batched black-box function optimization. They address model bias by quantifying reliability and automatically selecting models of appropriate complexity *via* cross-validation. They propose a visitation-based exploration bonus and show that it is more effective than entropy-regularization in identifying multiple local optima, and they have presented a new optimization task for benchmarking methods for biological sequence design based on protein energy Ising models (statistical thermodynamic "nearest-neighbor" models).

The cost and latency of wet-lab experiments requires methods that find good sequences in few experimental rounds that each contain large batches of sequences, a setting that off-the-shelf black-box optimization methods are ill-equipped to handle. The performance of existing methods varies drastically across optimization tasks. To improve robustness, Colwell and her co-workers have proposed population-based black-box optimization (P3BO), which generates batches of sequences by sampling from an ensemble of methods. The number of sequences sampled from any method is proportional to the quality of sequences it previously proposed, allowing P3BO to combine the strengths of individual methods while hedging against their innate brittleness. Adapting the hyperparameters of each of the methods online using

evolutionary optimization further improves performance. Through extensive experiments on *in silico* optimization tasks, the team has shown that P3BO outperforms any single method in its population, proposing higher quality sequences as well as more diverse batches.

The potential of this approach is illustrated through the design and experimental validation of viable adeno-associated virus (AAV) capsid protein variants for gene therapy applications. Gene therapy using AAV as a vector has emerged as a novel therapeutic modality that has the potential to lead to substantial disease modification in many monogenic disorders. To realize this potential, diverse new capsid proteins are required that prevent attack by the host immune system and deliver the DNA cargo to specific tissues or cell types.

Together with collaborators at Harvard, Colwell's team have applied deep learning to design highly diverse AAV2 capsid protein variants that remain viable for packaging of a DNA payload.[9] Sixty copies of a single monomer of about 700 amino acids from the CAP gene from AAV assemble into a very large and highly symmetric capsid. Simulation is a seemingly impossible task but machine learning provides a new approach to AAV capsid design. Colwell's team made changes in "tile 21" (about 120 base pairs) of the CAP gene. The team tested single and multimutant variants in this region. DNA was synthesized and built into a plasmid library, and after transformation, cloning and extraction, $P$ copies of each variant were made. Every member of the library was turned into a virus and put through a round of viral replication, and after viral DNA extraction, $V$ copies of each variant were produced. The selection coefficient is $V/P$.

In an additive baseline model, multimutants were made by combining mutations with good single site outcomes. The team evolved sequences guided by this model and tested about 56,000 baseline sequence designs. They measured single site mutations, used the additive model to choose multimutants, and assayed them experimentally. A plot of precision *versus* number of mutations showed that as you take further steps in the space the ability to find viable mutants drops off. The team therefore built three new models of the relationship between sequence and viability using logistic regression, a convolutional neural network, and a recurrent neural network. Training neural networks (NNs) on a limited dataset of single plus random double mutants (near the wild type) actually did a better job of finding new sequences than logistic regression.

In prospective validation, the team designed new sequences using all nine possible combinations of three models and three different training sets. They ranked sequences based on the model and then evolved the sequences using each model. They tested 100 model-selected and 900 model-designed sequences at each distance 5-29 steps from wild type. The additive model performed fairly well, but the neural networks performed a lot better: mutations up to 15 away from the wild type were reliably found, and with a larger training set, sequences even further away could be found. The logistic regression model can do very well but is not as robust. Sequences found by the logistic regression model were not as diverse as those from the neural networks, across all three training sets.

The Harvard researchers in the team have founded Dyno Therapeutics to take these approaches into the clinic through partnerships with Roche, Sanofi, Sarepta and Novartis.

# 3 Machine learning applications for macromolecular X-ray crystallography at Diamond

**Melanie Vollmar, Postdoctoral Researcher at Diamond Light Source, Oxfordshire, United Kingdom**

The full video of Vollmar's talk can be viewed here: https://youtu.be/3cRclaxOHWE.

Diamond Light Source is a not-for-profit limited company funded as a joint venture between UK Research & Innovation and the Wellcome Trust. It provides national science infrastructure that is free at the point of use. Primary facilities are the national synchrotron along with cryo-electron microscopy (cryo-EM) at the Harwell Science and Innovation Campus. Over 14,000 researchers from across life and physical sciences both from academia and industry use Diamond to conduct experiments, assisted by approximately 700 staff.

For more than 50 years, X-ray diffraction has been used to investigate protein crystals and the resulting diffraction images have been analyzed to reveal the structure of proteins to atomic level. Detailed information about the structure can be obtained if the "phase problem" can be solved for the molecule under study. The phase problem arises because it is only possible to measure the amplitude of diffraction spots: information on the phase of the diffracted radiation is missing. "Phasing" can be done experimentally, or by molecular replacement using phases from related proteins. Unfortunately, the vast majority of recorded diffraction data do not yield a protein structure. Vollmar and her co-workers have used machine learning to address this problem by producing a proof of concept[10] for a tool that could help crystallographers assess their datasets in order to determine which should be put forward for full analysis and structure solution using experimental phasing.

Other motivations for the use of machine leaning in decision making include:

- Advances in detector technology toward very high frame rates and read-out speeds
- Large amounts of raw diffraction data in a very short period of time
- Automated sample exchange through robotics
- Automated sample centering and screening
- Automated and unattended data collection
- The requirement for downstream analysis (data integration, scaling, phasing, model building and refinement)
- The excess of data for manual assessment, even for an expert crystallographer
- Limitations on computational resources and data storage
- Provision of guidance to nonexpert users
- Enabling high-throughput data analysis.

The research team used raw diffraction data for 810 structures in the Protein Data Bank (PDB): 303 from the Structural Genomics Consortium and 507 from the Joint Center for Structural Genomics. The metadata for these structures were retrieved from the published PDB files and stored in a database, METRIX_DB, created using the SQLite3 programming language accessed through a standard library within Python. Additional information was created concerning data integration and reduction, experimental phasing, and sequence analysis, and was likewise stored in METRIX_DB.

In a pre-assessment step, the most important features in decision making were identified using statistical tools such as Pearson's linear correlation coefficient and recursive feature elimination. Classifiers tried were a support vector machine with a linear kernel or with a radial base function kernel, a decision tree with or without bagging, a decision tree with AdaBoost, a random forest, and an extreme random forest.

Based on the results of automated analysis, 440 structures successfully produced 703 samples, each of which represents a crystallographic dataset at a single wavelength. These were split 20:80 between test and training sets. The six features with highest frequency of occurrence proved to be optimal. The reduced feature set identified was used to retrain all classifiers, and the best-performing classifier was a decision tree with AdaBoost. Vollmar showed its confusion matrix and a radar plot for classification accuracy and error, F1 score, and area under the receiver operating characteristic curve (ROC AUC), compared with those for a perfect classifier. Twenty-four new samples were used to challenge the classifier. Classification accuracy achieved was 79%. Sensitivity and specificity were 64% and 92%, respectively. The false-positive rate was 8% with a precision of 86% and an F1 score of 74%, and ROC AUC was 75%.

The method has been tested in Diamond's data analysis pipelines. A series of scripts has been developed to run xia2 (an expert system for macromolecular crystallography data reduction),[11] using Diffraction Integration for Advanced Light Sources(DIALS)[12] and AIMLESS[13] for diffraction image integration and data reduction. The xia2-3dii system also requires X-ray Detector Software (XDS). The predictions have been running since early in 2020. Vollmar presented results for the nine months from January 2020. At first, the results were disappointing. It seemed that this could have been because the public data used for training were out of date, so the system was retrained on user data from run1 2020 on the Diamond beamline and new predictions were run. The results (Table 1) are now better but still leave room for improvement.

Table 1. Results of Runs 2-4 in 2020

| | XIA2-3ddi (XDS) | XIA2-DIALS |
|---|---|---|
| Class accuracy (%) | 55 | 61 |
| Class error (%) | 45 | 39 |
| Sensitivity (%) | 70 | 9 |
| Specificity (%) | 43 | 94 |
| False positive rate (%) | 57 | 6 |
| Precision (%) | 46 | 46 |
| F1 score | 56 | 15 |
| True positives | 284 | 19 |
| True negatives | 254 | 325 |
| False positives | 330 | 22 |
| False negatives | 119 | 198 |
| Positives | 403 | 217 |
| Negatives | 584 | 347 |
| Total | 987 | 564 |

A one-size-fits-all predictor is not suitable: one predictor for each integration package is needed. Predicting experimental phasing success alone is too coarse: it is better to predict the chances of success for combinations of integration and experimental phasing software. Using current user data for training rather than public data avoids a technology gap. The DIALS predictor works very well to identify data where experimental phasing will not work (it identifies 94% of samples correctly) but the team needs to work on identifying success cases. The 3dii (XDS) predictor works well to identify data where experimental phasing is likely to work (it identifies 70% of samples correctly) but the team needs to work on identifying failure cases. Not all downstream experimental phasing packages (here Crank2, AutoBuild/AutoSol, and autoSHARP) produce a result with every dataset and there are few identical cases for either integration package.

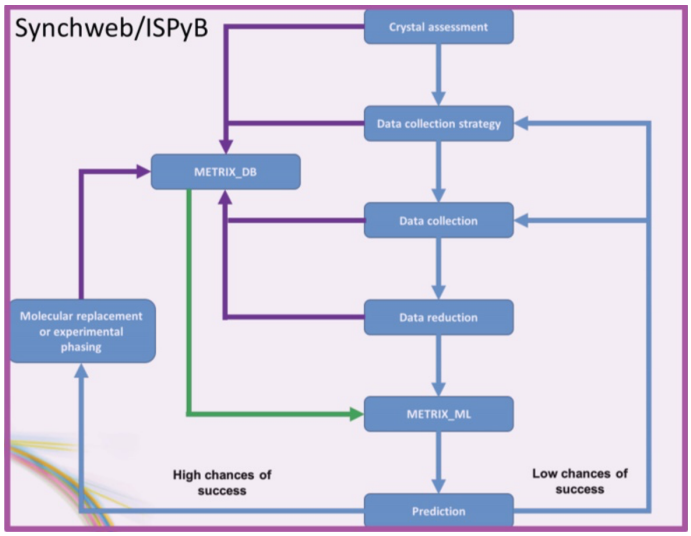



Figure 1. Future application.

The flowchart in Figure 1 gives a schematic outline for an interactive user system. Although a user will always be able to ignore the recommendations and trigger data analysis manually, including the trained algorithm in the analysis pipelines is expected to help in balancing the workload on the computing infrastructure in a more intelligent way than the brute-force approach currently in use.

## 4 Machine learning for molecular spectroscopy study

**Jun Jiang, Professor in Physical Chemistry, University of Science and Technology of China, Hefei, Anhui, P. R. China**

The full video of Jiang's talk can be viewed here: https://doi.org/10.5258/SOTON/P0094.

Quantum mechanics (QM) was built on the interpretation of atomic spectra. Optical spectroscopy provides powerful toolkits to decipher molecular structures and their configuration evolutions. Circular dichroism spectra for protein folding,[14] Raman spectra for molecular vibration,[15] and surface-enhanced Raman spectroscopy (SERS) fluorescence[16] for charge transfer are all based on the known rules of quantum mechanics but the intrinsic complexity of spectroscopic signals of molecular systems makes it difficult to correlate spectral characteristics with the underlying molecular structure and dynamics. The theoretical analysis of spectroscopic signals and connecting them with structural detail has long been a challenging task.

Data-driven artificial intelligence (AI) provides a new tool to solve highly complex scientific problems based on simple, known rules.[17–19] Firstly, AI can be used for spectroscopy prediction, to accelerate QM computations for spectroscopy simulations and analysis. Spectroscopy is widely used to monitor protein structural evolution and transformation but theoretical prediction of spectroscopic signals is very difficult, limiting real-time structure detection.

The mapping of UV spectra to atomic structure of proteins relies on expensive theoretical simulations, circumventing the heavy computational cost which involves repeated QM simulations of excited-state properties of many fluctuating protein geometries. Jiang and his colleagues have shown[20] that a neural network machine-learning technique can predict electronic absorption spectra of N-methylacetamide, which is a widely used model system for the peptide bond (that constituting the backbone of a protein). Using ground-state geometric parameters and charge information as descriptors, they employed a neural network (TensorFlow 1.14.0 with three hidden layers) to predict photo-excitation energies, ground-state, and transition dipole moments of many molecular dynamics (MD) conformations at different temperatures, in agreement with time-dependent density functional theory (DFT) calculations.

After machine learning a one-to-one relationship between structure and Hamiltonians, the model can predict the Hamiltonian at first-principles level for an input protein and from that the spectrum can be predicted. The researchers were able to reproduce experimental UV spectra by a neural network, and, more importantly, explain structural evolutions based on many MD configurations. The neural network simulations are nearly 3000 times faster than comparable quantum calculations.

The team has used a similar method to predict protein infrared (IR) spectra.[21] The amide I region in the spectrum is dominated by the stretching vibration of the carbonyl group in the peptide bond, and provides a fingerprint of protein structure and dynamics. The machine learning protocol of Ye *et al.*[21] uses a few key structural descriptors to predict amide I IR spectra of various proteins rapidly and agrees well with experiment. Its transferability enabled the team to distinguish protein secondary structures, probe atomic structure variations with temperature, and monitor protein folding. For a protein with about 1000 amino acids a DFT calculation that might take 3.8 years can be replaced by a machine learning technique that takes 3.3 hours. The method is available online to predict the IR or UV spectrum of a protein input as a PDB file.

SERS can capture the electronic-vibrational fingerprint of molecules surfaces but *ab initio* prediction of Raman response is a long-standing challenge because of the diversified interfacial structures. Jiang and his colleagues have reported a machine learning random forest method[22] that can predict, 10,000 times faster than DFT, the SERS signals of a trans-1,2-bis(4-pyridyl)ethylene molecule adsorbed on a gold substrate. Using geometric descriptors extracted from quantum chemistry simulations of thousands of *ab initio* MD conformations, the machine learning protocol predicts vibrational frequencies and Raman intensities. The resulting spectra agree with DFT calculations and experiment. Predicted SERS responses of the molecule on different surfaces, or under external electric fields and solvent environment, demonstrated the good transferability of the protocol.

AI can also be used for spectroscopy interpretation, revealing unknown correlations between the properties of a molecule and its spectra (Figure 2).

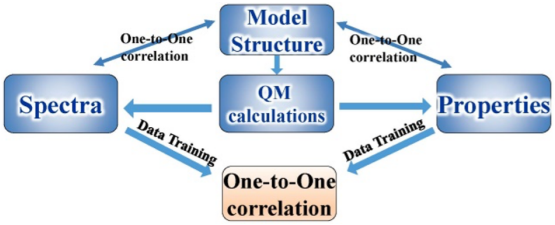



Figure 2. Correlation between chemical properties and spectra.

Jiang and his colleagues have demonstrated that surface chemistry can be understood and predicted using a simple adsorbate-surface interaction descriptor that relates charge polarization to chemical reactivity.[23] The product of two electric dipole moments of catalyst surface and molecule provided a parameter that allowed the team to predict the key catalytic properties for different adsorption sites and reaction pathways. Further findings have validated the effectiveness of the electric dipole descriptor. By training a machine learning neural network with a large dataset of first-principles calculations, the researchers achieved quick and accurate predictions of molecular adsorption energy and transferred charge, validating the effectiveness of the electric dipole descriptor.[24] The training model can be extended to study additional substrates.

Once a model has been developed for a one-to-one correlation between spectra and properties (Figure 2) it should be possible to use it to deduce chemical properties by measuring spectra. Jiang's team is preparing to publish a mathematical expression for predicting chemical properties from molecular IR or Raman spectroscopic features. Using such a protocol, AI could be used to predict new chemical structures from spectra.

Jiang's team has also experimented with chemical group recognition of the intramolecular hydrogen bond and the hydroxyl (O-H) bond, based on a machine learning study on the stretching vibration in IR and Raman spectra. They studied hydroxyl-carbonyl recognition with K-means clustering, validated with the silhouette coefficient. Transferring from training sets of molecules with eight heavy atoms (QM8) to test sets of 9-16 heavy atoms (QM9), they obtained accuracies of over 98% for combined IR and Raman spectra. It follows that unsupervised machine learning could be used to distinguish protein secondary structure. A convolutional neural network (CNN) model was trained and tested on a dataset consisting 87,993 spectra computed from protein peptide segments with $\alpha$-helical, $\beta$-sheet, and other typical secondary structures. The secondary structure classification accuracy reached nearly 100% and over 98.7% on spectra sets of new segments extracted from the same and homologous proteins, respectively.

## 5 Molecular dynamics simulations of proteins

**Jonathan Essex, Professor of Computational Systems Chemistry, University of Southampton, United Kingdom**

MD simulations are now widely used to explore protein structure and function. These simulations are underpinned by physical models of the interactions between the system components, and Newtonian dynamics to capture flexibility and entropic effects. In classical molecular mechanics (MM), the interactions between the atoms in the system are represented by a force field described by a potential with terms for bond stretching and vibration, angle bending, torsional motion, and van der Waals and electrostatic nonbonded interactions (eq 1).

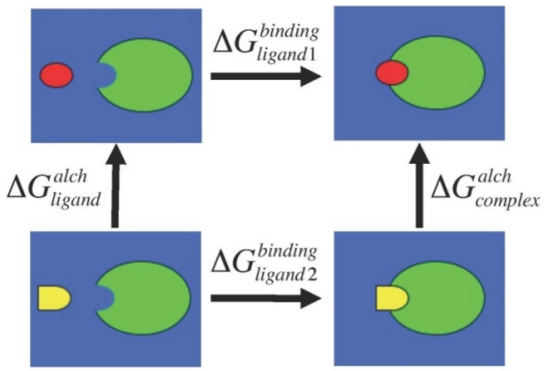
$$E_{potential} = E_{bond} + E_{angle} + E_{torsion} + E_{vdW} + E_{coulomb} \quad (1)$$

Newtonian dynamics is used to study how the system evolves as a function of time, that is, to propagate the energies into forces and hence into dynamic motion.

Essex's group is particularly interested in modeling membrane systems. The complexity of the calculation can be reduced by subsuming groups of atoms into interaction beads. This coarse graining model can be used to explore lipid phases for membranes, proteins, or sugars. The team has also taken dual-resolution modeling approaches, mixing the levels of representation of, for example, a small molecule antimicrobial and a membrane system which it disrupts. Additionally, Essex's group is studying big biological problems, for example modeling a lipid antigen bound to a protein in the immune system, to study T-cell or non T-cell response. The team is also looking at protein-protein complexes such as antigen-antibody complexes, with subtle and unusual effects in differential response.

In the current talk, Essex concentrated on protein-ligand binding. Predicting the experimental binding free energy allows a ligand to be optimized for affinity and selectivity. There is a strong theoretical foundation for identifying possible drug binding geometries and estimating drug binding affinities, but some structural information and an accurate model for the interactions are required. Use of MD in the calculations is important.

Ligand binding is driven by changes in the Gibbs free energy. Relative binding free energy with alchemical mutation makes use of the thermodynamic cycle to calculate the binding free energy difference between end-points in the cycle (Figure 3). Free energy calculation methodologies are used to estimate the free energies of proteins, their ligands and their complexes using conformations generated *via* MD simulations. $\Delta\Delta G^{binding}$ is either the difference between two processes usually measured experimentally, or two processes that can be simulated with MD and free energy calculations.




$$\Delta\Delta G^{binding} = \Delta G^{binding}_{ligand2} - \Delta G^{binding}_{ligand1} = \Delta G^{alch}_{ligand} - \Delta G^{alch}_{complex}$$

Figure 3. Alchemical free energy calculation.

These calculations are expensive and need extensive sampling. They also suffer from accuracy, precision, and reproducibility issues. Essex and his co-workers have performed a high-throughput study of 1296 such calculations, using over 220 $\mu$s of total sampling time, on three different protein systems (dihydrofolate reductase (DHFR), protein tyrosine phosphatase 1B (PTP1B), and factor Xa (FXa)), to investigate the impact of the initial crystal structure on the resulting binding free energy values.[25] They also considered the influence of equilibration time and discovered that the initial crystal structure can have a significant effect

on free energy values obtained at short timescales that can manifest itself as a free energy difference of more than 1 kcal/mol. At longer timescales, these differences are largely overtaken by important rare events, such as torsional ligand motions, typically resulting in a much higher uncertainty in the obtained values.

The team has also studied the impact of the initial tautomer or rotamer state of histidine (His) in trypsin and heat shock protein 90 (HSP90) using 777 free energy calculations. They found that free energies are highly sensitive to binding site tautomers and rotamers: differences of up to 1.5 kcal/mol were found for His57. His40 is 1.2 nm away from the binding site and free energy differences were up to 1 kcal/mol. Differences were insignificant for His91 which is 2 nm away from the binding site.

Essex next discussed the role of water molecules in ligand binding. A strategy in drug design is to consider enhancing the affinity of lead molecules with structural modifications that displace water molecules from a protein binding site. The binding may be primarily entropy-driven (release of ordered solvent molecules around the isolated molecule results in a net increase of entropy of the system) or enthalpy-driven, where noncovalent forces such as electrostatic attraction, hydrogen bonding, and van der Waals forces are primarily responsible for the formation of a stable complex.

Crystallographic structures can be used to locate the waters (if a good crystal and a suitable complex are available) but there are limitations (unclear electron density, the static nature of the snapshot, and the possibility of missing delocalized water sites). Other experimental methods exist but there is not much available data. The Grand Canonical Monte Carlo (GCMC) method, which uses physics-based force fields, is an alternative.[26,27] In essence, GCMC can be considered as an enhanced sampling method, as water molecules can be created and annihilated within a given cavity, circumventing the kinetic barriers that would be encountered in MD. Essex presented the example of galantamine (a treatment for Alzheimer's disease) bound to acetylcholinesterase. The X-ray structure (PDB structure 4EY6) has 11 waters; GCMC found 16 waters, including all 11 in the PDB structure (within <1.5 Å). Of the five novel structures, two are of particular interest.

Why does this matter? In the case of scytalone dehydratase (SD), Chen *et al.*[28] found that replacement of a nitrile in the ligand with a hydrogen atom lowered binding affinity 100-30,000-fold.They confirmed that the nitrile functionality displaced the water molecule as intended and that a favorable orientation was created with tyrosines 30 and 50 which had been part of the hydrogen-bonding network with the water molecule. Jorgensen's team carried out free-energy perturbation calculations in the context of Monte Carlo statistical mechanics simulations to investigate ligand series that feature displacement of ordered water molecules in the binding sites of SD, mitogen activated protein kinase p38 (p38-$\alpha$MAP kinase), and epidermal growth factor receptor (EGFR) kinase.[29] The change in affinity for a ligand modification was found to correlate with the ease of displacement of the ordered water molecule. In the EGFR example, the binding affinity may diminish if the free-energy increase due to the removal of the bound water molecule is not more than compensated by the additional interactions of the water-displacing moiety.

Issues can arise in relative ligand binding free energy simulations if the ligands considered have different active site water networks, as simulations are typically performed with a predetermined number of water molecules in preassigned locations. If an alchemical perturbation is attempted where the change should result in a different active site water network, the water molecules may not be able to adapt appropriately within the time scales of

the simulations-particularly if the active site is occluded. By combining the grand canonical ensemble with conventional free energy methods, the water network is able to adapt dynamically to the changing ligand. Essex's team refer to this approach as grand canonical alchemical perturbation (GCAP).[30] They have demonstrated GCAP for SD and for adenosine A2A receptor.

Physics-based simulations of biological molecules can provide useful thermodynamic and structural information and are now widely used as part of the drug discovery process. Experimental structural information is often incomplete yet subtle differences in structure can have a significant impact. Flexibility and dynamics are important.

## 6 Predicting metalloproteomes by machine learning

**Chu Wang, Professor of Chemical Biology, College of Chemistry and Molecular Engineering, Peking University, Beijing, P. R. China.**

The proteome is the entire set of proteins that is, or can be, expressed by a genome, cell, tissue, or organism at a certain time. An organism's complete proteome can be conceptualized as the complete set of proteins from all of the various cellular proteomes. With the rapid development of modern mass spectrometry (MS) based proteomic technology, we already know the overall proteome composition and architecture of multiple organisms. A draft map of the human proteome using high-resolution Fourier-transform mass spectrometry was reported in 2014,[31] but more challenging questions remain to be answered. What are the functions of these proteins and how are their functions regulated?

In order to complete a functional map of the whole proteome, we need to develop new high-throughput methods to aid discovery of functional sites in these proteins, which include but are not limited to catalytic, ligand binding, and post-translational modification sites, as well as protein-protein interaction interfaces. Using his background of cross-research in chemical biology, proteomics and theoretical calculation, Wang has developed and integrated a series of tools and methods relating to chemical probes, quantitative mass spectrometry, and theoretical calculation, has found different types of protein functional sites, and has discovered the precise regulation of them.

His team has carried out chemoproteomic profiling of functional post-translational modifications (PTMs): carbonylations,[32,33] O-GlcNAcylations,[34–36] N-homocysteinylations,[37,38] and itaconations.[39–41] For example, they have developed a specific thiol-reactive probe for quantitative chemoproteomic profiling of cysteine modifications by itaconate, an anti-inflammatory metabolite, and provided a global portrait of its proteome reactivity.[39] They have also worked on target deconvolution of bioactive ligands, including natural products,[42] endogenous metabolites,[43] and clinical drugs.[44]

In the current talk, Wang concentrated on the computational aspects of his research. His team has performed proteomic data mining to extract unique features from raw mass spectra, carried out 1D sequence analysis to extract local motifs, and performed structural modeling to aid discovery and confirmation of new ligand binding sites.

One example concerned selenoproteins a class of proteins in which selenium is inserted in the form of selenocysteine. Wang's team has developed a computational method called selenium-encoded isotopic signature targeted profiling (SESTAR)[45] to mine raw proteomic data and provide a comprehensive picture of selenoprotein distributions in human primary

hematopoietic cells and tissues. In another collaborative work, they developed a computationally aided and genetically encoded proximal decaging strategy[46] that enables time-resolved activation of a broad range of proteins in living cells and mice.

Haobo Wang, Xuemin Chen, and Can Li have worked with Wang in a machine-learning project on predicting reactive cysteines by local sequence motifs. Cysteine is one of the most intrinsically nucleophilic and chemically active amino acids in proteins, and also one of the most rarely used amino acids. High reactivity and redox properties give cysteines the power to play crucial roles in the structure, function, or regulation of a protein. In addition to stabilizing protein structures by forming covalent disulfide bonds, cysteines also perform critical catalysis, regulate intracellular redox potential, and coordinate metal ion cofactors in an enzyme's active site.

Weerapana, Wang and others[47,48] have carried out the isoTOP-ABPP method to quantitatively profile reactivity of functional cysteines in proteomes. The sites of 75 hyper-reactive cysteines (ratio <2) and 624 low-active cysteines (ratio >5) identified by isoTOP-ABPP were used in motif analysis. The team detected F, W and C enrichment near the hyper-reactive cysteine sites, and a CXXC motif which is known to constitute a reversible redox switch. They found a way to filter the motif using an algorithm called composition of k-spaced amino acid pairs (CkSAAP), which attempts to find useful amino acid pairs from a peptide segment. They then constructed a method called sequence-based prediction of cysteine reactivity (sbPCR)[49] to convert the peptide segment into a string of numbers using their modified peptide segment coding. These numbers can be processed by support vector machines to obtain a cysteine active classifier which produces excellent results: accuracy 97%, precision 95%, recall 89%, F1-score 92%, and ROC AUC 0.9777.

Out of 260,042 cysteines in the human proteome, 1058 were validated by MS, and of these 75 were predicted to be hyper-reactive. For all 13,367 organisms in UniProt, there were 2,725,789 cysteines, of which 90,532 were reactive according to sbPCR. This is far beyond experimental achievement. Wang and his co-workers verified the predicted hyper-reactive cysteines in *E. coli* using isoTOP-ABPP. The algorithm was 4.5 times more accurate than random guessing. Precision was 30%, accuracy 92%. The six true positive proteins were AldA, AldB, DsbA, FabH, FadA, and YecH. AldA and AldB are homologous proteins from the same family of dehydrogenases, the cysteine of which is reactive. DsbA contains sulfur reductase, and its CXXC motif is also significant. FabH and FadA are two enzymes involved in fatty acid synthesis.[50,51] Their cysteine reacts with coenzyme A. The last protein, YecH, is functionally unknown and Wang hopes to deduce its function. Further experimental work led to the discovery of a novel potential metal-binding site for Fe and Zn, and proof that the cysteine is important.

Wang next discussed another (unpublished) project, with Haobo Wang, Yao Cheng, and Yuan Liu: predicting metalloproteomes by sequence coevolution. At least 33% of the proteomes are metal-binding proteins, and they are functionally diverse. In the aligned sequences of homologous proteins, some positions are conserved, some are variable, and some are coevolved, that is, changes in one of the residues are compensated by changes to the other. Global coevolutionary models of homologous proteins have the capacity to predict residue-residue contacts from sequence information alone and thereby facilitate tertiary and quaternary protein structure prediction. One such model is Generative REgularized ModeLs of proteINs (GREMLIN)[52–54] which allows coevolution to be viewed in a contact map. A point in a map can be studied in a 21x21 frequency matrix of 20 amino acids and a vacancy. The model is also generative, allowing for the design of new protein sequences that have the same statistical

properties as those in the multiple sequence alignments (MSA). High-throughput sequencing technologies now allow the construction of an MSA, and accurate coevolution signals can be disentangled. Detected coevolved pairs can be used as residue-residue contact constraints in protein structure modeling and prediction of protein-protein interactions.[55,56]

Chakrabarti *et al.*[57] have investigated the structural and functional aspects of coevolved residues. They found that metal binding and active sites are more frequently coevolved with other metal binding and active sites, respectively, and a high fraction of coevolved sites are located close to each other. Wang wanted to see if he could predict metal binding sites based on features from coevolution analysis. Metal-binding amino acids are cysteine (C), histidine (H), glutamic acid (E), and aspartic acid (D). Wang's team filtered 499 metalloproteins from the PDB, with 1275 metal-binding coevolved pairs of CHED residues (positives) and 7039 non-metal-binding coevolved CHED pairs (negative). In all 10 pairings (CC, CH, etc.), they found that metal-binding occurs more frequently in coevolved pairs than in random pairs. The frequency matrices showed that metal-binding pairs had unique coevolution patterns.

Using an analogy with deep learning classifiers involving images of cats and dogs, the researchers fed metal-binding amino acid frequency matrix data to a neural network and finally deployed the model to unknown amino acid pairs to determine whether a "dog" that has never been never seen is real or not. They detected coevolution signals from a contact map, extracted CHED residues, made frequency matrices for metal-binding and non-metal-binding pairs, and used the data in the matrices to train a CNN model to predict whether a pair of residues are likely to be involved in metal-binding or not. The classification model had 94% accuracy.

Another topic being studied is the characteristics of metal binding sites within a coevolutionary network. The predicted metal-binding pairs can be integrated into a network or cluster. The cluster patterns of metal sites can be studied. Clusters can be used to query a 2D motif bank and the topology search can lead to prediction of possible metal ions, for example, a zinc cluster or a metal-sulfur cluster.

Wang's team applied this method, named "MetalNet", to four prokaryotic species, *E.coli*, *B.subtilis*, *S.solfataricus*, and *H.salinarum*, predicting more than 2400 high confidence sites spread among 489 proteins. They have analyzed these sites in an attempt to rationalize the algorithm. First, they made a simple homologous protein comparison of predictive sites: the metal binding site in homologous proteins was transferred to the corresponding protein for functional annotation. Of the sites predicted with high confidence, 29% were present in metal binding homologous proteins, of which 83% did have a metal ion near the site and only 13% did not show up. The team also analyzed some corresponding sites and found that the distance distribution of predicted sites in homologous proteins is close to the situation observed in practice, showing that the algorithm is very reliable overall.

Examining the prediction dataset in more detail, and data that were not supported by the homologous structure, they found cases that were actually correctly predicted. They were able to make plausible explanations and show metal-binding sites missing in PDB. A large number of proteins await full annotation. The team has validated a novel zinc-binding site in Citx. MetalNet also predicted four out of the five known metal-binding proteins[58,59] in the human spliceosome complex and another 15 novel predictions await characterization. In future, MetalNet will be expanded to more eukaryotic proteomes. The team has been successful with cysteine but plan to make predictions on several other common catalytic amino acids. They also want to predict more complex catalytic sites or ligand-binding sites using sequence, structural and coevolutional information.

# 7 Multiscale simulation for chemical biology: from enzyme evolution to interactive drug design in virtual reality

**Adrian Mulholland, Centre for Computational Chemistry, School of Chemistry, University of Bristol, United Kingdom**

The full video of Mulholland's talk can be viewed here: https://youtu.be/wqP5tOmxAQs.

MD simulations are contributing to studies of the SARS-CoV-2 virus: they can help to characterize the function of viral and host proteins and have the potential to contribute to the search for vaccines and treatments.[60] A cryo-electron microscopy structure of the SARS-CoV-2 spike glycoprotein reveals that the receptor-binding domains tightly bind the essential free fatty acid linoleic acid in three composite binding pockets.[61] MD helped to identify the ligand, and characterize its binding; and helped show that linoleic acid binding stabilizes a locked spike conformation, resulting in reduced angiotensin-converting enzyme 2 interaction *in vitro*. MD simulations also suggest that retinoids, steroids, and vitamins may stabilize the closed conformation.[62] Other such simulations indicate that spike peptide Y674-R685 is solvent accessible and has affinity for nicotinic acetylcholine receptors (nAChRs): different interactions with different receptor subtypes have been suggested.[63] Mulholland showed a potential binding orientation of the spike protein with nAChRs, in which they are in a nonparallel arrangement to one another.

The main protease (Mpro) of SARS-CoV-2 is a key enzyme of coronaviruses and has a pivotal role in mediating viral replication and transcription, making it an attractive drug target for SARS-CoV-2. Jin *et al.* identified an inhibitor (N3) by computer-aided drug design, and then determined the crystal structure of Mpro of SARS-CoV-2 in complex with this compound.[64] Antiviral drug repurposing programs have targeted MPro.[65] Mulholland is part of a team that has simulated the inhibition process of SARS-CoV-2 Mpro with N3.[66] The free energy landscape for the mechanism of the formation of the covalent enzyme-inhibitor product was computed with QM/MM MD methods. The simulations showed a two-step mechanism, and gave structures and calculated barriers in good agreement with experiment. The team went on to design two new N3 analogues as potential inhibitors, and to model the mechanism of inhibition. The COVID-19 molecular structure and therapeutics hub provides a community-driven data repository and curation service for molecular structures, models, therapeutics, and simulations related to the disease. More than 200 groups have signed up to a community pledge on data sharing.[67]

Drug action is inherently multiscale: it connects molecular interactions to emergent properties at cellular and larger scales. Simulation techniques at each of these different scales are already central to drug design and development, but methods capable of connecting across these scales will extend our understanding of complex mechanisms and our ability to predict biological effects.[68] Karplus, Levitt, and Warshel won the Nobel Prize for Chemistry in 2013 for the development of multiscale models for complex chemical systems.[69–71] Combined QM/MM methods are important in modeling enzyme catalytic mechanisms. By treating the reacting species with a QM method (i.e., a method that calculates the electronic structure of the active site) and including the enzyme environment (protein and solvent) with simpler MM methods, enzyme reactions can be modeled.[72,73]

In one example, Mulholland's team carried out high-level *ab initio* QM/MM calculations to determine activation enthalpies and free energies for chorismate mutase and *p*-hydroxy-benzoate hydroxylase that were in excellent agreement with experimental results.[74] Enzyme reactivity was described quantitatively by transition-state theory. Using projector-based embedding, the team also combined coupled-cluster theory, DFT, and MM to compute energies for the proton abstraction from acetyl-CoA by citrate synthase.[75,76] By embedding correlated *ab initio* methods in DFT they eliminated functional sensitivity and obtained high-accuracy profiles in a procedure that is straightforward to apply. They have applied a related range of QM/MM methods to investigate the Claisen rearrangement of chorismate to prephenate, in solution, and in the enzyme chorismate mutase.[77]

Antibiotic resistance is a growing, global health threat. Carbapenems, "last resort" antibiotics for many bacterial infections, can now be broken down by several $\beta$-lactamases (carbapenemases). Mulholland and his co-workers have predicted carbapenemase activity through QM/MM dynamics simulations of acyl-enzyme deacylation, requiring only the 3D structure of the apo-enzyme.[78] QM/MM assays may help predict the effects of mutations and may be used in design of future antibiotics.[79] For class D serine $\beta$-lactamases, QM/MM simulations can reproduce differences in free energy barriers between enzymes; small differences in active site hydration determine activity against cephalosporins.[80]

Digressing, Mulholland discussed research on nAChRs. These receptors modulate synaptic activity in the central nervous system. The $\alpha 7$ subtype is a target for several conditions, including Alzheimer's disease and schizophrenia. Extensive equilibrium and nonequilibrium MD simulations, enabled by cloud-based high-performance computing, have revealed the molecular mechanism by which structural changes induced by agonist unbinding are transmitted within the human $\alpha 7$ nAChR.[81] The simulations reveal the sequence of coupled structural changes involved in driving conformational change responsible for biological function. Comparison with simulations of the $\alpha 4\beta 2$ nAChR subtype identifies features of the dynamical architecture common to both receptors, suggesting a general structural mechanism for signal propagation. Darya Shchepanovska and Rob Arbon have used machine learning to produce a colorful visualization of nicotinic acetylcholine receptor dynamics mimicking Georgia O'Keeffe's painting style.

Mulholland went into more detail about nonequilibrium MD simulations. Understanding allostery in enzymes and tools to identify it offer promising alternative strategies to inhibitor development. In conjunction with two other groups, Mulholland's team has identified allosteric effects and communication pathways in two class A $\beta$-lactamases, TEM-1 and KPC-2, through a combination of equilibrium and nonequilibrium MD simulations.[82] The nonequilibrium simulations reveal pathways of communication operating over distances of 30 Åor more. Propagation of the signal occurs through cooperative coupling of loop dynamics. Notably, 50% or more of clinically relevant amino acid substitutions map onto the identified signal transduction pathways. This suggests that clinically important variation may affect, or be driven by, differences in allosteric behavior, providing a mechanism by which amino acid substitutions may affect the relationships among spectrum of activity, catalytic turnover, and potential allosteric behavior.

Arcus and Mulholland have coined the phrase "macromolecular rate theory" (MMRT) to describe the temperature dependence of enzyme-catalyzed rates independent of stability or regulatory processes. The heat capacity ($C_p$) for the enzyme-substrate complex is generally larger than the $C_p$ for the complex of enzyme and transition state. A negative value for $\Delta C_p^{\ddagger}$ is the result of the enzyme binding relatively weakly to the substrate and very tightly to the

transition state. A number of hypotheses arise directly from MMRT including a theoretical justification for the large size of enzymes and the basis for their optimum temperatures.[83] Psychrophilic enzymes show significantly different activation parameters (lower activation enthalpies and entropies) from their mesophilic counterparts.[84] The increase in enzymic rates with temperature up to an optimum temperature is widely attributed to classical Arrhenius behavior, with the decrease in enzymic rates above optimum temperature ascribed to protein denaturation or aggregation. Mulholland's team has shown that it is the change in heat capacity associated with enzyme catalysis ($\Delta C_p^{\ddagger}$) and its effect on the temperature dependence of $\Delta G^{\ddagger}$ that determines the temperature dependence of enzyme activity in many cases.[83–86]

The researchers have also shown, by a combination of experiment and simulation for two enzymes (dimeric ketosteroid isomerase and monomeric alpha-glucosidase), that the activation heat capacity change for a catalyzed reaction can be predicted through atomistic MD simulations. The simulations reveal subtle and surprising underlying dynamical changes: tightening of loops around the active site is observed, along with changes in energetic fluctuations across the whole enzyme including important contributions from oligomeric neighbors and domains distal to the active site.[87]

Temperature influences the reaction kinetics and evolvability of all enzymes. To understand how evolution shapes the thermodynamic drivers of catalysis, Bunzel *et al.*[88] optimized the modest activity of a computationally designed Kemp eliminase[89] by nearly four orders of magnitude over eight rounds of mutagenesis and screening. The catalytic effects of the original design were almost entirely enthalpic in origin, as were the rate enhancements achieved by laboratory evolution, but the large reductions in $\Delta H$ were partially offset by a decrease in $T\Delta S$ and unexpectedly accompanied by a negative activation heat capacity, signaling strong adaptation to the operating temperature. Extensive MD simulations show that evolution results in the closure of solvent exposed loops and better packing of the active site with transition state stabilizing residues.[90] These changes give rise to a correlated dynamical network involving the transition state and large parts of the protein. This network tightens the transition state ensemble, which induces an activation heat capacity and thereby nonlinearity in the temperature dependence.

Mulholland's team use a framework, Narupa, developed in Bristol for interactive molecular dynamics in a multiuser virtual reality (VR) environment, combining rigorous cloud-mounted atomistic physics simulations with commodity VR hardware. It allows users to visualize and sample, with atomic-level precision, the structures and dynamics of complex molecular structures on the fly and to interact with other users in the same virtual environment.[91] A series of controlled studies has quantitatively demonstrated that users within the interactive VR environment can complete sophisticated molecular modeling tasks more quickly than they can using conventional interfaces, especially for molecular pathways and structural transitions whose conformational choreographies are intrinsically 3D. Interactive MD in virtual reality (iMD-VR) is facilitating research, communication, and creative approaches within the molecular sciences, including training machines to learn potential energy functions, biomolecular conformational sampling, protein-ligand binding, reaction discovery using on-the-fly quantum chemistry, and transport dynamics in materials.[92]

Mulholland's team has outlined an experimental protocol which enables expert iMD-VR users to guide ligands into and out of the binding pockets of trypsin, neuraminidase, and HIV-1 protease, and recreate their respective crystallographic protein-ligand binding poses within 5-10 minutes.[93] Following a brief training phase, iMD-VR novices were able to generate unbinding and rebinding pathways on similar timescales as iMD-VR experts. The approach

has also been applied to both an Mpro inhibitor and an oligopeptide substrate,[94] using experimentally determined crystal structures. Docking with iMD-VR gives models in agreement with experimentally observed structures. The docked structures have also been tested in MD simulations and found to be stable.

iMD-VR has been applied in undergraduate education in computational chemistry,[95] teaching enzyme catalysis. A student survey indicated that most students found the iMD-VR component more engaging than the traditional approach, and also that it improved their perceived educational outcomes and their interest in continuing in the field of computational sciences.

## 8 An AI solution to the protein folding problem: what it is, how it happened, and some implications

**John Moult, University of Maryland, Rockville, MD, USA**

Since Anfinsen's famous experiments[96] in the 1960s, it has been known that the complex 3D structure of protein molecules is encoded in their amino acid sequences, and the chains autonomously fold under proper conditions. Moult summarized some of the methods that have been used since then. Structural principles include pathways, 2D structures, and fragment assembly. There have been notable successes in search methods: machine learning, genetic algorithms, reduced representations, and swarms. Physics-based methods (molecular dynamics, interatomic potentials, potentials of mean force, and funnels) have had no significant success. In evolutionary relationships from structures, single templates and multiple templates were useful, and loop building and refinement were used, but threading did not work. In evolutionary relationships from sequences, contacts and deep learning are being used.

In the 1980s, the community was not very good at rigor. In 1986, Moult himself made a fallacious claim ("...it is found that there is a wide spread of energies among the accepted [loop] conformations, and the lowest energy ones have satisfactorily small root mean square deviations from the X-ray structure..."). [97] Computational biology differs from traditional science in that it takes place in a virtual world. Achieving rigor in a computational world which the scientist controls is much harder than when dealing with the inflexible realities of the physical world. Community assessment experiments in computational biology were introduced to help achieve the same rigor as in real world science. Critical Assessment of Structure Prediction (CASP), the first framework for these experiments, is an organization that conducts double-blind, community-wide experiments to determine the state of the art of computational methods for modeling protein structure from amino acid sequence and other information. CASP1 was held in 1994; further CASPs have been held at two-yearly intervals since then, with continuing high participation rates (98 groups from 19 countries in 2020), and it has been accompanied by an enormous improvement in the accuracy of the protein modeling methods. Results are published in *Proteins: Structure, Function, and Bioinformatics*. Moult is founder and chair of CASP.

CASP encourages rigor, transparency, collaboration, and communication. The last two are now partially successful. CASP requires many targets (currently about 100). Each participant makes many predictions and there are many participants. A gold standard is needed. Methods are compared head-to-head, using clear metrics and authoritative evaluation. CASP categories are protein structure, protein assemblies, refinement, contacts and distances, accuracy estimation, and deriving function. The organizers and assessors are detailed on the Prediction Center website. Independent assessment is critically important. CASP1 was a rude awakening

but much progress has been made since then. Moult used geological timescale as a metaphor (Figure 4). We are not yet at the stage where no crystallography is needed but we are approaching it.

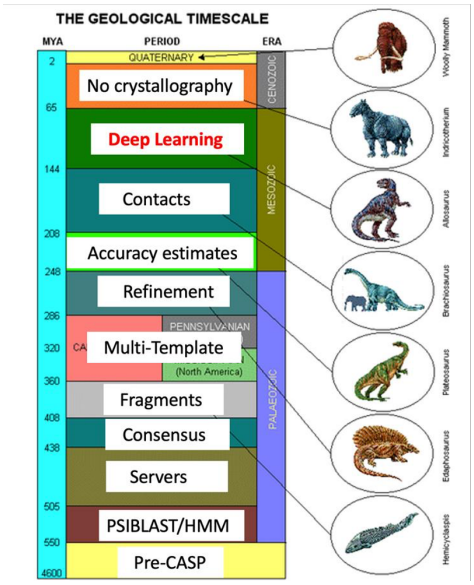



Figure 4. CASP progress.

In 1994, Göbel *et al.* interpreted correlated mutations observed in a sequence family as an indication of probable physical contact in three dimensions.[98] They reported a simple and general method to analyze correlations in mutational behavior between different positions in a multiple sequence alignment, used these correlations to predict contact maps for each of 11 protein families, and compared the result with the contacts determined by crystallography. For the most strongly correlated residue pairs predicted to be in contact, the prediction accuracy ranged from 37% to 68% and the improvement ratio relative to a random prediction from 1.4 to 5.1. This method was included in CASP2. Long range contacts, 24 residues apart can be relevant. There was no big improvement in precision in the years 2000-2014, although impressive results were reported for two targets 300 amino acids long in CASP11 in 2014.

The statistical model had been oversimplified: strong statistical dependencies are also observed for many residue pairs that are distal in a structure. In 2010, Burger and van Nimwegen[99] reported a Bayesian network model for disentangling direct from indirect statistical dependencies. Another model by Marks *et al.*[100] also made an impact. In CASP12, in 2016, precision increased from 25% to 47%. It is not necessarily true that contact accuracy leads to structural accuracy so a GlobalDistanceTest_TotalScore (GDT-TS)[101] metric is used in CASP. This metric is like an inverted ROC AUC. Methodological improvements led to an increase in precision[102] to 70% by CASP13.

More recent approaches are related to work on convolutional neural networks for image recognition. Convolution feature maps built to distinguish a dog from a cat are akin to a pile of contact maps which can be used predict protein structure based on alignments. Google DeepMind entered AlphaFold[103] into CASP13 (Figure 5). It was placed first in the free modeling (FM) category which assesses methods on their ability to predict novel protein folds. The company's approach builds on the use of co-evolutionary analysis to map residue covariation in protein sequence to physical contact in protein structure, and the application of deep neural networks to identify patterns in protein sequence and co-evolutionary couplings, and convert them into contact maps.[103] The Zhang Lab at the University of Michigan was

placed first in the template-based modeling (TBM) category which assesses methods on predicting proteins whose folds are related to ones already in the Protein Data Bank.

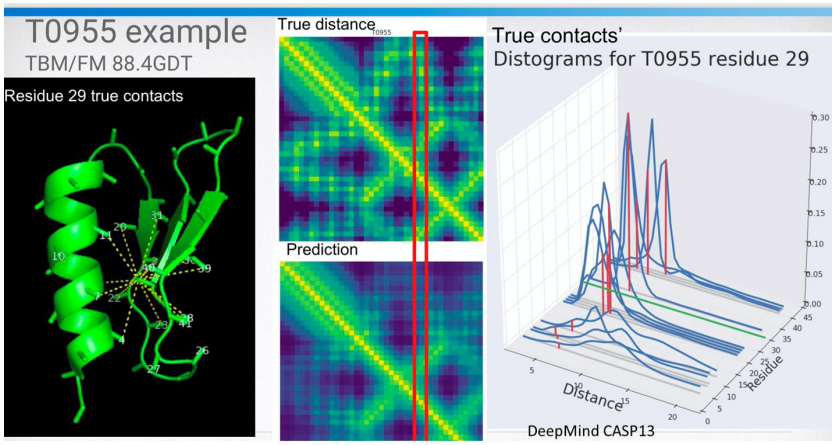



Figure 5. AlphaFold results in CASP13.

In CASP14 AlphaFold was the top-ranked method,[104] with a median GDT score of 92.4 across all targets and 87.0 on the challenging free-modeling category, compared to 72.8 and 61.0 for the next best methods in these categories. Figure 6 shows CASP14 results with and without DeepMind (group 427). According to Dan Rigden's CASP14 refinement assessment, the AlphaFold2-derived models were largely unimprovable, many of their apparent errors being found to reside at domain and, especially, crystal lattice contacts.[105]

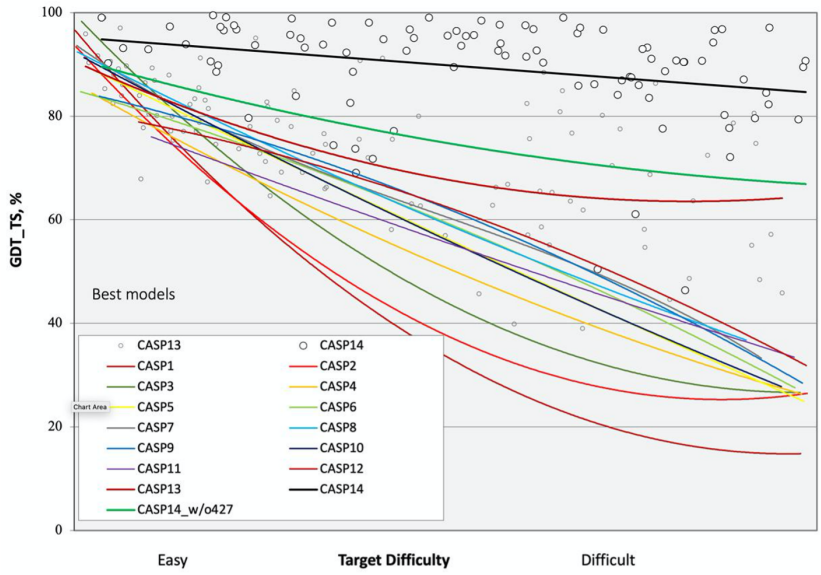



Figure 6. CASP best models.

What has DeepMind done to move so fast since CASP13? The CASP organizers do not demand code from the participants and at the time of Moult's presentation, DeepMind had not yet published details in a peer-reviewed journal.[106] Reportedly, key differences are as follows. The software does not stop at distograms: this is a full, end-to-end system with an iterative recycling stage for refining the structure. There is no separate gradient descent optimization (as there was in AlphaFold in CASP13). AlphaFold uses an attention-based neural network (instead of convolutions) that infers the implicit graph structure between

residues which is important for understanding the physical properties of the system. Some of the properties of protein physics are built into the network architecture so that it has information about factors such as local geometry constraints. The system has quite reliable uncertainty predictions, which is not very common in deep learning models, but these predictions are crucially important for downstream usability.

For CASP14, Baker's group has developed a deep learning framework (DeepAccNet)[107] that estimates per-residue accuracy and residue-residue distance signed error in protein models and uses these predictions to guide Rosetta protein structure refinement. The network uses 3D convolutions to evaluate local atomic environments followed by 2D convolutions to provide their global contexts and it outperforms other methods that similarly predict the accuracy of protein structure models. Incorporation of the accuracy predictions at multiple stages in the Rosetta refinement protocol considerably increased the accuracy of the resulting protein structure models.

One fairly large protein, a bacterial kinase with 402 residues, was very accurately modeled by DeepMind. The structure has one region that is experimentally disordered, and DeepMind produced more than one conformation for that region, showing that the method may be able to generate alternative conformations where they exist. The next step might be interdomain movement. Moult showed AlphaFold's best and worst structures using the T1024 interdomain angle (Figure 7). DeepMind submitted five models with a range of interdomain angles.

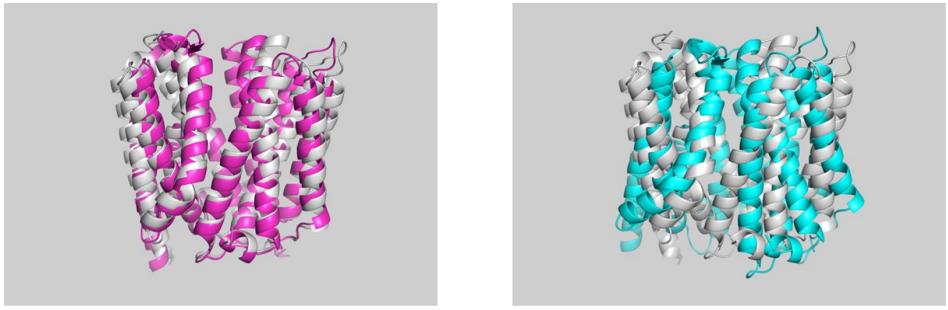


Figure 7. T1024 interdomain angle. AlphaFold best (on left) and worst (on right) models. (Experimental in white.)

Another example is SARS-CoV-2 ORF8 at atomic level (side chains in Figure 8). AlphaFold achieved subatomic accuracy almost everywhere, at the level of resolution in Figure 8. There are no training data like this in the PDB; the algorithm has generalized from the training data.

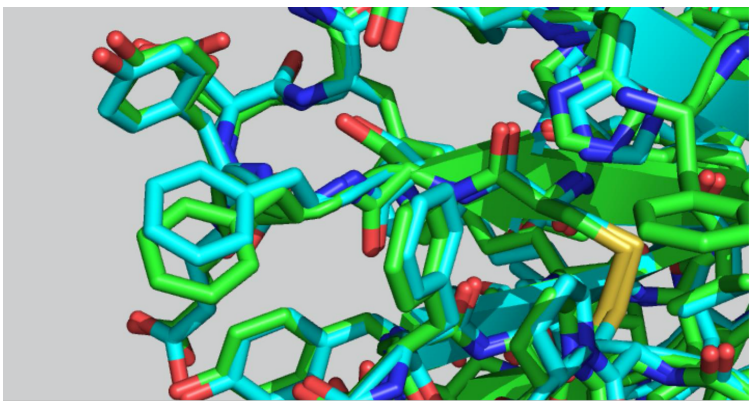


Figure 8. SARS-CoV-2 ORF8.

The physicist's approach to folding by first principles depends on representation at the atomic level, scoring using a force field, and searching using Newton's equations of motion. The method does not work well for proteins since they do not have a conventional force field.

CASP14 results have implications for structural biology. Models can be used for molecular replacement. Five structures were well-resolved so structure replacement could be used. Marcus Hartmann extended the method to all targets with diffraction data. This helped DeepMind. Crystallography will be speeded up by this approach.

What will be the impact of CASP14 on complexes? Some CASP14 model-model results are shown in Figure 9. DeepMind (participant 427) did extremely well. The Baker group was also successful in the docking section but they still have some way to go. There is a fold change when two proteins move together. Perhaps we should try both proteins as one molecule.

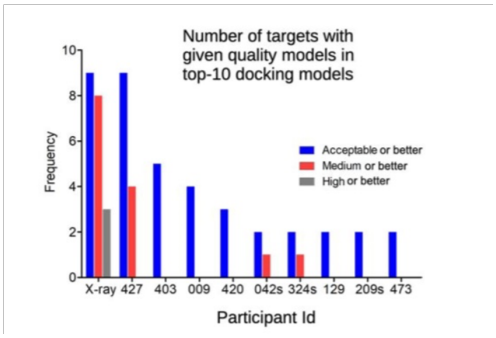



Figure 9. CASP14 model-to-model docking.

Moult turned to CASP14 function assessment and ligand docking.[108] In CASP13, David Koes used a voxel representation of a protein-ligand complex. A voxel grid is a geometry type in 3D that is defined on a regular 3D grid. A voxel can be thought of as the 3D counterpart to the pixel in 2D. Compare a red, green, blue pixel with a carbon, nitrogen, oxygen voxel. The only parameters for this representation are the choice of grid resolution, atom density, and atom type. Edwards' Critical Assessment of Computational Hit Finding Experiments (CACHE)[109] is looking at these methods.

CASP14 is not an end but a beginning. The door is now open to advances in protein complexes, accuracy estimation, protein design, protein conformational change, preferred conformations of disordered proteins, protein dynamics, ligand docking, and mutation interpretation. Do we really have a solution to the folding problem? Moult listed some comments that have been made. "Proteins have only one sequence, not a family." "What about the physics of folding?" "It's complexes that matter." "But this is only single, static structures." "Only DeepMind can do this." Moult is of the opinion that other groups are catching up with DeepMind.

Results from the CASP14 experiment show that new deep-learning methods have now provided a dramatic solution to the folding problem, with many computed structures comparable, likely sometimes better, representations of *in vivo* protein structures to those obtained with the state-of-the-art experimental techniques of crystallography and cryo-electron microscopy. These models have already demonstrated an ability to solve problematic crystal structures, and the results suggest the methods will be successfully applied to other areas of structural biology and more generally. The CASP methodology has now been adopted in a wide range of computational biology areas, including protein-protein interactions, genome sequence annotation, biological networks, and protein function annotation.

# 9 So you predicted a protein structure, what now?

**Thomas Steinbrecher, Schrödinger, Mannheim, Germany**

The full video of Steinbrecher's talk can be viewed here: https://youtu.be/e5ahKMf-inE.

Recent advances in technologies such as cryo-EM structure resolution and protein *de novo* folding prediction have resulted in a wealth of macromolecular structures that have not been resolved to the level of detail a high-resolution X-ray crystal structure could provide. Taking full advantage of these structures for rational drug design would benefit from additional validation and refinement. Schrödinger has investigated if computational refinement and structure-based modeling methods can be used to generate reliable complex poses.

Accurate 3D models of a protein-ligand interaction are essential for accurate structure-based drug discovery. Neither the protein nor the ligand has a single static 3D structure: both exist as an ensemble of conformations and those ensembles can change based on the environment. Proteins can induce conformational changes in ligands and *vice versa*. The goal of induced fit docking is to predict such changes in the 3D conformations of both the protein and the ligand.

A common scenario is obtaining a structure of novel chemical matter in a receptor the structure of which is known. Steinbrecher showed two structures, the crystal structure of thrombin in complex with a potent P1 heterocycle-aryl based inhibitor, PDB ID 1SL3, and a D-Phe-Pro-Arg-type thrombin inhibitor and the thrombin receptor, 1NZQ. The structures are nearly identical except for the motion of a few side chain atoms in Glu192. Without moving these few atoms, the 1NZQ crystal structure is incompatible with the ligand in 1SL3. The job of induced fit docking (IFD) is to predict this slight motion far more cheaply and quickly than experimentally obtaining the same result.

Schrödinger's initial IFD approach to cross-docking (IFD2006) was significantly better than rigid receptor fit (see below). Based on this promising, but insufficient performance, Schrödinger developed a significantly expanded IFD approach, IFD-MD. A successful docking tool must perform two tasks reliably: generate an accurate 3D model of the ligand-receptor complex (a sampling problem) and correctly rank the model (a scoring problem). Schrödinger has made improvements in both technologies.

Consider sampling first. The fundamental question in IFT methods is predicting which receptor atoms need to move. In the case of aldose reductase 2 (ALR2), for example, there are 32 non-Ala, and non-Gly residues within 8 Åof the ligand. If one were to permit no more than five residues to move, then 207,376 combinations would be involved. If each side chain had three rotamers, that would result in 50 million receptor conformations. Common attempts to tackle this problem involve inferring which side chain to sample based on solvent-exposed surface area, or crystallographic temperature factors, or motion observed over MD. These approaches are not reliable enough for prospective drug discovery projects. IFD2006 used only the first two approaches.

The correct motions are induced by the ligand yet the ligand cannot be docked in without first altering the receptor. This is a coupled problem which must be solved iteratively. The first step is to decouple the problem by docking the ligand into the receptor without considering receptor interactions. IFD-MD accomplishes this by ligand-to-ligand pharmacophore docking into a holo structure. (The structure of the protein bound to the ligand is known as a holo structure, the unbound protein as an apo structure.)

Pharmacophore docking enumerates thousands of conformations of the ligand. For each pharmacophore-docked conformation, clashes with the receptor are identified. These clashes are resolved in the presence of the ligand. The result is thousands of alternative unique receptor conformations. The ligand is redocked into these alternative receptor conformations, generating an ensemble of putative ligand-receptor complexes.

Scoring an IFD-generated complex is much more difficult than with rigid receptor docking. The receptor may make conformational changes to accommodate non-native binding modes. These changes could have unfavorable free-energy, but quantifying this is difficult. The IFD scoring function has to infer non-native contacts through MD investigation.

Metadynamics is an enhanced MD sampling method that artificially increases the potential energy of the system the longer it remains at a particular coordinate. The idea is to increase sampling of the MD simulation along this coordinate. The coordinate sampled over can be a complicated function of Cartesian coordinates. Similar to a reaction coordinate, it is called a collective variable (CV). Metadynamics allows the calculation to escape local minima over short time-scales.

In this case, the CV being sampled over is the RMSD of the ligand relative to its starting position. The system is trying to push the ligand away from its starting point. A native-like pose should resist this perturbation better than non-native poses over the same fixed simulation time scale. Unfortunately, the induced receptor changes can trap a non-native ligand into the binding site. This causes non-native ligands to appear artificially stable under metadynamics. These trapped poses invariably have bad ligand-receptor contacts, typically desolvation of polar groups. IFD-MD detects these bad contacts and counters the metadynamics stability term. For detection of desolvation to be reliable, highly accurate placement of explicit water molecules is necessary. Accurate placement of waters is done using GCMC. This allows hydration to be sampled on very short time-scales and it maintains computational efficiency.

The scoring function[110] is a sum of functions that consider molecular mechanics energy, rigid receptor docking terms, desolvation penalties, metadynamics stability, and side chain large motion penalties. A total of 18 parameters are optimized. Optimization is done on a grid but, because many terms are expected to balance one another, subsets of parameters are optimized together, to reduce dimensionality of the grid to something reasonable. For example, metadynamics stability and desolvation penalties are optimized together. In the workflow shown in Figure 10, the horizontal axis denotes time and dashed vertical lines denote a restartable checkpoint. The procedure takes 50 GPU hours and 400 CPU hours. Phase, Prime, Glide and WaterMap are Schrödinger products for, respectively, pharmacophore modeling, protein structure prediction, docking, and mapping the locations and thermodynamic properties of water molecules.

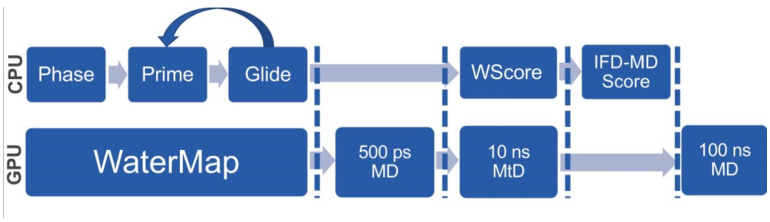
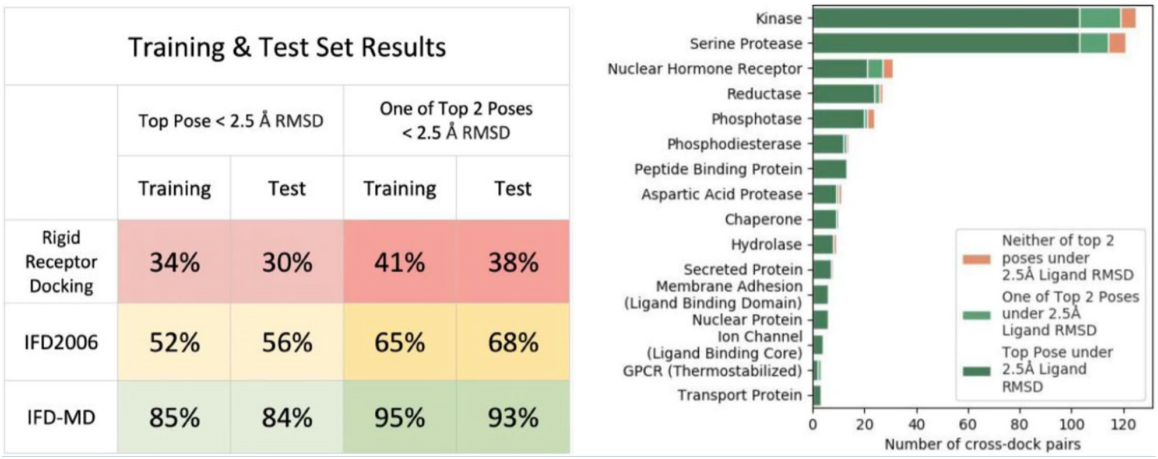



Figure 10. IFD-MD scoring workflow.

To test performance, IFD-MD was trained on a set of 258 cross-docks across 41 targets. The test set consisted of 157 novel cross-docks across a subset of these targets. The results for 415 retrospective cross-docking experiments taken from publicly available structures are shown in Figure 11. For comparison, the results from IFD2006 are included.



Training & Test Set Results

| | Top Pose < 2.5 Å RMSD | | One of Top 2 Poses < 2.5 Å RMSD | |
|---|---|---|---|---|
| | Training | Test | Training | Test |
| Rigid Receptor Docking | 34% | 30% | 41% | 38% |
| IFD2006 | 52% | 56% | 65% | 68% |
| IFD-MD | 85% | 84% | 95% | 93% |



Figure 11. Performance of IFD-MD.

In a usage scenario, the steps to create a validated ligand-receptor structure are to:

- select a known binder for structure determination
- run IFD-MD to generate the top two models for evaluation
- challenge the models with retrospective free energy perturbation (FEP), and
- select a good model (measured by using $R^2$ and RMSE) for prospective work.

The top two structures are selected, not just the top structure, because FEP should be run on two dissimilar ligand poses. If FEP performance is similar for both models, then the ligand dataset is insufficient to validate the structure with FEP. (FEP was explained by Jonathan Essex earlier in this report).

Steinbrecher presented a protein tyrosine phosphatase 1B (PTP1B) example (Figure 12) where there is a congeneric series with binding affinity data available. The IFD-MD with FEP task was to predict the PTP1B structure of the ligand with PDB ID 2QBS, starting from the PTP1B holo structure PDB ID 1C84.

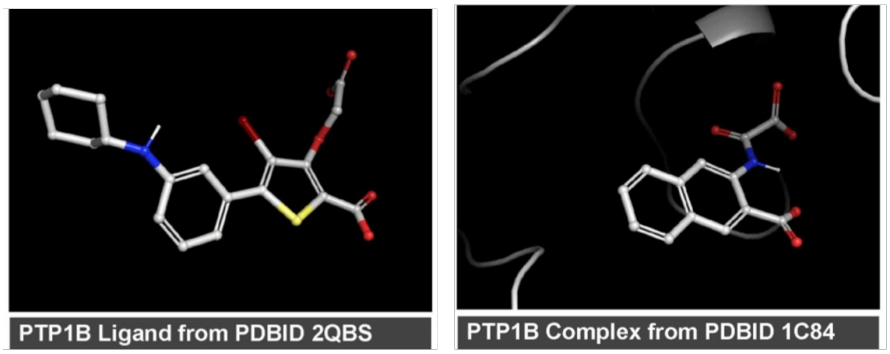



Figure 12. IFD-MD with FEP task.

FEP can distinguish between the top two ranked poses. In Figure 13, the pose on the right incorrectly flips the thiophene.

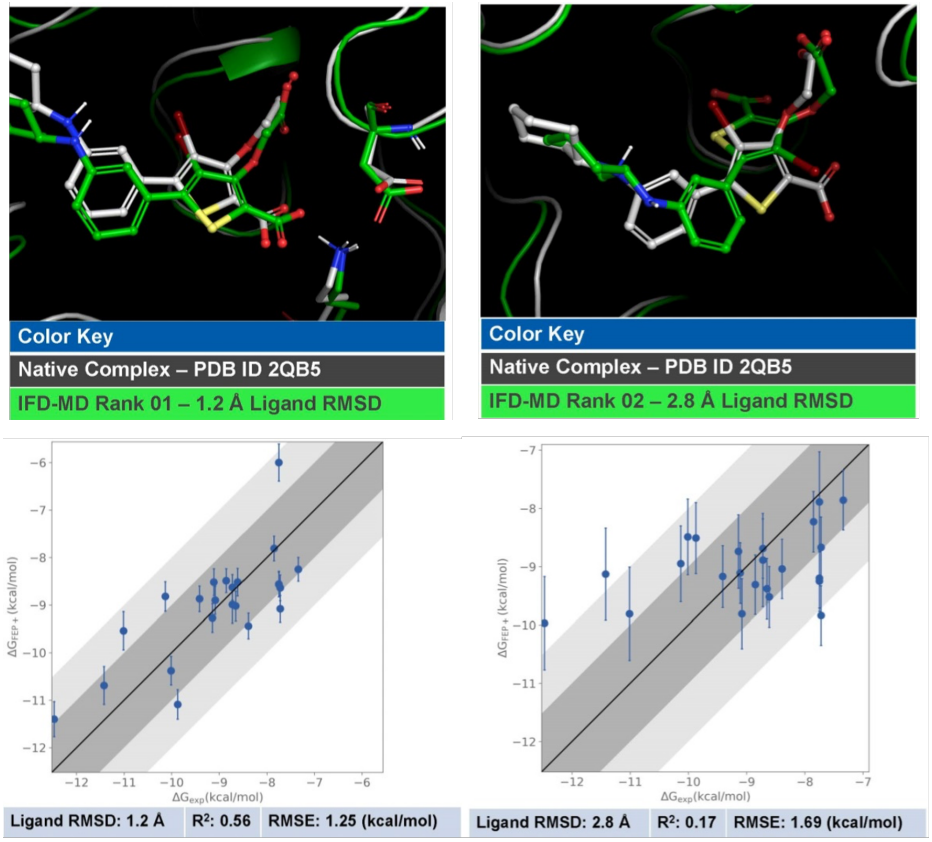



Figure 13. Distinguishing the top two poses.

In summary, in IFD-MD, a unique sampling strategy exploits existing holo structures to guide the generation of new receptor conformations, and rigorous scoring of ligand-receptor complexes is done by combining a docking-based scoring function and MD methods. IFD-MD allows for reliable generation of ligand-receptor complexes across a number of scenarios, including off-target prediction, HTS rationalization, and homology modeling.

## 10 Deep Learning enhanced prediction of protein structure and dynamics

**Martina Audagnotto, AstraZeneca, Gothenburg, Sweden**

The full video of Audagnotto's talk can be viewed here: https://youtu.be/E7eGHVybCH4.

Garegin Papoian has defined protein folding as an "intellectual problem at the intersection between biology, genetics, evolution, polymer physics, and statistical mechanics".[111] The dominant driving forces are H-bonding network, Van der Waals and electrostatic interactions, and hydrophobic and backbone angle preferences. Anfinsen showed[96] that the complex 3D structure of protein molecules is encoded in their amino acid sequences, and the chains autonomously fold under proper conditions.

It is well known that proteins in solution can reliably fold from a random coil to a unique native conformation on a biologically relevant timescale. Levinthal's paradox is an apparent contradiction between the number of possible conformations for a protein chain and the fact that proteins can fold to their native conformation quickly (less than a second). Levinthal initially estimated that for a protein there are $10^{300}$ possible conformations. If the protein

randomly sampled conformation space, it would not be able to fold correctly in a person's lifetime. This is the paradox: properly folded proteins are needed for the person to exist in the first place. For an unbiased random search, Levinthal's protein folding estimate is essentially correct, but if bias is introduced, the first-passage time to the fully correct state can be very much shorter.[112,113]

The Plaxco Simons Baker theory[114] is that folding kinetics can be predicted using simple, empirical, structure-based rules, suggesting that the fundamental physics underlying folding may be quite straightforward and that a general and quantitative theory of protein folding rates and mechanisms may be possible.[115]

Recent findings show that co-evolutionary analysis coupled with machine-learning techniques improves prediction precision by providing quantitative distance predictions between pairs of residues. The predicted statistical distance distribution from MSA has revealed the presence of different local maxima, suggesting the flexibility of key residue pairs. (See the talks by Wang and Moult earlier in this report.) Knowledge-based approaches, heavily based on the PDB, assume that similar sequences lead to similar structures. *De novo* prediction can be used even for proteins where there is no homologous PDB structure (or low similarity). See Moult's comments on the CASP13 experiment, earlier in this report. In CASP14, multiple groups used deep learning-based methods for *de novo* prediction.

In the current work, Audagnotto has investigated the ability of the residue-residue distance prediction to provide insights into the protein conformational ensemble.[116,117] She combined deep learning approaches with mechanistic modeling and applied the methodology to a set of proteins that experimentally showed conformational changes. The predicted protein models were filtered based on their energy score, and clustered by RMSD, and the centroids were locally refined. The models were compared to the experimental structure relaxed by classical atomistic MD simulation of the X-ray structure in explicit solvent, by analyzing the backbone residue torsional distribution and the side chain orientations. The workflow is shown in Figure 14.

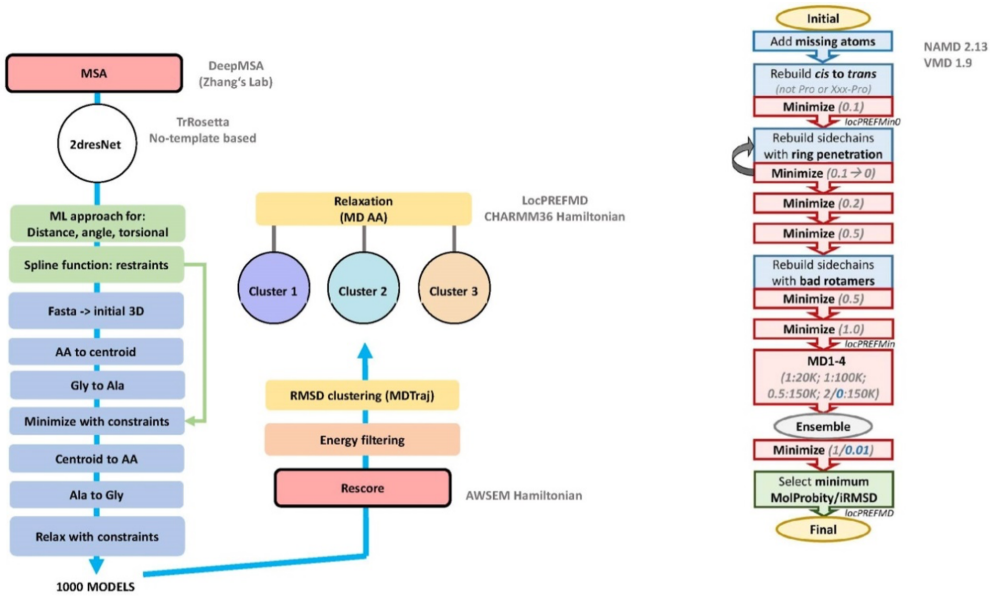
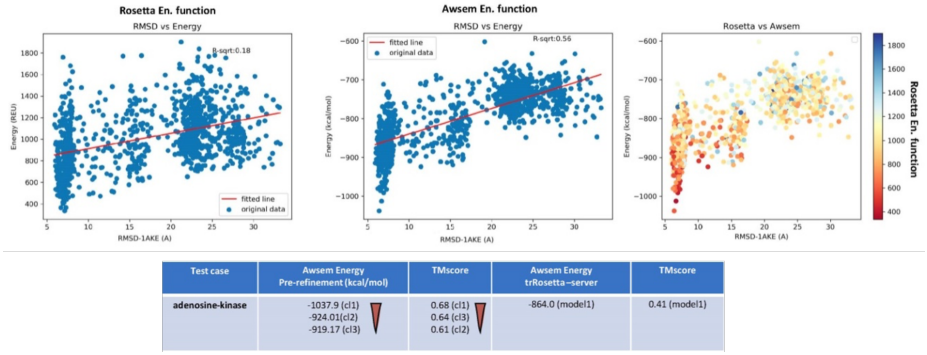



Figure 14. Pipeline from a 1D to a 3D structure.

DeepMSA[118] constructs deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. TrRosetta[117] is a Rosetta-constrained energy-minimization protocol from the Baker group. Associative memory Water Mediated Structure and Energy Model (AWSEM)[119] is a coarse-grained protein force field. PREFMD improves the overall quality of the predicted model with MD simulation, whereas locPREFMD[120] uses force field based minimization and sampling *via* MD simulations with a modified force field to bring bonds, angles, and torsion angles into an acceptable range for high-resolution protein structures. Another algorithm, ProSPr,[121] is an implementation of the AlphaFold protein distance prediction network. Adiyaman *et al.* have recently reviewed methods for the refinement of protein structure 3D models.[122]

Audagnotto had time to present only one of four case studies: adenosine kinase: PDB ID 1AKE in the apo form (R, 2.0 Å) and 4AKE in the holo form (R, 2.2 Å), using two experimental conformations. MD simulations were for AA (all atom), force field Amber99SB, without the small molecule. The challenge was the dynamic rearrangement as a function of small molecules (Figure 15). The other three case studies are compared in Table 2.




| Test case | Awsem Energy Pre-refinement (kcal/mol) | TMscore | Awsem Energy trRosetta-server | TMscore |
|---|---|---|---|---|
| adenosine-kinase | -1037.9 (cl1)<br>-924.01(cl2)<br>-919.17 (cl3) | 0.68 (cl1)<br>0.64 (cl3)<br>0.61 (cl2) | -864.0 (model1) | 0.41 (model1) |

Figure 15. Case study 1.

Table 2. Four Case Studies Compared

| Protein | Number of exp. Confs. | Number of pred. Confs. | CαRMSD Post-refinement | Challenge |
|---|---|---|---|---|
| adenosine-kinase | 2 | 2 | 3.6-4.2A | Dynamic rearrangement as function of small molecules |
| I-domain from the CD11a/CD18 | 2 | intermediate | 3.3-3.9A | Helix unfolding |
| Artificial Metalloproteins | 2 | intermediate | 3.6-4.2A | Pocket loop rearranagement |
| phage T4 lysozyme | 2 | 2 | 2.7-5.2 | N-terminal dynamics |

Switching between structural distinct states is common in proteins with the function of catalysis or molecular recognition. Alternative protein states are a largely open question in the field of protein folding. Current methods focus on the prediction of the experimental, static X-ray structure such as the one that is the best represented in the training set. MSA seems to encode the dynamic information necessary to explore the protein flexibility landscape. Audagnotto's pipeline was obtained from the combination of deep NN algorithms (DeepMSA and trRosetta), the molecular mechanics Hamiltonian AWSEM, and MD (in the relaxation step), exploring the potential correlation between residue-residue distance prediction and protein flexibility. Preliminary results showed a potential correlation between the dynamics of

the experimental structures and the dynamics of the predicted models. The pipeline not only allows one to retrieve the global experimental folding but also the experimental structural dynamics due to local and global conformational changes.

Based on these insights, Audagnotto is proposing a protocol that allows *in silico* investigation of protein dynamics for application in pharmacological research on catalysis and molecular recognition. In future, the work could have application in drug discovery, as a starting point for studying the binding motif of a small molecule on an unresolved protein structure; on the effect of mutagenesis on the protein folding prediction; and in membrane protein prediction (as a cryo-EM training test).

## 11 Fireflies Lévy flights algorithm for conformational optimization of peptides

**Zied Hosni, Antonio de la Vega de Leon, and Val Gillet, Information School, University of Sheffield, United Kingdom**

The full video of Hosni's talk can be viewed here: https://youtu.be/DWNjgmCaUQc.

Optimization algorithms are frequently used to guide the search in a conformational space of complex molecules such as proteins. It is a crucial step in accessing molecular properties corresponding to the most stable conformer but the optimizers are usually buried in docking software with limited tuning possibilities. Tabu search is a metaheuristic algorithm that can be used for solving combinatorial optimization problems (problems where an optimal ordering and selection of options is desired). The cuckoo search algorithm is another. Hosni has implemented an AI (metaheuristics) firefly algorithm with Lévy flights[123] distribution to search for the lowest energy conformations of peptides. His aim was to implement this algorithm (FLF), to design a fitness function (the force field energy of a peptide conformer); to optimize the performance of the algorithm (by tuning the hyperparameters); and to compare the performance of the FLF with the RDKit optimizer.

In the firefly algorithm, one firefly can be attracted to any other firefly, and attractiveness is proportional to the brightness of the firefly as determined by the landscape of the objective function. Firefly is a nature-inspired algorithm (Figure 16).[124] Table 3 is from a review of such algorithms by Kapur.

Table 3. Comparison of Techniques

| Algorithm | Inspiring organism | Inspiring behavior process | Communication action |
|---|---|---|---|
| ***PSO*** | Birds & fishes | Birds flocking & fish schooling | Coordinated behavior of swarm while flying |
| ***ACO*** | Ants | Food gathering phenomena | Pheromone deposition |
| ***FA*** | Fireflies | Attraction phenomena | Flash |
| ***ABC*** | Bees | Nectar gathering phenomena | Waggle dance |
| ***GA*** | - | Genetic phenomena | - |
| ***SA*** | - | Physical annealing | - |

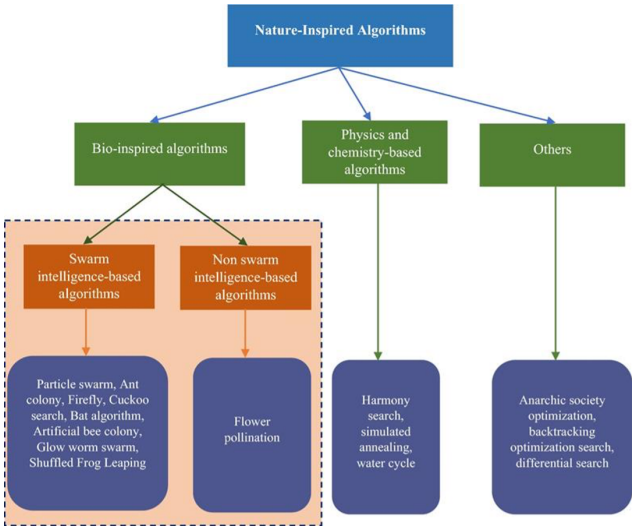



Figure 16. Nature-inspired algorithms.

An energy landscape[125] of protein folding and misfolding is shown in Figure 17. Hosni designed experiments as follows. A set of nine peptides with a broad range of sizes allowed him to challenge the performance of the FLF. The time of calculation was measured for each optimization. The absorption coefficient and the weight of the Lévy flights step were adjusted. Each optimization run was repeated five times to check the stability of the search to find the lowest energy conformer.

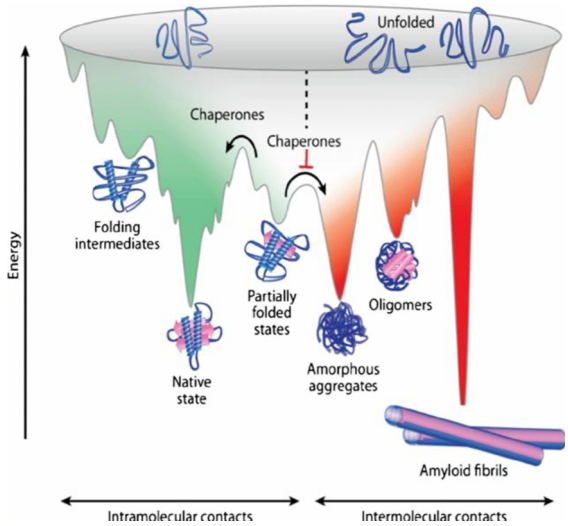



Figure 17. Energy landscape of protein folding and misfolding.

Figure 18 illustrates how an artificial firefly moves in space. Laboratory Virtual Instrument Engineering Workbench (LabVIEW) is a system-design platform and development environment for a visual programming language from National Instruments.

$$x_{i+1} = x_i + \beta_0 \exp(-\gamma r_{ij}^2)(x_j - x_i) + a\ sign\left(rand - \frac{1}{2}\right) \oplus Levy$$

Original position | Attractiveness | Randomness

Exploitation | Exploration

- $\beta_0$ is the attractiveness at $r = 0$.
- $\gamma$ is the light absorption coefficient.
- $i$ and $j$ are the index of the considered fireflies in the outer and inner loop of the algorithm, respectively.
- $r$ is the distance between the two fireflies of interest.

The product $\oplus$ means entrywise multiplications of matrices by the sign of a random number varying between -0.5 and +0.5. The random walk of the fireflies is a move in which the step-lengths have a probability distribution corresponding to the "Inverse Gamma Cumulative distribution" implemented in LabVIEW.

Figure 18. How an artificial firefly moves in space.

Lévy flights distribution is characterized by very small steps in all directions and occasionally very big jumps in the space. This approach allows a balance to be created between the local search that needs infinitesimally small steps and the global search that requires huge jumps, allowing the search in other areas of the space. The firefly algorithm is illustrated in Figure 19. An analogy is made between the artificial firefly and the peptide. The biological firefly moves in 3D space and the artificial firefly moves in $n$-dimensional space. The conformation of the peptide can be changed by changing the torsion angles. So, the number of dimensions is equal to the number of dihedral angles.

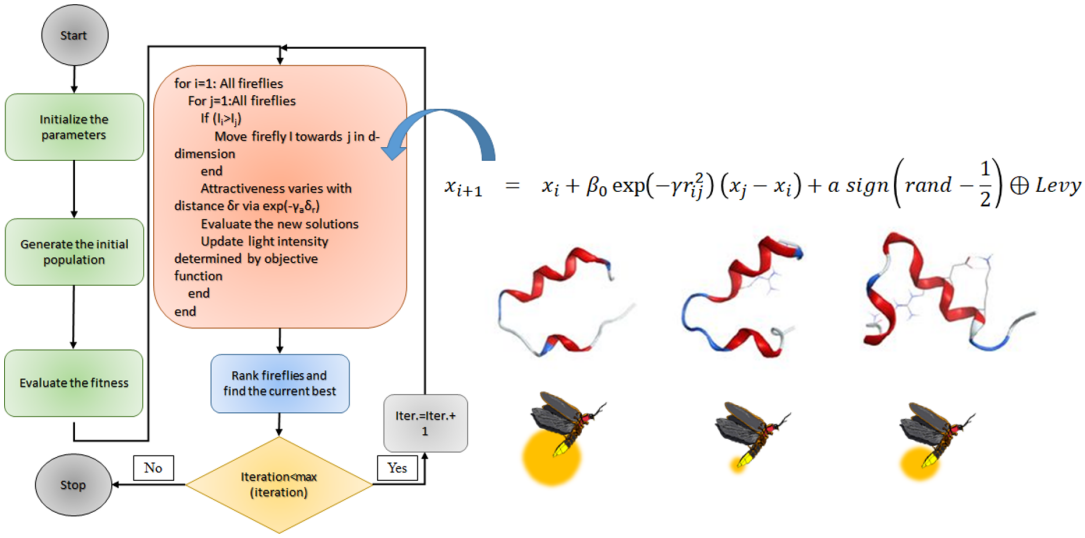



Figure 19. The firefly algorithm.

The Ackley function[126] is a non-convex function used as a performance test problem for optimization algorithms. Hosni used it to check the effect of the different hyperparameters. For example, reducing the population size reduces the number of evaluation functions but increases the instability of the algorithm. So the right compromise for the population size has to be found. Also, increasing the randomness weights increases the number of iterations, but reducing randomness weights increases instability.

FLF outperforms RDKit optimization in two thirds of the cases (Figure 20). For five runs on a subset of peptides it was shown that 2000 iterations were sufficient to reach convergence and, as expected, when the number of dimensions (torsion angles) was increased, more time was needed to reach convergence.

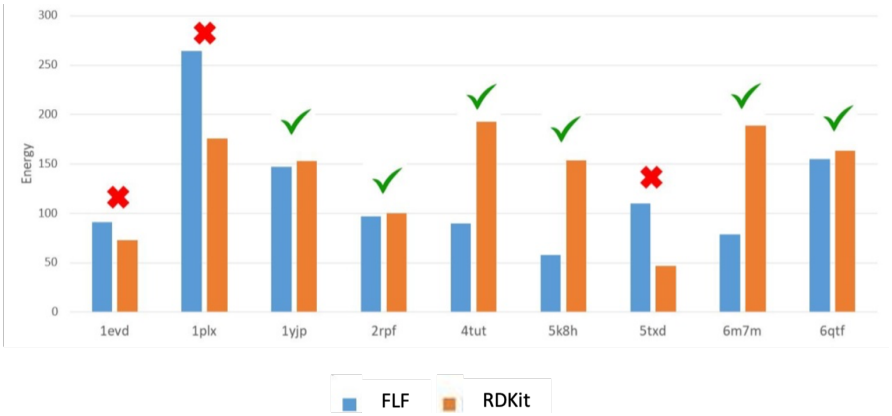



Figure 20. Final energy from FLF and RDKit optimizations.

In summary, the FLF algorithm was implemented successfully in LabVIEW and Python and was compared with and challenged against a state-of-the-art optimizer (RDKit). FLF outperformed RDKit in two thirds of the cases. The results show that the FLF algorithm is able to improve upon the genetic algorithm method with fewer energy evaluations; 2000 iterations were sufficient to reach the convergence in most cases. Hyperparameters, and especially the Lévy flights term, are crucial to avoid being trapped in local minima. Fine-tuning the right amount of randomness and balancing local search and global search are crucially important in controlling the performance of any metaheuristic algorithm.

In future, the research team plan to apply the FLF algorithm to a for real-life problem to by extracting certain fragments of bigger enzymes. Hosni showed a loop which is experimentally highly disordered; experimentalists would like to know the mechanism by which the loop expands to allow the DNA strand to go through a gap in the enzyme. By optimizing the structure of the linear loop and comparing low energy conformers with high RMSD to the initial linear conformations, FLF can give an explanation. The team also plan to implement the toolkit for online use.

## 12 How good are protein structure prediction methods at predicting folding pathways?

**Carlos Outeiral Rubiera, Department of Statistics, University of Oxford, United Kingdom**

The full video of Rubiera's talk can be viewed here: https://youtu.be/XBz9KZMDApE.

Google DeepMind's AlphaFold 2 indisputably won the CASP14 competition. The AlphaFold results are so incredibly accurate that many have hailed this code as the solution to the long-standing protein structure prediction problem. An item on the Oxford Protein Group's blog discusses what Google DeepMind's AlphaFold 2 really achieved, and what it means for protein folding, biology and bioinformatics. Deep learning has achieved unprecedented success in predicting a protein's crystal structure, but whether this achievement relates to a better modeling of the folding process is an open question.

AlphaFold 2 has solved the structural prediction problem but not the protein folding problem. More importantly, while AlphaFold 2 provides a general solution for protein structure prediction, this does not mean that it is universal. Several of the CASP14 targets were not predicted successfully. Folding is itself a fascinating question, of interest for basic biology but also for biomedicine, where it may better our understanding of the many diseases where misfolding is a cause or a terrible consequence. In designing drugs we do not know enough about targets and this problem depends on protein folding.

Outeiral and his colleagues have compared the dynamic pathways from six state-of-the-art protein structure prediction methods to experimental folding data.[127] The methods were Rosetta,[128] the classic fragment replacement program; SAINT2[129] from the Deane team at Oxford (a fragment replacement program with some enhancements); EVFold,[130] written by the Marks group at Harvard, using contact prediction and the CNS molecular dynamics software suite, DMPfold,[131] which uses deep learning, and RaptorX[132] and trRosetta[117] which use deep learning to predict distances, and also use relative torsion angles. RoseTTAFold,[133] was added in later comparisons. Outeiral *et al.* modified the programs so that they would predict folding trajectories. They used 200 decoys for the simplest programs and 10 in other cases.

They compiled a dataset of 170 proteins for which experimental folding kinetics data were available, by collating entries from the Protein Folding Database (PFDB)[134] of kinetic constants and the Start2Fold[135] directory of hydrogen-deuterium exchange (HDX) experiments. Many structures looked "unreal" so Outeiral reduced the data to fraction of native contacts between secondary structure elements, analyzed the trajectories using the fraction of native contacts between secondary structure elements (with STRIDE),[136] identified the main structure elements, mostly $\alpha$-helices and $\beta$-strands. Then he performed simple time trace analysis to identify which phases of folding were present, and interpreted them to get a coarse-grained view of what was happening in the trajectory. He compared his structures with experimental data: sequences and reference structures were downloaded from the RCSB PDB.

His first experiment was to try and predict the folding kinetics of the different proteins. The results are shown in Table 4. Unsupervised metrics employ a simple rule $c(x)$ that assigns a protein the most frequent kinetics. If 50% or more of the decoys display multistate kinetics, the protein is taken to fold in multiple steps; otherwise it is considered two-state. Supervised metrics fit a logistic regression on $c$(x) and report the average of 1000 five-fold cross-validation experiments. The baseline is a logistic regression that uses only the length of the protein (the number of amino acids). Accuracy reports the average recall per class, to account for the slight imbalance of the dataset (90 two-state folders and 80 multistate folders). The F1-score is the harmonic mean of recall and precision. The area under the receiver-operating curve (AUROC) for length is computed by projecting the values to the [0,1] interval.

Table 4. Performance of the Different Protein Structure Prediction Methods at Determining Folding Kinetics

| | trRosetta | RaptorX | DMPfold | EVfold | SAINT2 | Rosetta | Length | Contact order |
|---|---|---|---|---|---|---|---|---|
| *10 decoys* | | | | | | | | |
| Unsupervised accuracy | 0.615 | 0.568 | 0.534 | 0.632 | 0.629 | 0.633 | - | - |
| Unsupervised F1-score | 0.670 | 0.526 | 0.694 | 0.594 | 0.562 | 0.563 | - | - |
| Supervised accuracy | 0.608 | 0.575 | 0.612 | 0.588 | 0.614 | 0.639 | **0.656** | 0.500 |
| Supervised F1-score | 0.637 | 0.576 | 0.602 | 0.610 | 0.620 | 0.638 | **0.731** | 0.695 |
| AUROC | 0.714 | 0.631 | 0.574 | 0.668 | 0.705 | 0.725 | **0.725** | 0.574 |
| *200 decoys* | | | | | | | | |
| Unsupervised accuracy | 0.517 | 0.625 | 0.514 | 0.643 | - | - | - | - |
| Unsupervised F1-score | 0.675 | 0.605 | 0.688 | 0.611 | - | - | - | - |
| Supervised accuracy | 0.551 | 0.603 | 0.534 | 0.658 | - | - | **0.656** | 0.500 |
| Supervised F1-score | 0.644 | 0.617 | 0.691 | 0.688 | - | - | **0.731** | 0.695 |
| AUROC | 0.647 | 0.671 | 0.626 | 0.730 | - | - | **0.725** | 0.574 |

All the methods are significantly better than random, and are better than contact order (Plaxco, Simons and Baker[114] demonstrated that the average contact order of the native structure is strongly correlated with the folding rate constant of two-state proteins), but note that chain length outperforms any of the protein structure prediction methods at predicting folding kinetics.

Outeiral and his colleagues next examined whether the protein structure prediction methods can predict the folding rate constant ($k_f$) of the two-state processes. Plaxco, Simons and Baker showed that ln $k_f$ is strongly correlated with contact order. Outeiral found that most programs that he studied exhibit only a very weak correlation between the simulated trajectories and the folding rate constant. The Spearman correlation coefficients are not significant, at the 95% level of confidence, for trRosetta and RaptorX and DMPfold, and while EVfold, RaptorX and Rosetta display significant correlation, the correlation has the wrong sign: later folding events lead to larger (faster) rate constants. In contrast, the correlation between trajectories produced by RoseTTAFold and folding kinetics, although weaker in magnitude, has the correct sign. Nevertheless, all of the methods are significantly worse than the length of the protein chain at predicting the folding rate constant.

As, on occasion, structure predictors do correctly identify folding kinetics, the team next examined if in these cases the structures predicted in the pathway are consistent with experimental data. They hypothesized that if the structure predictor has insight into the multistate process, it should (1) predict structures that are congruent with experimental measurements, and (2) produce consistent predictions of the intermediates across independent replicas for the same protein. Hydrogen-deuterium exchange (HDX) experiments probe unfolded regions of a protein at different stages of the folding process and allow identification of which regions of an intermediate are structured and which have not yet folded. The researchers compared the predicted folding trajectories to these data.

They used the predicted trajectories to identify which pairs of secondary structure elements are interacting closely in the intermediate. This allows comparison between the noisy protein structure prediction pathways and the low structural resolution provided by experimental HDX data. For every protein and program, they considered a binary vector whose elements

correspond to pairs of secondary structure elements that are in contact in the native structure, with the Jaccard coefficient used to quantify similarity. They used the same trajectory analysis as was used previously to identify which pairs interact in the folding intermediate. The metrics of these classifiers are summarized in Table 5.

Table 5. Comparison Against HDX Characterization of Intermediates

| | RoseTTAFold | trRosetta | RaptorX | DMPfold | EVfold | SAINT2 | Rosetta | Random |
|---|---|---|---|---|---|---|---|---|
| *10 decoys* | | | | | | | | |
| Accuracy | 0.448 | **0.532** | 0.486 | 0.474 | **0.532** | 0.479 | 0.522 | 0.502 |
| F1-score | 0.210 | 0.192 | 0.217 | 0.000 | 0.125 | 0.000 | 0.039 | **0.252** |
| Jaccard | 0.052 | 0.052 | 0.052 | 0.054 | 0.052 | 0.052 | **0.166** | 0.094 |
| AUROC | 0.432 | **0.512** | 0.477 | 0.505 | 0.500 | 0.447 | 0.470 | 0.498 |
| | | | | | | | | |
| *200 decoys* | | | | | | | | |
| Accuracy | 0.453 | 0.534 | 0.495 | 0.489 | **0.540** | - | - | 0.502 |
| F1-score | 0.222 | 0.169 | 0.110 | 0.026 | **0.307** | - | - | 0.252 |
| Jaccard | 0.052 | 0.052 | 0.052 | 0.052 | 0.052 | - | - | **0.094** |
| AUROC | 0.441 | 0.503 | 0.502 | 0.492 | **0.530** | - | - | 0.498 |

Intermediates predicted by protein structure predictors are erratic and incompatible with available HDX data. There are only three or four cases for each method where the similarity coefficient is close to 1. Sometimes there is a different intermediate in every run. Outeiral concludes that these protein structure prediction methods are not good for studying kinetics, or rate constants, or intermediates.

Finally he discussed a case study concerning the mechanism of two-state folding of the small protein ubiquitin. The team analyzed the protein structure prediction trajectories for ubiquitin generated during their analysis, and as a physically-inspired baseline they considered a coarse-grained molecular dynamics (CGMD) simulation. Figure 21a shows the native structure of human ubiquitin (PDB: 1UBQ). Figure 21b shows the proportion of trajectories generated by each program that exhibit two-state dynamics. Figure 21c shows pairs of secondary structure elements that are formed in the identified intermediate. Black and white shaded cells represent the proportion of trajectories where an intermediate presents a given interaction, where white is 0 (does not appear in any trajectory) and black is 1 (appears in all trajectories). None of the RoseTTAFold trajectories exhibits an intermediate; hence none of the interactions is identified.

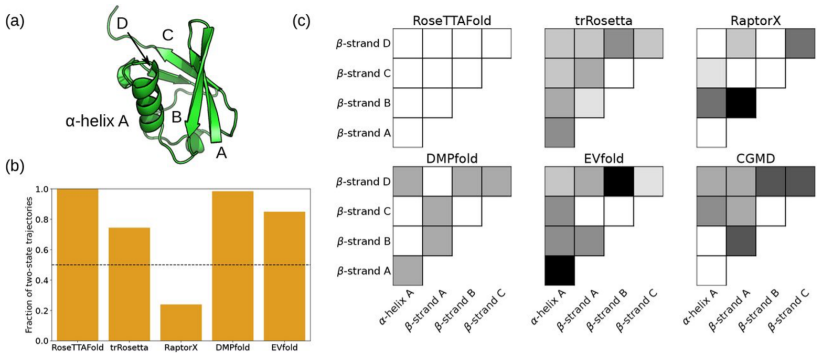



Figure 21. Simulations of human ubiquitin folding.

The researchers found that ubiquitin displays two-state folding kinetics in most of the simulated trajectories of RoseTTAFold, trRosetta, DMPfold and EVfold, but only in less than one quarter in RaptorX (Figure 21b). Surprisingly, CGMD simulations also suggested that the folding is multistate, with 62% of the trajectories exhibiting an intermediate. These results suggest that all codes, except potentially RoseTTAFold and DMPfold, are generating stable intermediates that do not reflect experimental kinetics.

Outeiral concludes that protein structure predictors are not telling us anything about the folding pathway. The "physics" induced by current deep learning methods is unrealistic, and does not correlate with experiment. Even if the protein structure prediction problem has been "solved", it looks as if the quest to understand protein folding has only just started. Refolding experiments are the utmost simplification of folding: how do we get better at modeling? Is there any other dynamical information of interest in deep learning?

## 13 Protein-ligand structure prediction for GPCR drug design

**Chris De Graaf, Head of Computational Chemistry, Sosei Heptares, Cambridge, United Kingdom**

The full video of De Graaf's talk can be viewed here: https://youtu.be/m8jwVmD4THg.

Sosei Heptares is an international biopharmaceutical group focused on the discovery and early development of new medicines originating from its proprietary Stabilized Receptor (StaR) technology, targeted at G Protein-Coupled Receptors (GPCRs), and its structure-based drug design (SBDD) platform capabilities.

GPCR drug discovery remains challenging. There are low expression levels, often with complicated expression and secretion pathways. Purification is difficult because of loss of structural integrity outside the membrane. Also, GPCRs are inherently flexible, changing conformation depending on the bound ligand. Sosei Heptares introduce point mutations into a GPCR, leading to increased thermostability. The receptor is trapped in a relevant conformation to match the drug product profile. Finally, the StaR can be extracted from the membrane and purified, with function retained. More than 70 stabilized receptors have been generated in agonist or antagonist conformations, leading to a new era of GPCR structure-based drug design.[137] The PDB has structures for 95 GPCRs, and 330 unique GPCR-ligand complexes; Sosei Heptares: has 300 structures in-house, 33 GPCRs, and 70 StaRs.

SBDD improves GPCR drug quality. Atom by atom optimization improves ligand efficiency. Polar contacts can be designed to control lipophilicity. The design of multiple structures allows for receptor flexibility and selectivity. Design for druglike properties increases *in vivo* efficacy and safety. An example is HTL1071/AZD4635 which is in a phase 2 clinical trial for patients with metastatic, castration-resistant prostate cancer.[138,139] The prediction of diverse binding modes has been a revelation.[140,141]

De Graaf gave examples of progressing from GPCR structure prediction to structural GPCR-ligand interaction prediction. IT1t is a potent C-X-C chemokine receptor type 4 (CXCR4) antagonist. Predicting the CXCR4-IT1t crystal structure was challenging, and the structure was featured in the GPCR DOCK 2010 assessment.[142] The CXCR4 amino acid sequence, some GPCR X-ray templates, and the 2D structure of the ligand were known. There were also some CXCR4-ligand structure activity relationship (SAR) and mutation data.[143]

The dilemma is that there is two-fold symmetry in the binding site and also symmetry in the ligand, and there are many different ways in which the ligand can fit.

The mutagenesis studies show that basic amine groups in the ligand match acidic $D^{2.63}$, $D^{4.60}$, $D^{6.58}$, and $E^{7.39}$ residues. SAR helped Sosei Heptares to prioritize plausible binding modes, and chemokine receptor broad chemogenomics and structural analysis were useful. In the new era of GPCR SBDD the modelers could tackle the sampling problem (GPCR-ligand binding mode diversity) and the scoring problem (GPCR binding sites are far more complex than just shape) and they could combine knowledge-based and physics-based approaches. In the GPCR DOCK 2010 assessment the Sosei Heptares team was the only one to predict the correct binding mode as one of their solutions.[142] From the GPCR DOCK challenges,[142,144,145] the team has learned that predicting the folds and overall helical TM conformation is no longer such a big challenge: prediction of the contacts is the big challenge.

X-ray and cryo-EM structures have revealed surprising GPCR ligand binding sites and binding modes. Water molecules play an important role in GPCR-ligand binding mode prediction, and membrane lipid molecules form a part of many GPCR ligand binding sites. Ligand dependent GPCR binding site conformational changes are important.

Another example where folding is difficult to predict is a Family A orphan GPCR with low sequence and structure similarity to GPCR templates. There were limited direct polar GPCR-agonist interactions: the fit was stereospecific, and the lipophilic GPCR-membrane interface played a role. The GPCR-agonist binding mode was challenging to predict based on knowledge transfer from other GPCR-ligand complexes. The GPCR TM helical conformation and loop conformations were impossible to predict from low homology GPCR X-ray or cryo-EM structure templates. An agonist StaR protein with improved expression and stability was used to determine the cryo-EM structure to 3 Åresolution.

De Graaf turned to an appreciation of the finer details of GPCR SBDD, for example, the small structural changes that SBDD can make in atom by atom optimization etc. GPCR structures allow detailed views of molecular interactions. Lipophilic hotspots have been found to be critical for binding in all structures. The lipophilic hotspots drive binding, displacing "unhappy" waters. These are often forgotten or underrepresented in pharmacophore models, leading to wrong models (as now shown by multiple ligand X-ray structures). WaterMap, from Schrödinger, and WaterFLAP, from Molecular Discovery, are complementary methods to predict water networks and estimate energy relative to bulk water. WaterMap is a physics-based method using MD simulations with explicit waters. WaterFLAP is a rapid empirical approach using GRID molecular interaction fields (MIFs), CRY lipophilic and hydrophobic probe, and entropy estimation. Other approaches are also available.[146,147] It is very important to model waters as well as the ligand and protein receptor to perform effective SBDD. Waters also influence pharmacophore models, where they often mediate some H-bonding from the ligand, and sometimes all the polar (H-bond) interactions.

Protein binding sites are far more complex than shape. One approach to modeling waters is water network prediction: studying the apo protein to locate unhappy waters, and designing a ligand which targets lipophilic regions and displaces unhappy water. Another approach is to look at the way the ligand influences the water network. In the adenosine $A_{2A}$ antagonists example[138,139] mentioned earlier, water network energetics determine receptor selectivity (Figure 22): $A_1$ traps an unhappy water molecule in a pocket which is not present in $A_{2A}$.[147] Very small changes in the shape and the properties of the binding site can have a dramatic effect on the water network. The discovery of the $A_{2A}$ receptor antagonist HTL1071/AZD4635

was a success story for AstraZeneca and Sosei Heptares in immuno-oncology.[138,148]

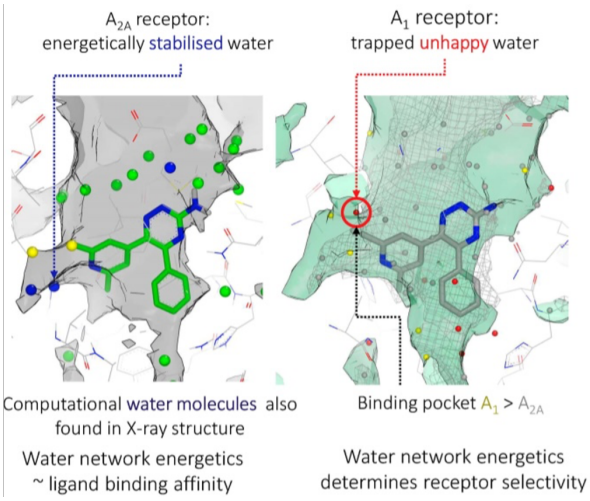



Figure 22. Ligand-perturbed water network.

Another example is work by Sosei Heptares on the orexin system, which consists of the two G protein-coupled receptors OX1 and OX2. Multiple PDB and in-house structures are available. Sosei Heptares has reported X-ray structures of the receptors in complex with 10 new antagonist ligands from diverse chemotypes, which complement the existing structural information for the system and highlight the critical importance of lipophilic hotspots and water molecules for these GPCR targets. Rappas *et al.* have discussed learnings from the structural information regarding the utility of pharmacophore models and how selectivity between OX1 and OX2 can be achieved.[149]

Predicting protein ligand binding affinities is of great interest for drug discovery. Sosei Heptares has been collaborating with Schrödinger on A2A antagonist affinity prediction with FEP,[150] not just using FEP+ out of the box but taking account of the importance of GCMC set-up. (See the talk by Essex earlier in this report for an explanation of FEP and alchemical free energy calculation). Standard protocols do not work for GPCRs: there are limited structural data for the majority of GPCRs (whereas, ideal targets have multiple high resolution structures), and small changes in the binding site have a big impact on FEP model predictivity. Membrane embedded systems have a complex input system; system plasticity, receptor flexibility and stability must be considered during simulations. A variety of natural ligands bind in a variety of binding sites with different physicochemical characteristics. Careful consideration of binding site solvation, shape, and protonation are critical.

De Graaf's final theme was a chemogenomic view to navigate structural GPCR-ligand interaction space: the identification of novel patterns and relationships between GPCR biological, chemical and structural data. How can we extrapolate the wealth of data on diverse binding modes and pharmacological data to other receptors? An example is finding new allosteric binding sites using StaRs..[151] C5a exerts a pro-inflammatory effect *via* the GPCR C5a anaphylatoxin chemotactic receptor 1 (C5aR1). Peptide antagonists based on the C5a ligand have progressed to phase 2 trials but these compounds exhibited several problems. NDT9513727 is a nonpeptide inverse agonist of C5aR1, and is highly selective for the primate and gerbil receptors over those of other species. To study the mechanism of action of C5a antagonists, Robertson *et al.* determined the structure of a thermostabilized C5aR1 in complex with NDT9513727. They found that the small molecule bound between transmembrane helixes 3, 4 and 5, outside the helical bundle. One key interaction between the small molecule and

residue Trp213$^{5.49}$ seems to determine the species selectivity of the compound. The structure demonstrates that NDT9513727 exerts its inverse-agonist activity through an extra-helical mode of action.

Systematic cheminformatics analysis of structurally and pharmacologically characterized GPCR ligands shows that cocrystallized GPCR ligands cover a significant part of chemical ligand space, despite their limited number. Many GPCR ligands and substructures interact with multiple receptors, providing a basis for polypharmacological ligand design.[140] Experimentally determined GPCR structures represent a variety of binding sites and receptor-ligand interactions that can be translated to chemically similar ligands for which structural data are lacking. This integration of structural, pharmacological, and chemical information on GPCR-ligand interactions enables the extension of the structural GPCR-ligand interactome and the structure-based design of novel modulators of GPCR function.

De Graaf and his colleagues[152] have developed scientific KNIME tools and workflows that enable the extrapolation of very unusual binding modes[151,153–156] to other systems. The tools enable structure-based bioactivity data mapping; structure-based identification of scaffold replacement strategies for ligand design; ligand-based target prediction; protein sequence based binding site identification and ligand repurposing; and structure-based pharmacophore comparison for ligand repurposing across protein families.

Site-directed mutagenesis data provide an invaluable complementary source of information for elucidating the structural determinants of binding of different ligand chemotypes. Vass *et al.*[157] have performed a comparative analysis of 6692 mutation data points on 34 aminergic GPCR subtypes, covering the chemical space of 540 unique ligands from mutagenesis experiments, and information from experimentally determined structures of 52 distinct aminergic receptor-ligand complexes. The analysis enables detailed investigation of structural receptor-ligand interactions and assessment of the transferability of combined binding mode and mutation data across ligand chemotypes and receptor subtypes.

Another project is BioGPS which is based on the software FLAP which combines GRID MIFs and pharmacophoric fingerprints. It comprises the automatic preparation of protein structure data, identification of binding sites, and subsequent comparison by aligning the sites and directly comparing the MIFs. Chemometric approaches are included to reduce the complexity of the resulting data on large datasets, enabling focus on the most relevant information. Individual site similarities can be analyzed in terms of their pharmacophoric interaction field (PIF) similarity, and the differences in their PIFs can be extracted.[158] A collaboration between Molecular Discovery and Sosei Heptares began in connection with this research.

In another example, de Graaf and his colleagues[159,160] have reported a virtual screening method that combines an energy-based docking scoring function with a MIF[161] to identify new ligands based on GPCR crystal structures. The interaction bit strings for a protein-ligand complex are compared to the interaction bit strings of binding poses predicted for other molecules, in order to identify chemically different molecules that can adopt similar binding modes. This approach has been successfully applied to GPCRs, with good hit rates for many nonpeptide experimentally validated ligands.[148,159,160,162–165]

Finally, de Graaf's team is collaborating with workers at the University of Cambridge, aiming to move toward AI-augmented GPCR drug design. Many current generative models are restricted to relatively data-rich targets, and ligand-based approaches often bias molecular generation toward previously established chemical space. Thomas *et al.*[166] have used Glide

docking (a structure-based approach) as a scoring function to guide the deep generative model REINVENT[167] and they compared model performance and behavior to a ligand-based scoring function. Also, they modified a known benchmarking dataset to remove any induced bias towards nonprotonatable groups and they proposed a new metric to measure dataset diversity. They report improved predicted ligand affinity. In addition, generated molecules occupy complementary chemical and physicochemical space compared to the ligand-based approach, and novel physicochemical space compared to known actives. Furthermore, the structure-based approach learns to generate molecules that satisfy crucial residue interactions. Overall, this work demonstrates the advantage of using molecular docking to guide *de novo* molecular generation over ligand-based predictors with respect to predicted affinity, novelty, and the ability to identify key interactions between ligand and protein target.

## 14 Using icospherical input data in machine learning on the protein-binding problem

**Ella M. Gale, School of Chemistry, University of Bristol, United Kingdom**

The full video of Gale's talk can be viewed here: https://youtu.be/qAZtfhEZe9c.

Gale started with a perspective on the importance of 3D shape (Table 6). Determining the binding coefficients of ligands to proteins is an essential step in targeted drug development. The 3D structures of both the protein binding pocket and the ligand are crucial in solving this problem.

Table 6. Importance of 3D Shape

| Problem | Importance of 3D shape and chirality to solution | Types of algorithms used for solution |
|---|---|---|
| Suggesting target molecules | Critical | QM, MM |
| *De novo* drug design | Critical | NN, clustering, search, generative |
| Host-guest and protein-ligand binding constants | Critical | General ML, QM, NN |
| Structure assignment | Critical | General ML, NN |
| Retrosynthesis | High (needs to be tracked) | NN, search |
| Forward reaction prediction | Medium (needs to be tracked) | General ML, NN, search |
| Condition recommendation | Low | NN, search |

How difficult is the 3D problem in chemistry? A child can recognize a cat in 3D at the age of about six months, can draw a cat in 2D from about 18 months, but needs 3-4 years to cope with 1D and symbolic representations such as writing the word "cat". Humanity and chemistry work in the other direction. 1D representations such a hieroglyphics were used in 3400 BCE and 2D cave paintings in 30,000 BCE, but 3D objects might take an infinite number of years; chemistry is much easier in 1D than in 2D and much easier in 2D than in 3D. Chemistry featurizations in 1D include fingerprints, physicochemical properties, and SMILES. In 2D, they are derived from the 2D formula (a molecular graph), and in 3D the data are sparse and large and are rotationally invariant and permutation-invariant. Gale proposed that a new way of inputting 3D information was needed.

Gale's second proposition is that augmentation is necessary. Since chemistry datasets are small and chemistry data are hard to get, we need to get the most out of small datasets. Molecules are not like cats. Using different rotations of a molecule is acceptable augmentation; using reflection operations is not. We need to preserve rotational symmetry, chirality, global structure, and translational invariance.

Gale presented three of her research projects: icosahedron projection, Spherical_NN, and graphs and topology for chemistry, all part of Icospherical Chemical Objects Surpassing Traditional A.I. Restrictions (ICOSTAR) through Replacing Existing Representations (ICOSPHERER), a new methodology which can be applied to the protein binding problem. Spherical_NN runs a spherical neural network (a type of 3D NN) and uses TensorFlow1. Calculation of topological features uses DeepChem, sklearn machine learning in Python and gtda, and TensorFlow 1 and 2. Figure 23 shows where these methodologies fit into the abstraction of the protein-binding problem.

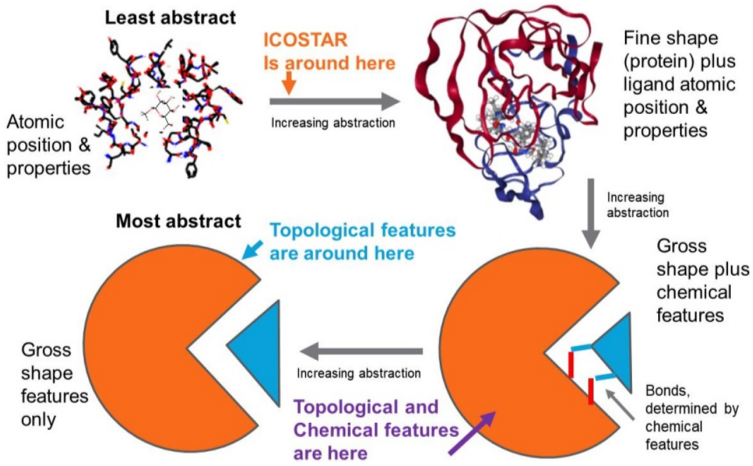



Figure 23. Abstraction of the protein-binding problem.

The PDBbind database is a collection of experimentally measured binding affinity data for the protein-ligand complexes deposited in the Protein Data Bank. It is a benchmark test set for the Comparative Assessment of Scoring Functions (CASF) challenge. The original PDBbind Core dataset had only 195 protein-ligand combinations: 1 bad, 1 medium, and 1 good ligand for each protein. The state of the art result with this set is Pearson $R^2 \sim 0.3$, root mean square error (RMSE) $\sim$ 1.9kcal/mol. The PDBbind Refined dataset has about 3000 protein-ligand combinations, of lower quality than those in the core set.

ICOSTAR uses icosahedron projection since icosahedrons offer the smallest distortion when going from 3D to 2D.[168] An icosahedron has 20 faces (equilateral triangles). It is the shape that gives the most symmetrical distribution of points, edges, and surfaces on a sphere. An icosahedron has 43,380 distinct nets. These are all the ways 20 equilateral triangles can be arranged to fold into the icosahedron. In ICOSTAR, a molecule is encapsulated in an icosahedron, atoms are projected from the center of mass to the surface of the icosahedron, the triangular faces of the icosahedron are colored (as if a light were in the center of the molecule and the atoms cast shadows on the surface of the icosahedron) and they are unfolded. For cubane there are 60 possible "nets" (unfoldings), some of which are shown (without hydrogens) in Figure 24. Another example is benzene, where some of the nets are the same because of the high degree of symmetry (Figure 25).

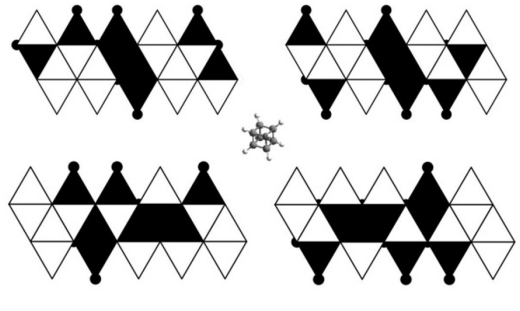


Figure 24. Some nets for cubane.

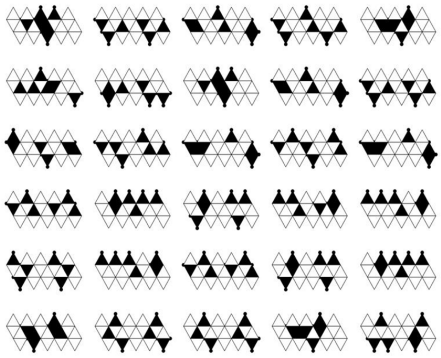


Figure 25. Some nets for benzene.

The method preserves chirality (Figure 26). In practice icosahedrons, with only 20 triangles, are not used but each triangle is divided into others to produce icospheres. The molecule (Tamiflu in Figure 27) is rotated and the surface is colored on the basis of which atoms hit it.

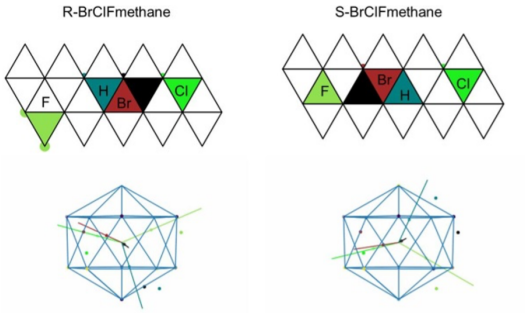



Figure 26. Preservation of chirality.

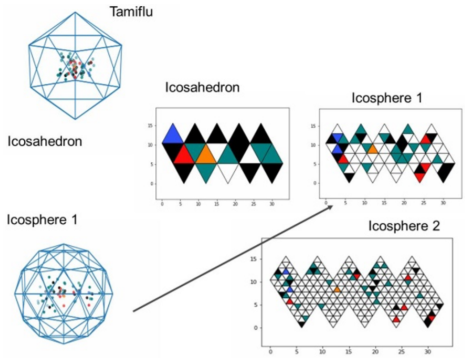



Figure 27. Icospheres: adding rotations.

Symbols (e.g., C, -COOH, $^{12}$C, +) are not used. Red, green, and blue (RGB) are used where R is the mass of the innermost atom, G is the mass of the outermost atom, and B is the sum of atom masses. Thus, C-H is [12.0107, 1.0078, 13.0185]; C=O is [12.0107, 15.999, 28.0097] and O-H is [15.999, 1.0078, 17.0068].

Gale applied ICOSTAR to PDBbind. She trained on 155 complexes from the core dataset and tested on 20 complexes from the core set. She also trained on the core dataset, and did a generalization test on the refined dataset of about 3000 complexes of less good quality. In the simplest architecture for the PDBbind problem, an icosphere for the protein is put into one spherical neural network and an icosphere for the ligand is put into another. Augmented data were used. The combined output from the two networks was compared with the ground truth. Control architectures were random forest and a neural network. Gale compared her results with the results from other algorithms used in CASF ("the state of the art").

Augmentation has the effect of equaling the state of the art and ICOSTAR generalizes: training on a small dataset gives the same level of accuracy for the largest dataset as for the small test set (Figure 28). Training takes around 20-30 epochs, fewer than the state of the art algorithms used. ICOSTAR trains quickly and after just one epoch (where the machine learns from seeing each example only once) the errors are not that far from the RMSE after final training.

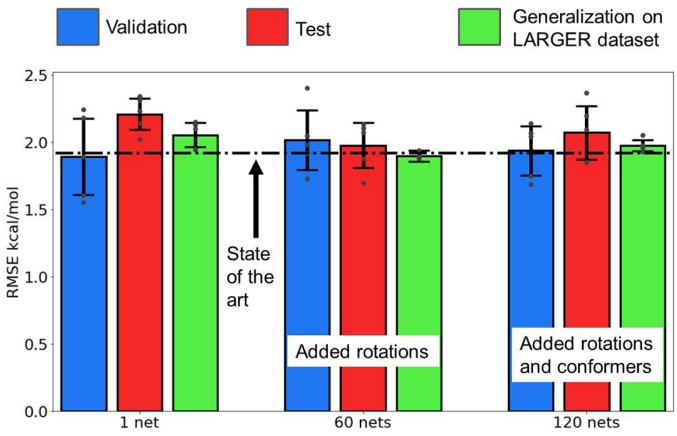



Figure 28. ICOSTAR and PDBbind.

Finally, Gale discussed topological features and first referred back to Figure 23. Topology is the mathematics of shapeshifting, stretching and pulling, but not tearing: a doughnut is not the same as a bread roll because it has a hole in it and a doughnut is not the same as a pretzel because a pretzel has two holes. The number of holes in a thing is called a topological invariant: it cannot be changed without tearing. A doughnut can be shaped into a teacup; both have one hole.

Persistence diagrams count things like the number of connected components. 0D persistent homology in Euclidean space can be explained as growing balls simultaneously around each point. As the balls around the points expand, 0d persistent homology notes when the balls touch. As a threshold is swept from negative infinity to infinity, the first interesting threshold value is 0. At 0, a connected component for each point is born; each one of these is represented by a ball with none of the balls intersecting. 0D persistent homology is tracking when these balls intersect. More specifically, it records when the ball in one connected component first intersects a ball of a different connected component. When the first set of two balls is touching, they will become part of the same connected component. This is the first death of a connected component, and thus the first point on the persistence diagram. As the process continues, further (birth, death) pairs are added to the diagram.

Persistence diagrams can be thought of as "mass spectroscopy for shape". They reveal local shape structure, counting all contiguous parts, loops and voids. Some of these loops will correspond to formal chemistry rings, some will not. Previous work[169,170] using topological methods in protein-ligand binding has been published.

In the current work, the points where all the atoms are in a protein binding pocket are made into a point cloud (other information is discarded). Connected points, loops, and voids are counted. Some chemical information is actually picked up (Figure 29). A similar procedure can be carried out with the ligand. Since persistence diagrams encode shape properties, they can show how the ligand might fit into a protein.

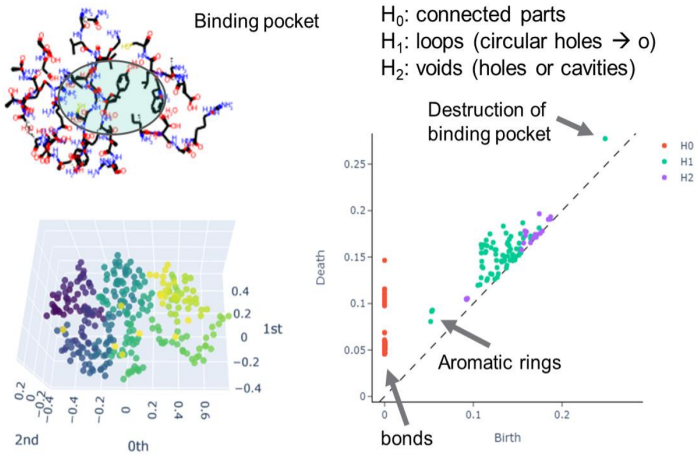



Figure 29. Protein topological features.

Persistence entropy, number of points, bottleneck distance, Wasserstein amplitude, and persistence image (each feature is 3D) are calculated from the persistence diagram to get topological features. The dimensionality of the topological features tested was either 6D (persistence entropy only) or 36D (all topological features). Topological features are fed into a simple, two hidden layer (2000, 1000 units) neural network and trained for 3000 epochs. Results from topological features equal the state of the art with small-sized features (Figure 30).

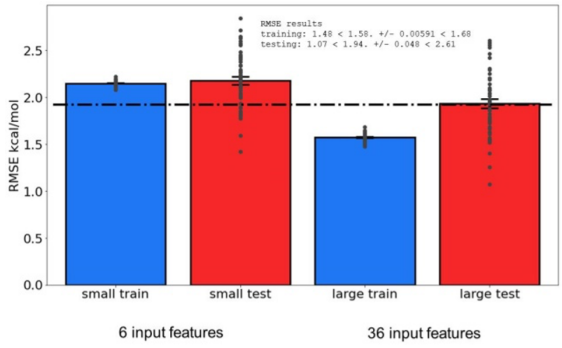



Figure 30. Topological features and PDBbind.

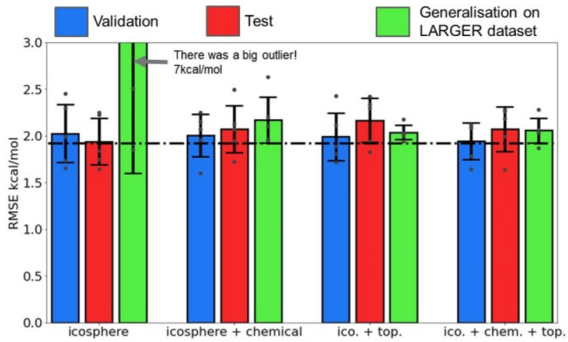



Figure 31. Use of combinations of features.

As a final experiment, Gale tried combinations of *all* features. Icospherical projection represents the scale-free, rotationally invariant 3D structure, atom types and masses, relative angles, and flexibility (with conformers). Topological features represent the scale of "chunks", 3D shape, and the presence of chunks, holes, voids, and tunnels. Chemical features represent physicochemical properties and the presence or absence, or counts of 15 functional groups. Use of all combinations of features equals the state of the art (Figure 31) and generalizes well (especially if the hyperparameters were adjusted to correct the green bar for the icosphere experiment). Gale also tried smaller and larger neural networks and the minimum energy (as opposed to RMSE) with encouraging results.

Possibly, Gale is approaching the limits of this particular dataset, but she *has* shown that augmentation improves generalization and that shape is important. Her methodologies have a number of advantages. ICOSTAR equals SotA with less training time, which might help in avoiding overtraining. When it is trained on only a small dataset, it can generalize to a larger, messier dataset. Even with small-sized features, results from the use of topological features are

equal to those of state of the art algorithms (meaning that the problem is being solved largely with shape information), and combining ICOSTAR and topological features beats state of the art algorithms.

In future, Gale would like to look at other small datasets. She also plans to put both sequence data and the 3D information encoded by moving atoms over the surface during rotation (Figure 27) into a recurrent neural network (RNN) as RNNs work on sequence data. This method does not lose any atom data, it might be faster, and it still allows for augmentation.

# 15 Lessons learned from generative models of biological sequences

**Aleksej Zelezniak, Associate Professor, SciLifeLab Fellow, Chalmers University of Technology, Gothenburg, Sweden**

The full video of Zelezniak's talk can be viewed here: https://youtu.be/3UeOFhsfFc8.

*De novo* protein design for catalysis of any desired chemical reaction is a long-standing goal in protein engineering because of the broad spectrum of technological, scientific, and medical applications. For example, we need an urgent solution to combat plastic pollution: 6.4 million tonnes are dumped into the ocean annually. If we continue like this, by 2050 there will be more plastic in the ocean than fish.

We could use the planet's microbiome potential to degrade plastics. We could identify all known plastic-degrading enzyme sequences from 30 million scientific articles, use bioinformatics to find more than 130 million related genes, and expand the known sequence space. Profile hidden Markov models (HMMs) are probabilistic models that encapsulate the evolutionary changes that have occurred in a set of related sequences (i.e., a multiple sequence alignment). To do so, they capture position-specific information about how conserved each amino acid is in each column of the alignment.

While biodegradation is a plausible route towards sustainable management of plastic waste, the global diversity of plastic-degrading enzymes remains poorly understood. Taking advantage of global environmental DNA sampling projects, Zelezniak and his colleagues have constructed HMM models from experimentally-verified enzymes and have mined ocean and soil metagenomes to assess the global potential of microorganisms to degrade plastics.[171] Metagenomics were studied in the Tara Ocean project, the global top soil project, and the Australian and Chinese microbiome projects.

An end-to-end synthetic biology framework for designing effective protein systems needs to tackle enzyme engineering and to enable the expression of proteins, maybe in new hosts. Frances Arnold and Kentaro Miyazaki presented saturation mutagenesis as an alternative method to random mutagenesis for obtaining enzymes with increasing stability. Both techniques were conceived to accomplish directed evolution, an approach honored by the Nobel Prize in Chemistry to Arnold in 2018. Directed evolution starts off with an enzyme that has properties similar to the desired ones. In her earliest work, Arnold created an enzyme that cleaves peptide bonds in organic solvents. The natural protein does this only in water: organic solvents change its structure and stop it working. Arnold then introduced random changes (mutations) into the gene that encodes the peptide-cleaving enzyme. Different versions of the mutated gene were then inserted into bacteria that started churning out many, slightly different enzymes. Arnold then selected the bacteria whose enzymes worked best in organic

solvents and subjected them to further rounds of test-tube evolution.

Generating functional protein diversity practically is very challenging: 75% of single amino acid mutations decrease the activity of enzymes and 50% of single amino acid mutations are deleterious to protein function. A protein of 100 amino acids can be made in $10^{130}$ ways; only 1 out of $10^{77}$ will be an active protein. Fitness landscapes depict how genotypes manifest at the phenotypic level and form the basis of our understanding of many areas of biology, yet their properties remain elusive.

Sarkisyan *et al.*[172] have visualized an extensive region of the local fitness landscape of the green fluorescent protein from *Aequorea victoria* (avGFP) by measuring the fluorescence of tens of thousands of derivative genotypes of avGFP. Fitness landscapes are often thought of as mountains but it seems that they are more like isolated islands of opportunity. It would be useful to be able to search in the right space and not in the dark matter between the islands.

There have been great advances in image recognition since 2014; the quality of human faces output has been greatly improved. Is current machine learning technology now ready for making the most complex molecules in the universe? Zelezniak and his co-workers have developed ProteinGAN,[173] a self-attention based variant of the generative adversarial network (GAN)[174] that is able to "learn" natural protein sequence diversity and enables the generation of functional protein sequences. ProteinGAN learns the evolutionary relationships of protein sequences directly from the complex multidimensional amino-acid sequence space and creates new, highly diverse sequence variants with natural-like physical properties.

The team used malate dehydrogenase (MDH) as a template enzyme. MDH is 300-350 amino acid long protein, active as a tetramer, with two active sites. It belongs to a diverse family with pairwise sequence identify as low as 40%. Zelezniak and his co-workers used about 15,000 sequences of training examples.

A discriminative model of statistical classification is a model of the conditional probability of the target $Y$, given an observation $x$, $P(Y|X=x)$. A generative model is a model of the conditional probability of the observable $X$, given a target $y$, $P(X|Y=y)$. A flow chart for ProteinGAN is shown in Figure 32.

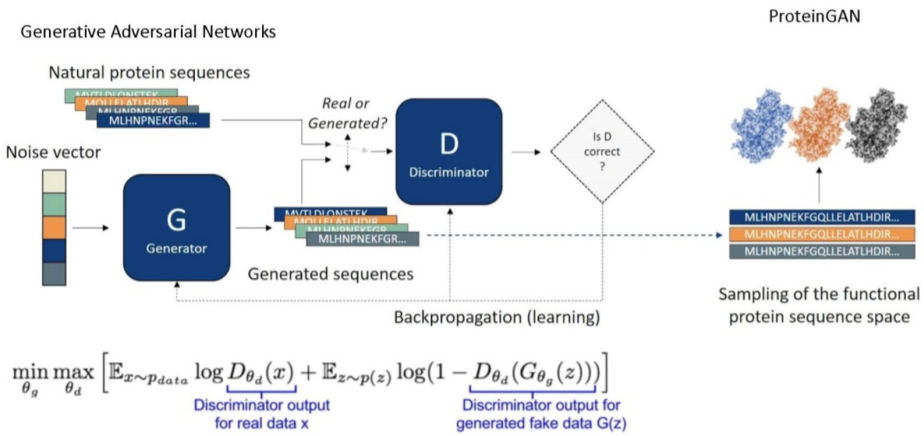



Figure 32. Teaching a computer to generate natural-like proteins.

It is not as easy to assess the quality of a generated protein as it is to judge if a facial image is of high quality. So the team started by looking at how the training progresses (Figure 33, left) and how overfitting is avoided. It was seen that after a week sequences were generated that looked

like real ones. Comparing a natural sequence from the test set with a generated one (Figure 33, right) was also good evidence.

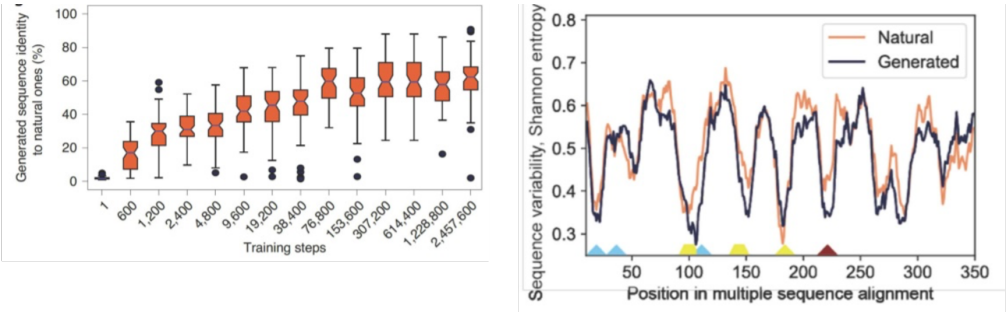



Figure 33. Assessing sequence quality and evolution of training.

HMMs are good discriminative models but they not capable of generating natural sequence variability (Figure 34).

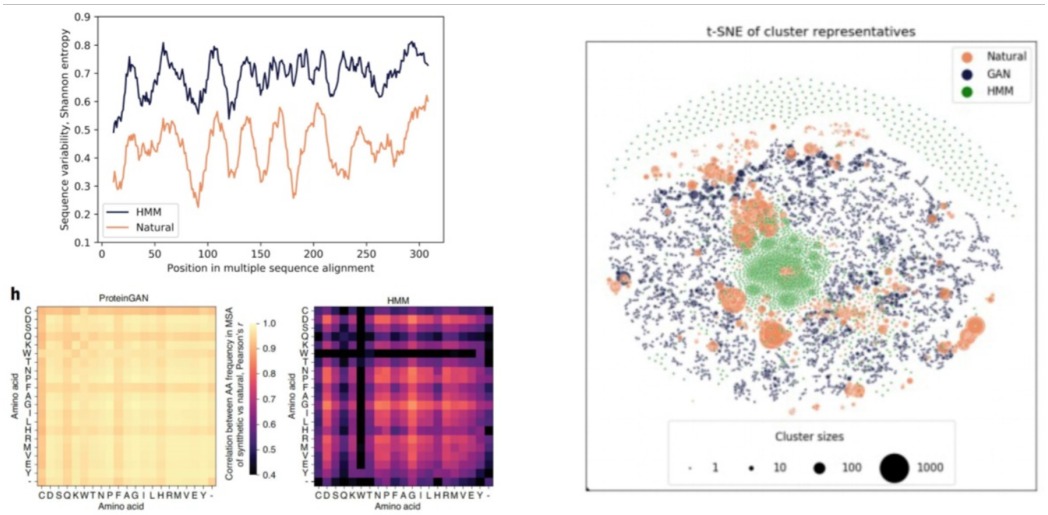



Figure 34. HMM model compared with ProteinGAN.

The choice of loss function is critical for sequence quality. The researchers were interested to find a literature precedent for this conclusion.[175] Architectures, on the other hand, did not influence sequence quality with one important exception: attention is important for long-range interactions (Figure 35). The sequences were generating active sites (Figure 35, right). The attention mechanism enables learning of long-range sequence interactions (compare "The animal didn't cross the street because it was too tired"). In a plot of median attention between pairs of amino acids (Z score) against median attention between pairs of amino acids (Å), $r = 0.5$, $p <$ 2.2e-16. Interpolation of latent spaces allows control of sequence diversity (Figure 36).

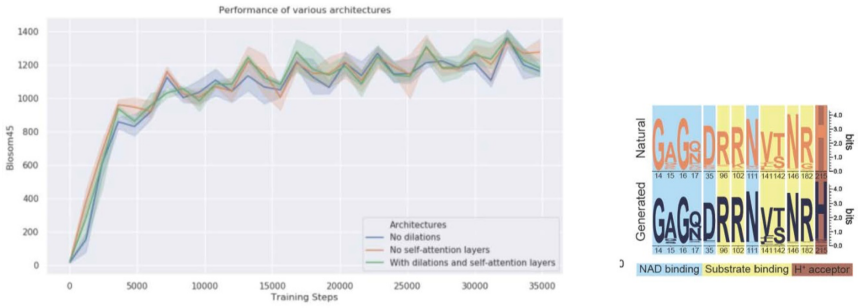



Figure 35. Attention is important for long-range interactions.

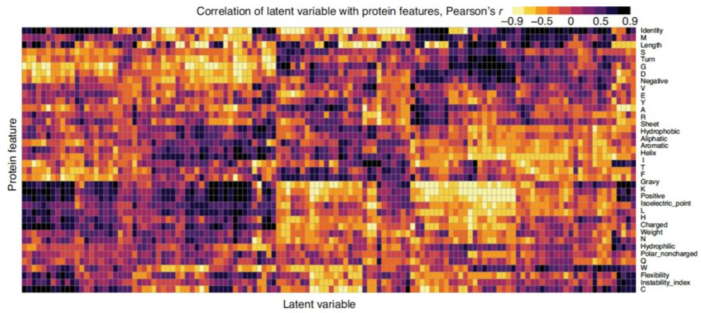



Figure 36. Latent spaces and sequence diversity.

Synthetic sequences are more diverse than naturally found MDH sequences (Figure 37). ProteinGAN generates highly diverse proteins which are as active as natural proteins (Figure 38).

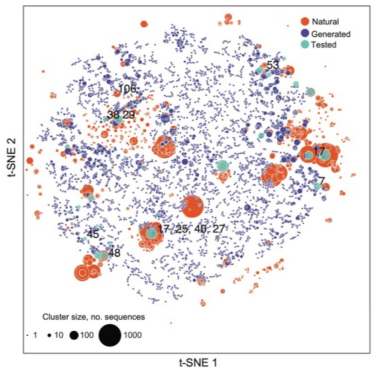



Figure 37. Diversity of sequences.

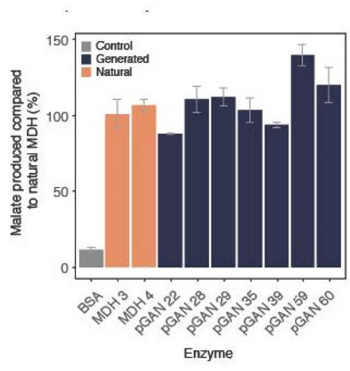



Figure 38. Generated proteins as active as natural proteins.

Zelezniak and his co-workers showed that 24% (13 out of 55 tested) of the ProteinGAN-generated and experimentally tested sequences are soluble and display MDH catalytic activity in the tested conditions *in vitro*, including a highly mutated variant of 106 amino-acid substitutions. ProteinGAN therefore demonstrates the potential of artificial intelligence to generate highly diverse functional proteins rapidly, within the allowed biological constraints of the sequence space.

Challenges remain. How do we "look" at the proteins? Algorithms for learning long-sequence dependencies are an active area of research. There is a lack of functional data, for example, for enzymatic activity improvement. Physics needs to be combined with generative models.

Solubility is an unsolved problem in recombinant protein production. Nevertheless, AI generates highly diverse active sequences. It learns the physicochemical constraints of protein families and enables navigation of functional protein space to change protein properties.[173,176]

## 16 DeepDock: a deep learning approach to predict ligand binding conformations

**Oscar Méndez-Lucio, Mazen Ahmad, and Jörg Kurt Wegner, Janssen Pharmaceuticals, Beerse, Belgium; Antonio Ehecatl del Rio-Chanona, Centre for Process Systems Engineering , Department of Chemical Engineering, Imperial College London, United Kingdom**

The full video of Méndez-Lucio's talk can be viewed here: https://youtu.be/j4qV6aew9cs.

Understanding the interactions formed between a ligand and its molecular target is critical in guiding the optimization of molecules. Different experimental and computational methods have been used to understand better these intermolecular interactions. One way of deciding if an interaction is possible is to examine the distance distribution for the frequency of occurrence of that interaction in the PDB. It is impossible to do this for all combinations of all fragments and all amino acids in all binding sites, so Méndez-Lucio and his colleagues developed DeepDock,[177] a method based on deep learning. DeepDock learns each fragment and each point in the protein surface as embeddings; predicts the distance probability distribution for each combination; and uses these distributions as a score.

DeepDock is a neural network responsible for two main tasks: feature extraction from the input data and identifying key ligand-target interactions (Figure 39). In a first step, the neural network extracts relevant representations of the input data, namely ligand and target structures, using the molecular surface of the binding site in the form of a polygon mesh. In this mesh, a collection of nodes, edges and faces defines the shape of the molecular surface as a polygon (Figure 39a). Moreover, the nodes also contain features encoding chemical and topological information at that specific point of the molecular surface, whereas edge features encode the connectivity between nodes. The protein targets were processed using a pipeline based on the one previously described by Gainza *et al.*[178]

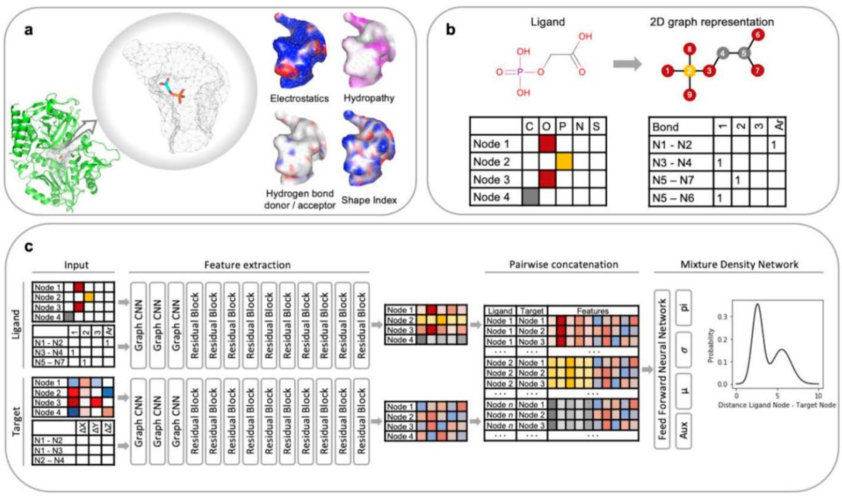



Figure 39. Deep learning model used to learn a potential to predict binding conformations.

Ligands are represented as a two-dimensional undirected graph, where atoms are designated by nodes and bonds are represented by edges (Figure 39b). In this case, node and edge features encode the atom and bond types, respectively. Both, the target mesh and the ligand graph, are processed by independent residual graph convolutional neural networks (GNNs). Through this procedure, the processed node features not only contain information of an individual atom or point in the molecular surface, but also have information about the other nodes around them. In other words, the processed atom features encode the whole atomic environment around a specific atom (Figure 40), whereas the target features encode a patch of the molecular surface around a specific point.

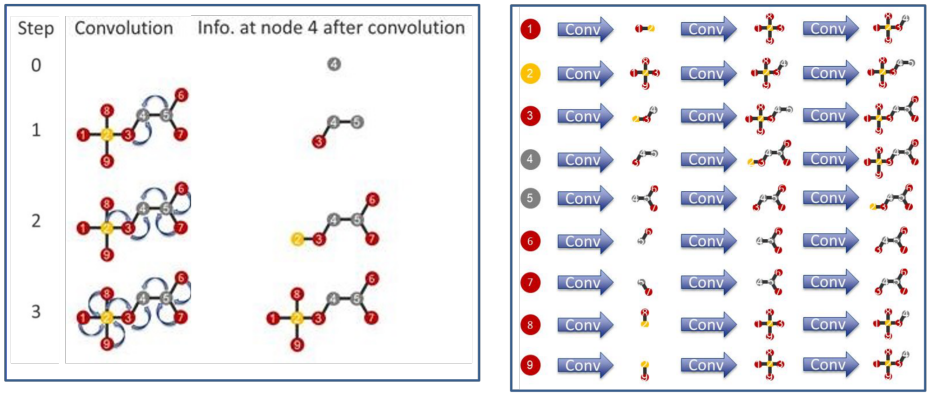



Figure 40. Graph convolution network.

In a following step, the processed node features from the target and ligand were combined in order to model the interaction of the ligand with the target (Figure 39c). All node features were concatenated in a pairwise manner, meaning each ligand atom was paired with each node in the molecular surface of the target. In a final step, these concatenated features were processed by mixture density network (MDN).

An advantage of this deep learning approach is that it can be easily combined with optimization algorithms in order to find the ligand conformation associated with the global minimum of the potential. In other words, it can find the ligand conformation with highest likelihood of binding. The optimization algorithm rotates each rotatable bond in the molecule, and at the same time it translates and rotates the whole ligand using a transformation matrix until it finds the conformation that best fits the binding pocket. The ligand conformation is represented as a vector of the Euler angles, the relative position of the ligand in the Euclidean space, and the dihedral angles of all rotatable bonds in the molecule (Figure 41). Differential evolution was employed to find the ligand conformation that minimizes the potential learnt by the model for that specific complex, although other methods of optimization could be used.

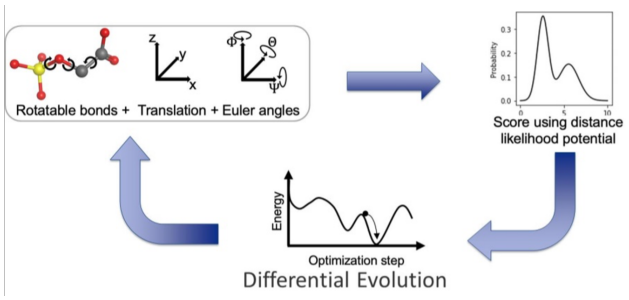



Figure 41. Predicting the binding conformation.

Figure 42 illustrates the use of distance likelihood potential to predict ligand binding conformations for 2-phosphoglycolic acid to rat PEPCK (PDB ID: 2RKA). The experimental binding conformation is in depicted in cyan lines and the polygon mesh in gray lines. Examples of predicted distance probability distributions between ligand atoms and target nodes for 2RKA are shown in Figures 42 d-f. The dashed line indicates the distance between ligand atoms and target node for the predicted binding conformation.

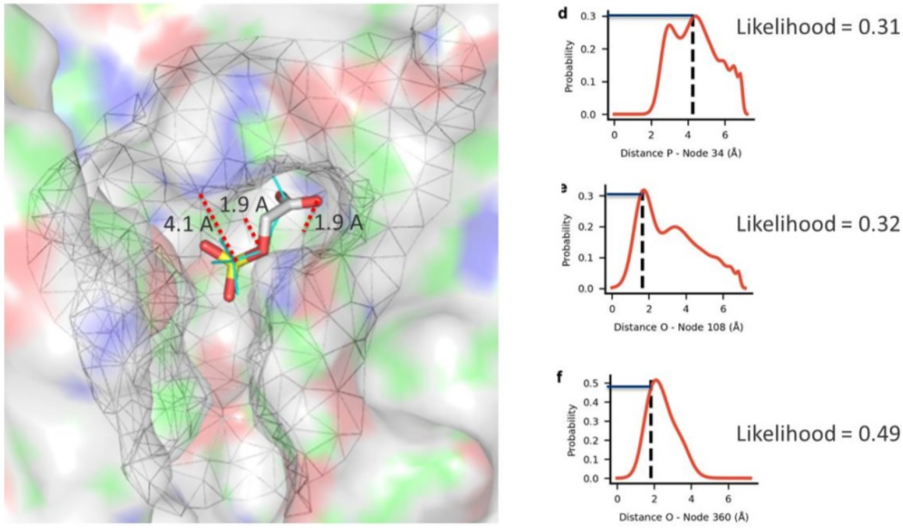



Figure 42. Use of distance likelihood potential to predict ligand binding conformations.

Méndez-Lucio and his co-workers used the CASF-2016 benchmark[179,180] in order to evaluate if their potential based on distance likelihood is suitable to be used as an accurate scoring function for an optimization algorithm. The CASF-2016 benchmark is composed of 57 protein targets each with a set of five true ligands (i.e., a total of 285 compounds), carefully selected to contain diverse proteins in terms of amino acid sequence and unique ligands, covering a wide range of binding affinity. It provides 100 precomputed (decoy) binding conformations for each of the 285 ligands in each of the protein targets (i.e., 28,500 conformations per target). The benchmark is designed to assess scoring functions in four demanding tasks, namely scoring, ranking, docking, and screening power. Since DeepDock is not specifically trained to predict binding affinities, only the docking and screening power tasks were relevant in this study.

The evaluation of docking power measures the ability of a scoring function to identify native ligand binding poses among a set of 100 decoys. The scoring function under evaluation is used to rank all decoys expecting those with similar conformations to the native ligand-binding pose (i.e. RMSD $< 2$ Å) to be among the top-ranked. The forward screening task involves find the correct binder for a target; the reverse screening task evaluates the reverse screening power of a scoring function, that is the ability of identify the real target of a molecule among a set of random targets. The comparative performance of DeepDock in the three tasks[180] is shown in Figure 43.

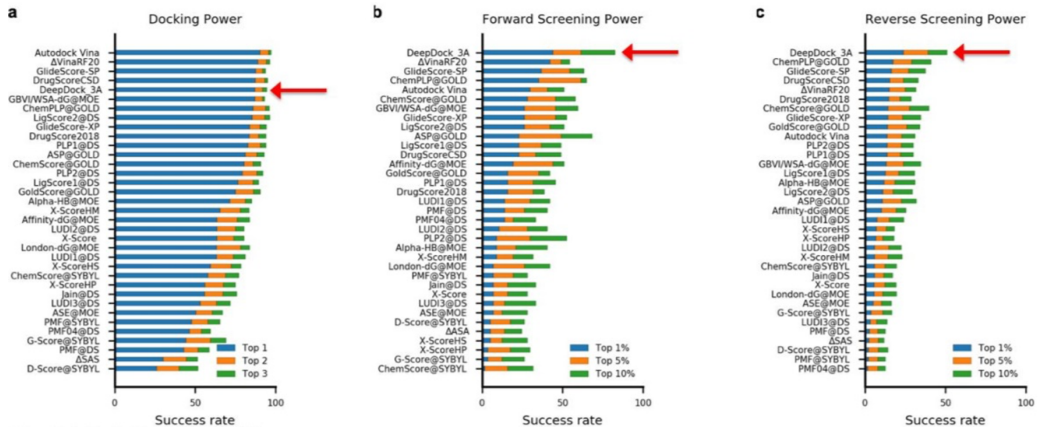



Figure 43. Model performance in CASF-2016 benchmark.

The number of compounds with successful optimization in CASF-2016 was 225 out of 285. The mean RMSD of all compounds was 1.87 ± 1.33. The mean RMSD of compounds with successful optimization was 1.42 ± 0.94. A scatter plot (Figure 44) shows that RMSD between predicted and experimental binding conformations is lower for compounds with fewer rotatable bonds. The optimization using differential evolution successfully finished for most compounds with fewer than 10 rotatable bonds.

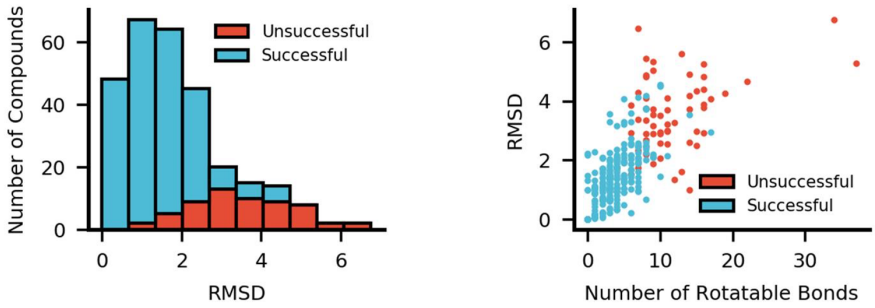



Figure 44. Optimization in CASP-16.

In another experiment, 1367 ligand-target pairs were used as the validation set to test the optimization. The number of compounds with successful optimization was 917 out of 1367, the mean RMSD of all compounds was 2.50 ± 1.85 and the mean RMSD of compounds with successful optimization was 1.62 ± 0.99.

Méndez-Lucio concludes that geometric deep learning can learn distance distributions that can be used as a potential for ligand-target interactions. This potential performs similarly to or better than well-established scoring functions. It can be coupled with global optimization algorithms to reproduce experimental binding conformations of ligands and is an example of how AI will help to significantly improve and speed up structure-based drug design.

# 17 Finding new *in silico*-based therapeutic strategies for IAHSP

**Matteo Rossi Sebastiano, Department of Molecular Biotechnology and Health Sciences, University of Turin, Italy**

The full video of Sebastiano's talk can be viewed here: https://youtu.be/ciR0bXvyIoE.

Infantile-onset ascending spastic paralysis (IAHSP) is a neurodegenerative, autosomal recessive, rare disease which affects fewer than 50 people worldwide. The pathogenesis starts in early childhood, with a progressive degeneration of the upper spinal motoneuron, progressively hindering deambulation until it spreads to the upper limbs and to the involuntary musculature.[181] Key events responsible for this condition are mutations to the gene ALS2, which encodes for the protein alsin related to cell trafficking. Failure of cell trafficking leads to defective neuron development and maintenance.

The relatively broad mutational landscape and the low number of reported cases make a complete understanding of the physiopathology and the search for suitable therapeutic strategies challenging. The majority of mutations described in literature result in a truncated form of alsin which is reputed to be degraded, thus depicting a scenario of loss-of-function pathogenesis. Nevertheless, some patients report missense mutation, leading to nondegraded, mutated forms. In those cases, the majority of amino-acid substitutions occur in the N-terminal RLD domain, essential for alsin localization to the plasma membrane and eventually to early and late endosomes upon activation of the RAC1 pathway. In endosomes, alsin binds to the small GTPase Rab5 and performs guanine nucleotide exchange factor (GEF) activity through its C-terminal VPS9 domain2 (Figure 45). In contrast with the majority of reports, Rossi Sebastiano presented a patient case harboring two alsin mutations in the C-terminal region: one allele translates a frame-shifted, truncated form which gets degraded; the other allele harbors the R1611W amino-acid substitution in the VPS9 domain (Figure 45 bottom right).

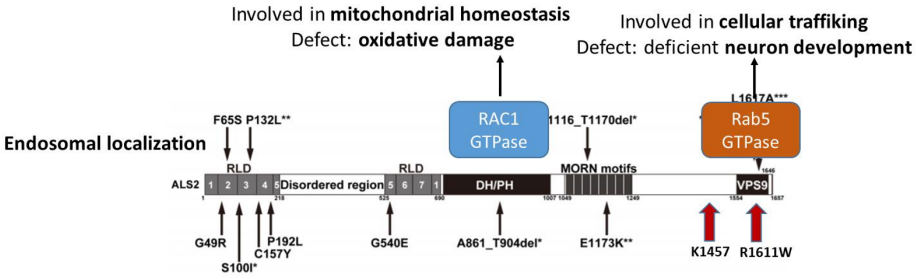



Figure 45. Alsin structure. functions and mutations.

This pathway is reputed to be the major strategy that mammalian cells follow, in order to assemble endosomes and exchange materials within the cell architecture. In dimensionally important cells such as motoneurons, coordinated and efficient cell trafficking is crucial for correct development and function maintenance. Alsin exists in cytoplasmic solution as a tetramer, firstly assembled by parallel dimerization through the VPS9 domain and subsequently by interaction of two dimers through their DH/PH domain, located upwards of the VPS9 region2 (Figure 46).[182] The first challenge that such a broad mutational landscape offers is that different mutations correspond to different multimers. These states do not just affect stability and solubility, but also subcellular localization and GEF activity. To make this situation more challenging, there is no experimentally resolved 3D structure of alsin, and a

homology modeling effort[183] to build the whole protein seems questionable because of the lack of a reliable template.

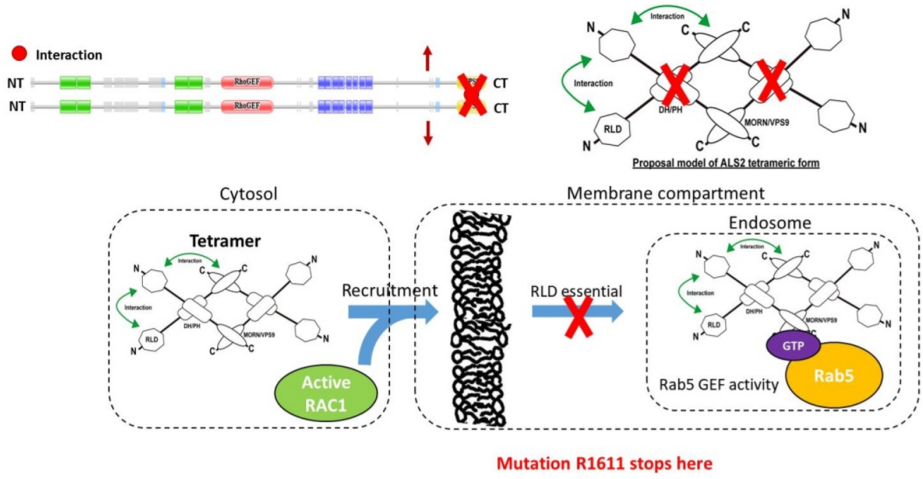



Figure 46. Localization upon RAC1 activation.

Rossi Sebastiano and his colleagues aimed to characterize key domains intervening in the pathogenesis of R1611W (VPS9), to characterize their structural dynamics, and to carry out drug repurposing in order to treat the patient. With the aid of *in silico* computational tools, Rossi Sebastiano and co-workers predicted the 3D structure of normal and mutated forms of the domain. The team carried out two-step homology modeling using MODELLER. First, they modeled the VPS9 core domain with wild-type (WT) and mutated (mut) sequences (Figure 47). Delprato *et al.*[184] had shown that the catalytic core of the Rab GEF Rabex-5 has a tandem architecture consisting of a Vps9 domain stabilized by an indispensable helical bundle (Figure 47C).

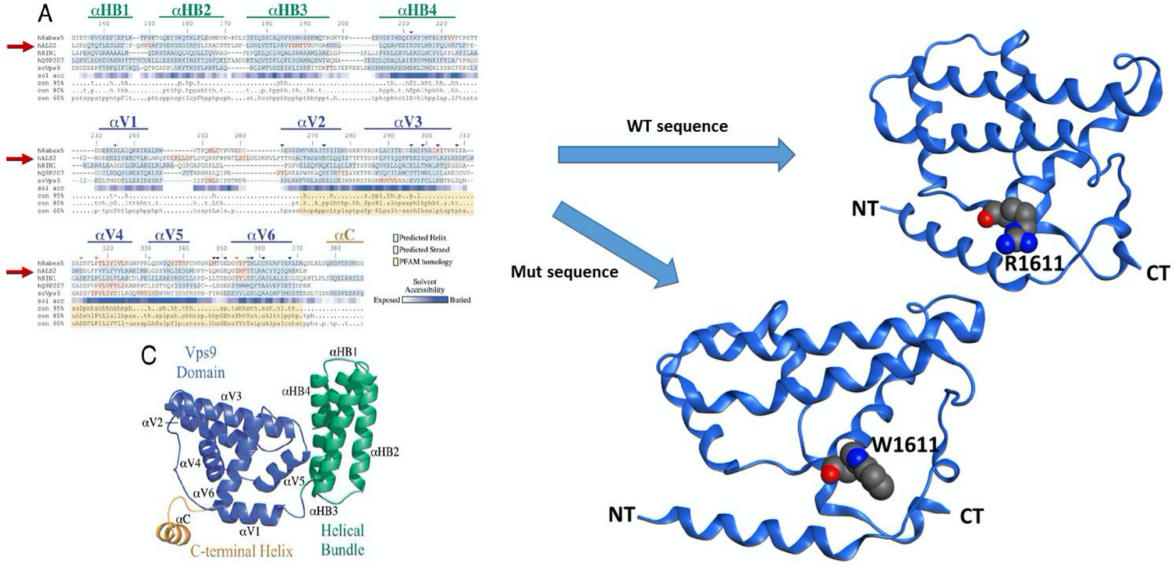



Figure 47. Homology modeling focused on the VPS9 domain.

Next Rossi Sebastiano *et al.* modeled the helical bundle (Figure 48) and showed no difference in flexibility (using CABS-flex 2.0).

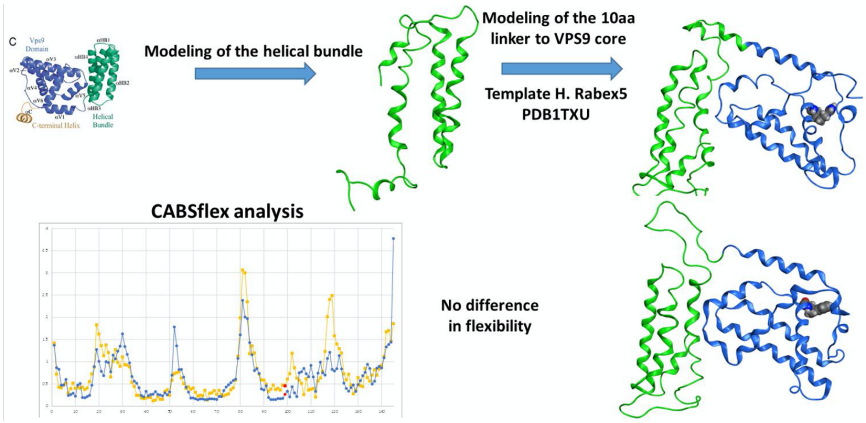



Figure 48. Homology modeling fstep 2.

They also characterized physiologic and pathologic dimerization modes. Using protein docking based on the ClusPro docking engine, and 30 dimer structures per condition, they showed that for both VPS9 and RLD domains, the mutated form is preferentially at the interface (Figures 49 and 50).

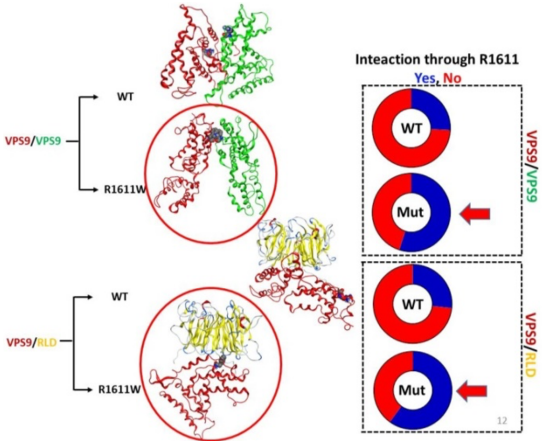



Figure 49. Dimerization capacity of W1611-VPS9.

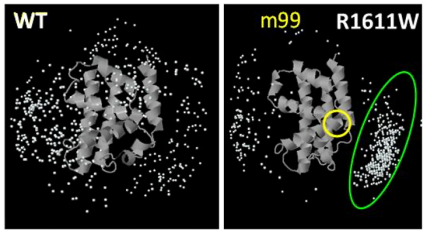



Figure 50. Dimerization and stability.

Binding analysis using ZDOCK showed that mutated VPS9 preferentially forms an antiparallel dimer by interacting with the RLD domain: VPS9-R1611W forms the most stable complex with RLD. In the mutated VPS9 domain, there is a cluster of interaction areas (green oval in Figure 50) corresponding to the mutated residue (in yellow). Substituting an arginine with a tryptophan might give rise to a hydrophobic surface and, indeed, after calculating the solvent-accessible surface area it can be seen that the area is reduced, indicating that the mutated sequence prefers to react with the RLD domain. The interaction energy calculated with the AMBER force field confirms the highest potential, as does the Caldararu model.[185] Rossi Sebastiano *et al.* linked this discovery to the experimentally determined loss of tetrameric aggregation and, more important, to the incorrect endosome localization. This finding corroborates and gives a mechanistic explanation for the experimentally characterized reduced Rab5 GEF endosomal activity. The team plans to use steered molecular dynamics in future (using NAMD2 and the CHARMM36 force field) in the hopes that the trajectory analysis will offer information about the stability of the interaction, but, thus far, some technical difficulties have been encountered.

Finally, Rossi Sebastiano's team has performed *in silico*, structure-based virtual screening. Having found a mutated model, they defined a region of interest constituted by nine residues within 5 Åof Trp1611 for a ligand search. They individuated the pharmacophore with PocketQuery, and used Cavity Prank, Proteins*Plus*, and Cavity CASTp to confirm the presence of a cavity. They then used the cavity centered on W1611 as the query to search 992 compounds from ZINC15 and 8823 from DrugBank. Docking was performed using GOLD with standard parameters and the final hits were screened for blood-brain barrier permeability. An already commercialized drug which is able to shield the pathologically acquired hydrophobic moiety is being repurposed.

Another issue that the team had to investigate was potential interference of the mutated residue with Rab5 activity of the protein. So, they modeled the interaction of the VPS9 domain with Rab5 using a structure (PDB ID 2EFC) of Rab5 bound to VPS9 of Rabex5 in A. *thaliana*. This has a high homology to the human protein. Rossi Sebastiano's team were able to obtain a series of interaction models by superimposition. Visual inspection of these showed that Arg1611 is on the other side of the protein from the interaction surface with Rab5, suggesting no interference. Moreover, a literature search revealed that six residues are essential for Rab5 activity of Rabex5 and none of them is in close contact with Arg1611 in the human protein or the residues interacting with the drug candidate.

In the team's hypothesis, the mechanism of action re-establishes physiological dimerization, tetramerization mode, subcellular localization, and Rab5 activity in R1611W-mutated patients. The drug candidate is currently under preclinical testing in an alsin R1611W cellular model.

## 18 Designing molecular models by machine learning and experimental data

**Cecilia Clementi, Einstein Professor of Physics, Freie Universität (FU) Berlin, Germany**

The full video of Clementi's talk can be viewed here: https://youtu.be/DWbJS4bQ3uQ.

A major challenge in biophysics is the broad range of interconnected length and time scales (Table 7). For a small molecule QM/MM can be used, and for larger biomolecules, all-atom MD, but for much larger scales, such as for a cell, tools such as reaction diffusion simulation might be used. In Clementi's group, they use coarse-grained (CG) models to try to bridge small and large scales in molecular biophysics.

Recently there has been an immense increase in high-throughput and high-resolution technologies for experimental observation as well as high-performance techniques to simulate molecular systems at a microscopic level, resulting in vast and ever-increasing amounts of high-dimensional data, but experiments provide only a partial view of macromolecular processes and are limited in their temporal and spatial resolution. On the other hand, atomistic simulations are still not able to sample the conformational space of large complexes, thus leaving significant gaps in our ability to study molecular processes at a biologically relevant scale. Clementi's group aims to bridge these gaps, by exploiting the available data and using state-of-the-art machine-learning methods to design optimal coarse models for complex macromolecular systems.

Table 7. Scales in BioPhysics

| Entity | Range | Method |
|---|---|---|
| Organism | $\sim 10^{20}$ atoms | Thermodynamics. Macroscale |
| Cell | $\sim 10^{10}$ atoms, $\sim$1-10$\mu$m | Thermodynamics. Mesoscale |
| System | $\sim 10^{4}$-$10^{5}$ atoms, $\sim$10-100nm | Mesoscale. Multiscale |
| Biomolecule (macromolecule) | $\sim 10^{3}$-$10^{4}$ atoms, $\sim$1-10nm | Mesoscale. Multiscale |
| Molecule | $\sim 10^{1}$ atoms, $\sim$1-10 Å | Quantum chemistry |
| Atom | $\sim$1 atom, $\sim$1 Å | Quantum physics |

As a metaphor for "coarse-grained", Clementi used the image of a cow and the amount of detail needed for recognition that it actually is a cow. At one end is a high resolution picture of a bull painted by Picasso, and at the other extreme is the simplest sketch of a mammal such as might be seen in a cave painting with a few brushstrokes, which is nevertheless recognizable as a cow. Similarly, in biophysics the level of detail depends on the goal and application. Behind a CG model is the mapping operator that changes the $N$ atoms in an atomistic representation to the $n$ beads in a CG representation (Figure 51).[186]

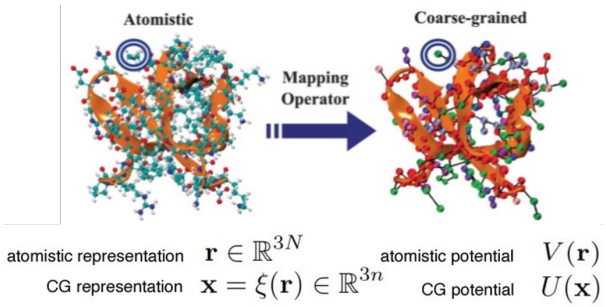



Figure 51. Coarse-grained mapping.

A coarse-grained model can be made "top-down" from real, experimental observations (often thermodynamic and macroscopic); or "bottom-up" from a classical, empirical, atomistic model of real material (a fundamental description); or by a knowledge-based approach from experimentally determined structures.[186] The choice of properties to be preserved determines the approach. If you want to preserve real, experimental observations you use a top-down approach; if you want to preserve the thermodynamics of an all-atom model, for example, you use a bottom-up approach Note that this talk does not cover kinetics, although the team has also worked on kinetics. Several approaches have been proposed to design effective CG energy functions for large molecular systems that either reproduce structural features of atomistic models (bottom-up) or reproduce macroscopic properties for one or a range of systems. In the latter class is the work Clementi's team did on topological and energetic factors in protein folding, studying the folding of simplified models of five small globular proteins.[187] Now her team has formulated the well-known force-matching procedure for coarse-graining as a supervised machine learning problem.

Noid *et al.*[186] established the theory behind coarse-graining with thermodynamic consistency (Figure 52). The goal is to optimize the parameters of a bottom-up CG model to satisfy this consistency as well as is possible. The right hand side of Figure 52 illustrates a two-dimensional potential used as a toy system; the 2D energy function has a 1D projection underneath.

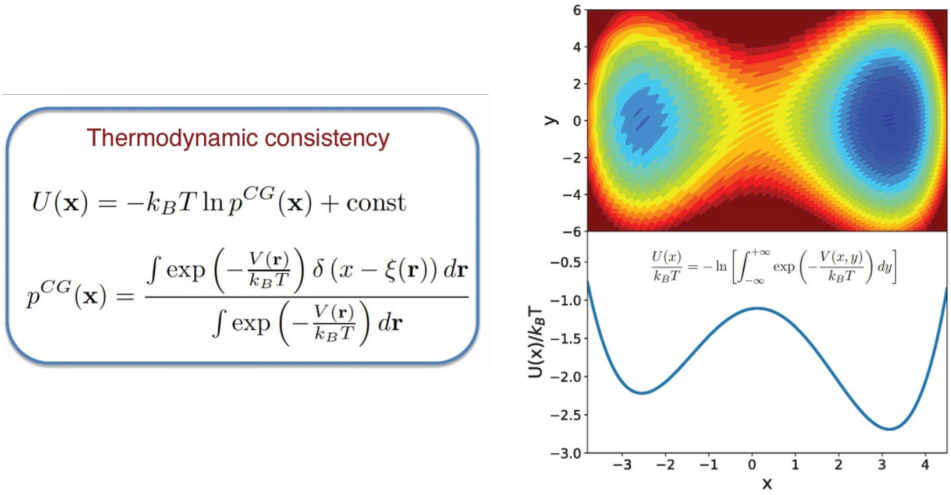



Figure 52. Coarse-graining with thermodynamic consistency.

Enforcing the thermodynamic consistency is equivalent to minimizing the force matching error. The force matching error is always greater than zero and can be decomposed into two parts: potential of mean force (PMF) error, and noise. The PMF error should be as close to zero as possible; the noise term is determined solely by the coarse-graining mapping. The gradient of the CG potential should be as close as possible to the "mean force". A loss function is defined to minimize the force matching error over a Boltzmann-distributed sample.[188]

Clementi's team have developed CGnets (Figure 53), a deep learning approach that learns coarse-grained free energy functions and can be trained by a force-matching scheme.[188] CGnets have many similarities to neural networks used to learn potential energy surfaces from quantum data, such as enforcing the relevant invariances (e.g., rotational and translational invariance of the predicted energy, equivariance of the predicted force). In contrast to potential energy networks, CGnets predict a free energy (PMF) and then use the gradient of this free energy with respect to the input coordinates to compute a mean force on the CG coordinates. As the CG free energy is not known initially, only the force information can be used to train the network.

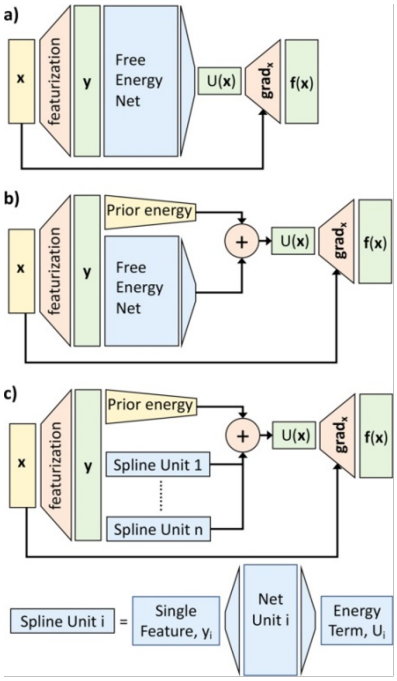



Figure 53. Neural network schemes. (a) CGnet. (b) Regularized CGnet with prior energy. (c) Spline model representing a standard CG approach, for comparison. Each energy term is a function of only one feature, and the features are defined as all the bonds, angles, dihedrals, and nonbonded pairs of atoms.

Clementi's team has demonstrated that CGnets succeed in learning the CG mean force and the CG free energy for the coarse-graining of all-atom, explicit-solvent simulations of the folding and unfolding of the polypeptide chignolin to a CG model consisting only of the protein $C_\alpha$ atoms and no solvent (Figure 54). CGnets maintain all physically relevant invariances and allow one to incorporate prior physics knowledge to avoid sampling of unphysical structures. They achieve a systematically better performance than classical CG approaches which construct the CG free energy as a sum of few-body terms. In the case of the chignolin protein, the classical few-body model cannot reproduce the folding and unfolding dynamics, but the inherently multibody CGnet energy function approximates the all-atom folding and unfolding landscape well and captures all free energy minima.

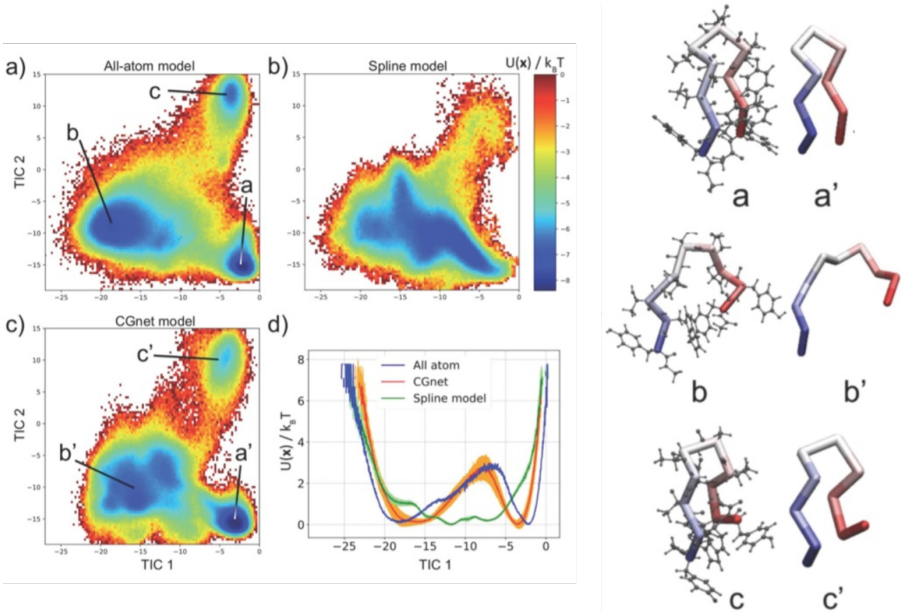



Figure 54. Free energy landscape of chignolin for the different models. (a) Free energy as obtained from all-atom simulation, as a function of the first two time-lagged independent component analysis (TICA) coordinates. (b) Free energy as obtained from the spline model, as a function of the same two coordinates used in the all-atom model. (c) Free energy as obtained from CGnet, as a function of the same two coordinates. (d) Comparison of the one-dimensional free energy as a function of the first TICA coordinate (corresponding to the folding/unfolding transition) for the three models: all-atom (blue), spline (green), and CGnet (red). (e) Representative Chignolin configurations in the three minima from (a–c) all-atom simulation and (a'–c') CGnet.

The framework, however, requires the manual input of molecular features to machine-learn the force field. So, the team has built upon it by introducing a hybrid architecture, CGSchNet. Schütt *et al.*[189] have reported a deep learning architecture (SchNet) that is designed to model atomistic systems by making use of continuous-filter convolutional layers. They demonstrated the capabilities of SchNet by predicting a range of properties for molecules and materials, where the model learns chemically plausible embeddings of atom types across the periodic table. They have also used SchNet to predict potential-energy surfaces and energy-conserving force fields for MD simulations of molecules. Clementi's team use the SchNet method in CGSchNet,[190] where the coarse-grained force fields learn their own features *via* a subnetwork that leverages continuous filter convolutions on a graph neural network architecture (Figure 55).

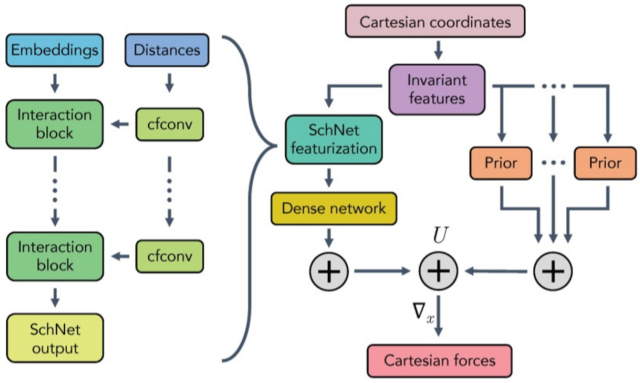



Figure 55. SchNet and CGSchNet.

Besides being transferable, CGSchNet works better than CGnet in terms of reproducing the free energy. Clementi showed two different ways of measuring the error (Figure 56) and a free energy landscape with three minima and the representative ensembles of structures (Figure 57). The three minima are also present in the atomistic method. Clementi also showed a movie of the trajectory in chignolin folding and unfolding.

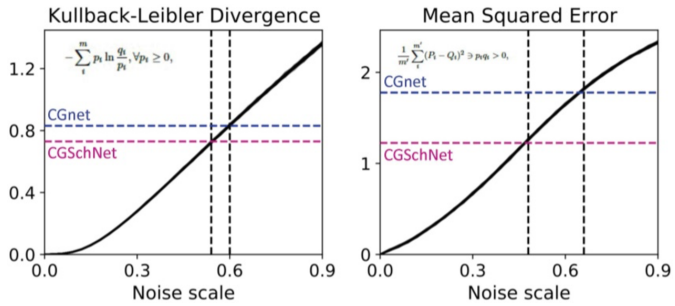



Figure 56. CGSchNet free energy results.

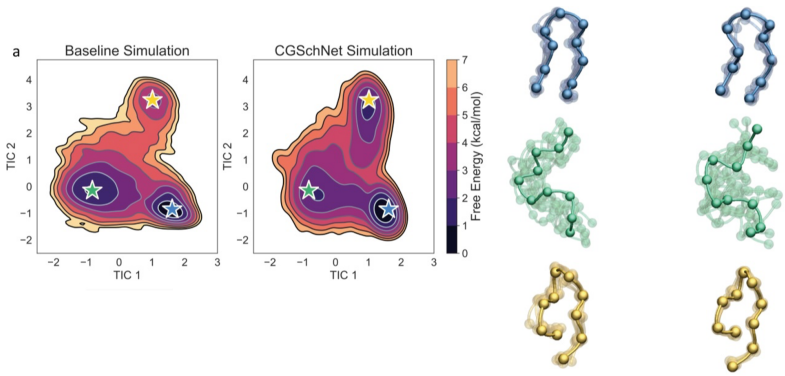



Figure 57. CGSchNet results.

The success of these models relies on the fact that they can reproduce multibody interactions. (Note that in the spline model (with only two bodies) in Figure 54 there is just one blob, not three minima.) It has been shown that the inclusion of multibody terms improves the accuracy of a CG model but no general approach has been proposed to construct systematically a CG effective energy that includes arbitrary orders of multibody terms. Clementi's team adopted a neural network based approach to address this issue and constructed a CG model as a multibody expansion (Figure 58).

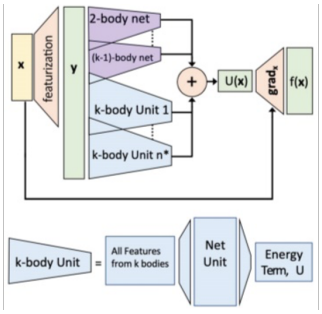



Figure 58. Multibody potentials.

By applying this approach to a small protein, they evaluated the relative importance of the different multibody terms in the definition of an accurate model. Using the same ways of measuring error as were used in Figure 56, Clementi presented the results shown in Figure 59 (where in 2,3,4C, the "C" indicates chiral). She also showed the corresponding free energy landscapes (Figure 60). The team observed a slow convergence in the multibody expansion,

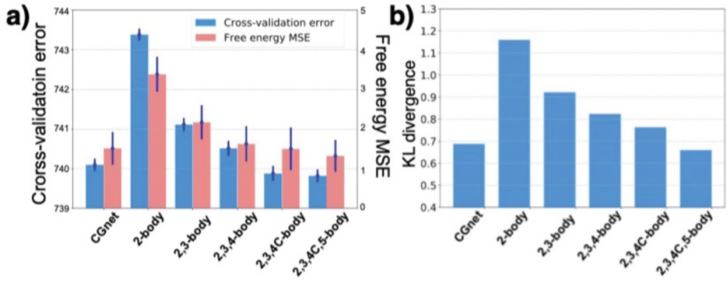



Figure 59. Number of bodies.

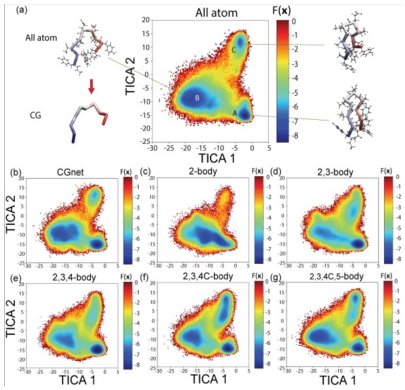



Figure 60. Free energy landscapes.

where up to five-body interactions are needed to reproduce the free energy of an atomistic model.[191]

Several outstanding challenges still need to be tackled to build upon these results and design a transferable, general CG model for proteins. Clementi's group is currently addressing these challenges on multiple fronts in their ongoing research.

# 19 The "almost druggable" genome

**Tudor Oprea, Professor, and Chief, Translational Informatics Division, Department of Internal Medicine, University of New Mexico School of Medicine, Albuquerque, NM, USA**

The full video of Oprea's talk can be viewed here: https://youtu.be/PezsuWDpvuI.

Informatics, data science, and machine learning can be used with the three pillars of drug discovery: diseases, targets, and drugs. Oprea presented some specific contributions in each of these three areas, starting with targets.

Drug targets are often poorly defined in the literature. Oprea's team have contributed to a comprehensive map of molecular targets of approved drugs. They curated 667 human-genome-derived proteins and 226 pathogen-derived biomolecules through which 1578 drugs approved by the U.S. Food and Drug Administration (FDA) act.[192] The data are captured in DrugCentral. Three updates (for the years 2018, 2019, and 2020) have been summarized in *Nature Reviews Drug Discovery* Biobusiness Briefs.

A large proportion of biomedical research and the development of therapeutics is focused on a small fraction of the human genome. In a strategic effort to map the knowledge gaps around proteins encoded by the human genome and to promote the exploration of currently understudied, but potentially "druggable", proteins, the U.S. National Institutes of Health launched the Illuminating the Druggable Genome (IDG) initiative in 2014.[193] The IDG consortium encompasses five different groups. On the experimental side, there are three Data and Resource Generating Centers (DRGC) and 57 R03-funded grants; a Knowledge Management Center (KMC), and a Resource Dissemination and Outreach Center (RDOC). On the data science side, there are three cutting edge informatics tools (CEIT).

IDG KMC quantifies data availability from a wide range of chemical, biological, and clinical resources, and has developed platforms that can be used to navigate understudied proteins (the "dark genome"), and their potential contribution to specific pathologies.[194] The Target Central Resource Database (TCRD)[195] is the central resource behind the IDG KMC. The multiple data sources (about 70 of them) in the integrated knowledge base are available through Pharos.[195] Since 2017, Pharos has attracted over 100,000 users and more than 500,000 page views.

Oprea and his colleagues classify proteins by the Target Development Level (TDL), a ranking system based on various target milestones. Targets annotated as drug targets are known as "Tclin"; proteins for which *potent* small molecules are known are "Tchem"; proteins for which biology is better understood are "Tbio"; and proteins that lack antibodies, publications, or National Center for Biotechnology Information (NCBI) Gene References Into Function (GeneRIFs) are "Tdark". In 2021, 29.4% of targets are Tdark; 58.3% are Tbio; 9% are Tchem; and 3.3% are Tclin. On average, it takes 15-20 years for a Tdark protein to become druggable. Nuclear hormone receptors, signaling proteins, and kinases were the first to be identified. Transporters, transcription factors, and olfactory GPCRs are the most recent. There is a knowledge deficit: about 30% of the proteins remain understudied (although that percentage is steadily decreasing) and about 3% of the proteome is currently targeted by drugs (although that percentage is slowly increasing).

The DISEASES database supports biological research by preserving freely-accessible, richly and accurately text-mined data of biomedical entities.[196] The system consists of a highly efficient dictionary-based tagger for named entity recognition of human genes and diseases, which is combined with a scoring scheme that takes into account co-occurrences both within and between sentences. The content of the text mining channel is boosted by full-text articles openly accessible in PubMed Central. The text mining approach is also combined with other types of evidence such as manually curated disease-gene associations, cancer mutation data, and genome-wide association studies (GWAS) from existing databases. Target Illumination GWAS Statistics (TIGA) rank-scores GWAS data. In DISEASES 2.0, the evaluation of full-text articles using the new version of the dictionary of entities, based on BioBERT,[197] achieved 91:1% AUROC, compared to 84:5% AUROC for the 2013 collection of PubMed abstracts using the old version of the dictionary.

Most rare diseases still lack approved treatments despite major advances in research providing the tools to understand their molecular basis, as well as legislation providing regulatory and economic incentives to catalyze the development of specific therapies. Addressing this translational gap is a multifaceted challenge, for which a key aspect is the selection of the optimal therapeutic modality for translating advances in rare disease knowledge into potential medicines (orphan drugs). Oprea's team have reviewed the technological basis and rare disease applicability of the main therapeutic modalities, and have also discussed selected overarching topics in the development of therapies for rare diseases, such as approval statistics, engagement of patients in the process, regulatory pathways and digital tools.[198]

A lack of robust knowledge of the number of rare diseases and the number of people affected by them limits the development of approaches to ameliorate the substantial cumulative burden of rare diseases. Oprea teamed up with the Monarch team (Melissa Haendel, Chris Mungall, and Peter Robinson) and revised the number of rare diseases from about 7000 to 10,393, using Disease Ontology, OrphaNet, the Genetic and Rare Diseases Information Center (GARD), the National Cancer Institute Thesaurus (NCIt), Online Mendelian Inheritance in Man (OMIM), and the Monarch Initiative Mondo Disease Ontology.[199]

Contrary to claims in the literature, the majority of human proteins lack verifiable associations with human diseases. Only 5583 protein-coding genes (27.35%) are likely to have relevant associations. There is an urgent need to improve our current repository of gene-phenotype associations. Oprea's team is currently tackling this issue.

Knocking out the activity of a gene provides valuable clues about what that gene normally does. Humans share many genes with mice. Consequently, observing the characteristics of knockout mice gives researchers information that can be used to understand better how a similar gene may cause or contribute to disease in humans. Oprea's team found that of phenotyped knockout mice genes mapped to 7039 human genes, 1011 have associated drugs or chemical probes (Tclin/Tchem), 4694 have biological roles that have been studied (Tbio), and 1334 are in the "ignorome" (Tdark). Moreover, 3956 of these genes (of which 2468 are Tbio and 1286 are Tdark) are understudied (i.e., have a PubTator Central score $\leqslant 50$).

Oprea presented some Tclin, Tchem, Tbio and Tdark distributions for 177 diseases related to gain-of-function (GoF) mutations in proteins (according to the International Mouse Phenotyping Consortium) and 949 rare diseases related to loss-of-function (LoF) mutations (Figures 61 and 62 respectively).

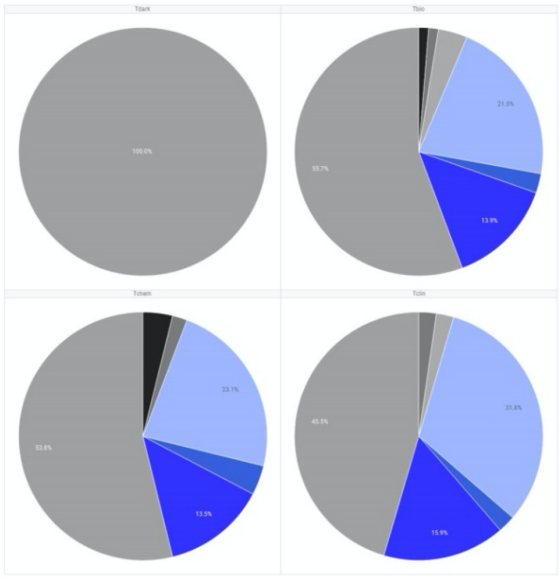


Figure 61. GoF diseases.

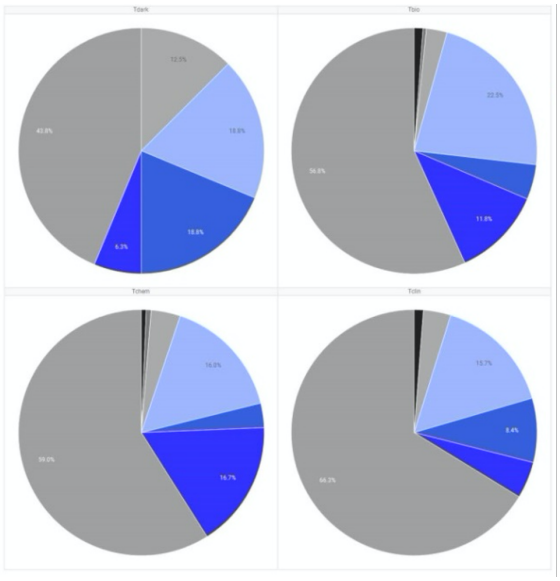
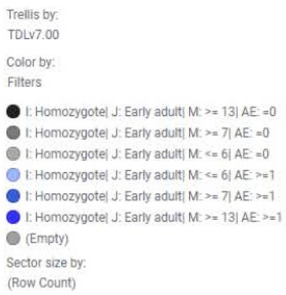


Figure 62. LoF diseases.

Tdark top left; Tbio top right

Tchem bottom left; Tclin bottom right




Monogenic rare diseases are likely to improve our understanding of therapeutic effects. Many drugs are antagonists or inhibitors, yet we do not associate their therapeutic success with "blocking GoF". When combined with phenotyped mouse models, GoF and LoF disease models can increase our understanding of pathology. This is a promising rational strategy for therapeutics.

Oprea described a machine learning workflow. A metapath (Figure 63) is a path consisting of a sequence of relations defined between different object types (i.e., structural paths at the meta level). Instances of metapath(s) are used to determine the strength of the evidence linking a gene to a disease or phenotype or function. This approach enables the transformation of assertions and evidence chains of heterogeneous biological data types into a format ready for machine learning.

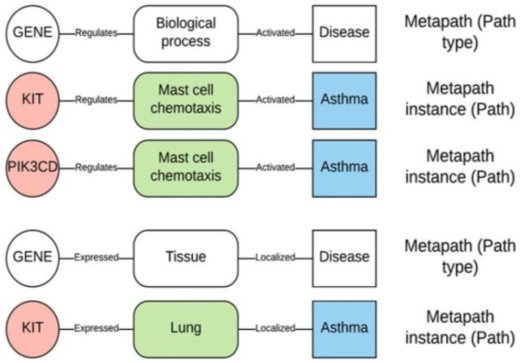



Figure 63. The metapath approach.

Oprea's team used a metapath-XGBoost workflow. Datasets are merged, *via* R scripts, into a PostgreSQL database (Figure 64). The database represents the knowledge graph of entities and relationships which connect as evidence paths. Graph embedding transforms evidence paths into vectors, converting data into matrices (Figure 65). Datasets are tables ready for machine learning, one row per gene, one column per embedded metapath instance feature. Genes associated with any OMIM phenotypic series (PS) are the training subset, divided into associated or not associated with this PS. Genes unassociated with any PS are the test subset.

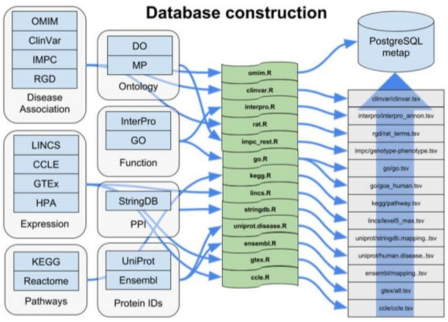



Figure 64. Database construction.

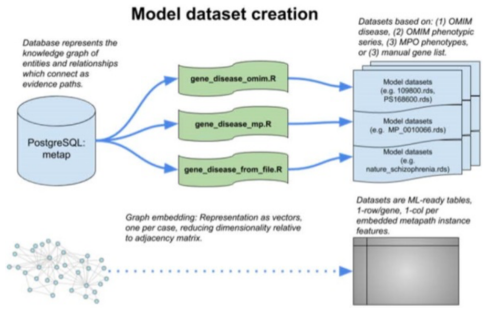



Figure 65. Model dataset creation.

XGBoost runs 100 models. The "median model" (by AUC and F1) is then selected for analysis and prediction to avoid overfitting (Figure 66).

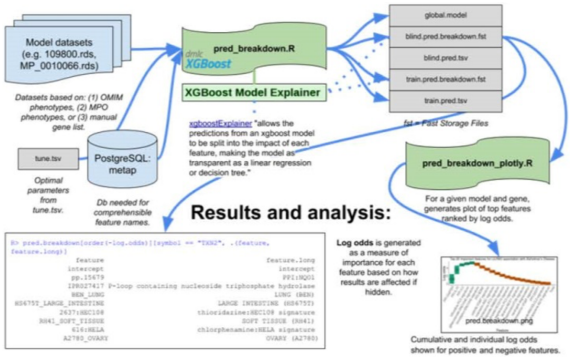



Figure 66. XGBoost model.

In the first model (Figure 67, left)) 51 OMIM genes associated with type 2 diabetes mellitus (T2D) *versus* 3954 OMIM genes associated with other pathologies, AUC was 0.72 ±0.08. VIP-ranked variables included HFE and HMOX1, which relate to hemochromatosis (which leads to T2D in 80% of cases), and IL1B and IL10 which suggests an immune component. Of the top 300 predicted proteins, about 90% were expressed in the pancreas and about 70% were expressed in beta-Langerhans cells. Later, Mark McCarthy suggested that the OMIM training set might not be a reliable source of information, so Oprea's team tried another training set.

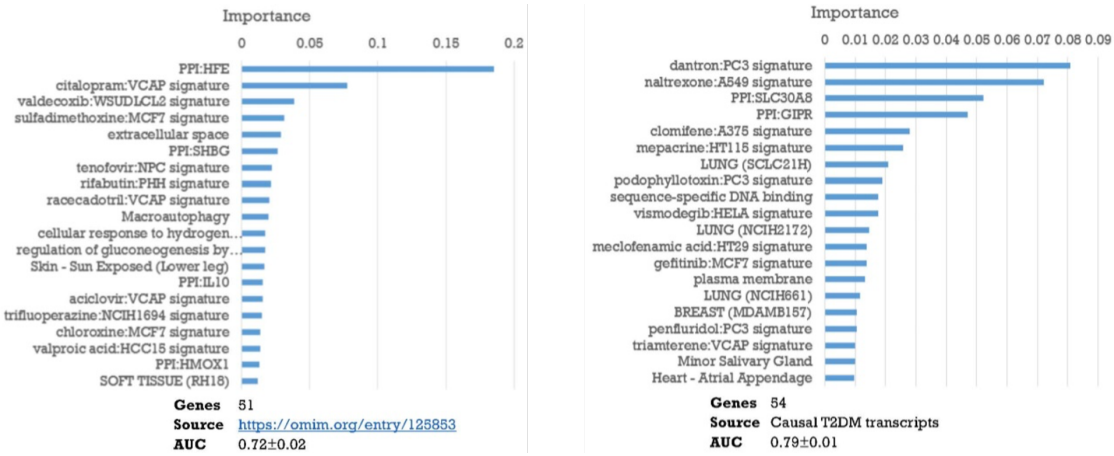



Figure 67. T2D classifier, variable importance plots.

In the second model (Figure 67, right), 54 causal transcripts provided by Anuba Mahajan and Mark McCarthy *versus* the 3954 OMIM genes, AUC was 0.79 ±0.01. Genes confirmed by GWAS (nine in the top 24) were C2CD4B, C2CD4A, JAZF1, ADAMTS9, CRY2, LINGO2, THADA, TMEM18 and SEC16B. Four genes had GO terms for insulin secretion (CPLX1, ADRA2A, SYT7 and SYTL4). The top four VIP-ranked variables included two protein-protein interaction nodes (SLC30A8(rs13266634) and GIPR(rs8108269)), which have GWAS-T2D associations.

Oprea concludes that machine learning model builders need to talk to experts: computers could not distinguish between the two models. His examples show that dark genome research can benefit both rare and common disease, and both patients and mice can help.

His final theme was drugs. DrugCentral[200] includes links to adverse events (AEs) from the FDA Adverse Event Reporting System (FAERS) mapped to the Medical Dictionary for Regulatory Activities (MedDRA) terminology: 2220 drugs are associated with 12,098 MedDRA terms, leading to 739,990 drug-AE combinations, while 1618 drugs associated with 8185 unique AEs lead to a total of 147,191 (20%) significant drug-AE combinations. DrugCentral also now incorporates REDIAL-2020, a machine learning platform that estimates anti-SARS-CoV-2 activities.[201]

Critical Assessment of Computational Hit Finding Experiments (CACHE)[109] is a public-private partnership benchmarking initiative to enable the development of computational methods for hit-finding. It is being conceptualized by scientists at Bayer, Google, the Structural Genomics Consortium (SGC) and the EBI, with active input from academics and scientists in dozens of companies and funders, and with advice from John Moult (the founder of CASP: see earlier in this report). The following questions were posed by Aled Edwards and the SGC team, for the CACHE program. Identify a list of human proteins that have:

- structure, no ligand
- structure with bound ligand (important ligand subcategories: druglike, fragment, peptide)
- structure, ligands and SAR
- no structure, close structural homologue with or without ligand.

In answer to the first two questions, Matthieu Schapira in Oprea's team found 1439 UniProt IDs captured by a PDB analysis (the "PDB-ome"). Of these, 583 have ligands bound and more than 75% desolvated upon binding to avoid shallow pockets; 384 have no ligand bound; and 622 have an endogenous ligand. There are 150 entries that have both endogenous and "desolvated" ligands. These data suggest that as many as 756 proteins are immediately "ligandable". Of these, 347 do not have ligands that meet the Tchem and Tclin criteria.

In answer to Edwards' third question Oprea found that 2644 UniProt IDs are associated with chemical matter (the "SAR-ome"), as captured by four resources: 2135 from ChEMBL, 272 from Guide to Pharmacology, 67 from DrugCentral, and 170 from World of Biomolecular Activity[202] (WOMBAT). Each of these proteins is associated with at least one ligand with measured bioactivity. Maximum bioactivity (-$\log_{10}$molar) was used to evaluate TDL. Five categories were identified (regardless of maximum bioactivity):

- significant SAR: 25 or more compounds
- good SAR: 11-24 compounds
- some SAR: 6-10 compounds
- limited SAR: 2-5 compounds
- no SAR: 1 compound.

Based on these data, 115 proteins might meet the ligandable criteria. Data on the intersection of the SAR-ome and PDB-ome suggest that as many as 759 proteins lack SAR. Of these, 721 have one or no ligand and a 3D structure, and are not Tchem or Tclin.

In an answer to Edwards' fourth question, Cristian Bologa has captured 4098 UniProt IDs by homology modeling (the "Pocket-ome"), from an analysis of UniProt and PDB using the Many-against-Many sequence (MMseqs) algorithm (after excluding the SAR-ome). Of these, 363 are ligandable Tbio or Tdark proteins that have a PDB-ligand association, in addition to 337 which are associated with a PDB structure *via* MMseqs analyses. There are 180 "ligandable-NoPDB" Tbio or Tdark proteins that do not have a PDB association but are associated with a Tchem or Tclin structure *via* MMseqs analyses. There are 1295 "3D-Model Likely" proteins that lack a PDB entry but could yield homology models given primary sequence identity over 60% within PDB. There are 1328 "3D-Model Maybe" proteins that lack a PDB entry but could be subjected to homology modeling given high primary sequence identity (30% -59.9%) with a protein in PDB. Thus, 700 proteins (363 + 337) are ligandable, another 180 may be ligandable *via* assays, though lacking immediate PDB associations, and 2623 proteins (1295 + 1328) could also be ligandable *via* homology models (depending on the cut-offs for primary sequence identity).

For the "phen-ome", Oprea examined all 6742 UniProt entries (4098 from "pocket-ome" and 2644 from "SAR-ome") with respect to phenotype information:

- 3294 in GO with experimental molecular-function or biological-process leaf term annotations
- 2010 with at least one IMPC knockout mouse model phenotype finding

- 3498 with GWAS traits (highest-ranked, one listed per gene; extracted from TIGA)
- 343 in the COSMICv92 cancer annotation system
- 2179 with rare disease annotations
- 148 with common disease annotations.

In total, 5425 proteins have at least one category. Of these phenotype-associated proteins that are not Tclin/Tchem, 606 proteins are ligandable with PDB; 144 may be ligandable *via* assays; and up to 1696 proteins may be ligandable as well, by homology modeling.

Thus, Oprea's answers to the four CACHE questions are that the "almost druggable" genome (Tbio and Tdark only) includes:

- 715 ligandable proteins
- 180 ligandable-noPDB proteins
- 100 Tbio or Tdark proteins with "no-SAR"
- 1295 "likely" and 1328 "maybe" proteins in the "homology model possible" category.

These constitute "low hanging fruit" that should be subjected to further biochemical screens. No doubt, the availability of AlphaFold2 predicted 3D models is likely to increase the number of "almost druggable" proteins.

To paraphrase William Gibson, "the truth is already here; it is just unevenly distributed".[194] High quality data are really hard to obtain and the weakest component is the ground truth. Specific domain expertise and feedback are needed. Informatics, data science, and machine learning are likely to accelerate drug discovery if high quality data are available.

## 20 Protein structure prediction: a drug discovery perspective

**Derek Lowe, Director, Chemical Biology and Therapeutics, Novartis Institutes for Biomedical Research, Cambridge, MA, USA**

The full video of Lowe's talk can be viewed here: https://youtu.be/YJg0WwUANEQ.

The advent of gradually more effective AI/ML techniques is already having effects on the traditional practices of medicinal chemistry and drug discovery in general. What can we expect as the process goes on and how will drug discovery scientists have to adjust their thinking and their research roles?

A medicinal chemist's immediate problems are which compounds to make next, how to make them, and whether the compound that was made was the one intended to be made. The first question is the traditional one during a project: interpreting SAR and deciding where to go. Of course, that decision might be driven by potency, by selectivity, by pharmacokinetics (PK), or many other factors. The second question is one of organic synthesis, and becomes an issue in scaffold-hopping, introducing a demanding new substituent, seeking a chiral synthetic route and so on, not to mention the implications for process chemistry. And the third question is the interface with analytical methods. Usually we are fairly sure of our footing, but mistakes here can be very costly indeed, and also embarrassing.

A medicinal chemist's larger problems are how potent and selective compounds need to be for the current project and which counterscreens (and toxicity data) have to be considered. The

first question may be difficult or near to impossible to answer, depending on how much is known about related proteins, about the "tone" of the system being targeted, and how well the project assays might translate to the clinic. The latter question is probably the single biggest success factor in a project, and one of the hardest to address. It gets into the "known unknowns" and "unknown unknowns" of the project. The toxicity issues that end many projects were not things that were the subjects of the counterscreens carried out.

The biggest problem for medicinal chemists is which project they should be working on. Should the time and effort be spent somewhere else, and if so, where might that be, and how sure can the chemist be about that decision: this therapeutic area, or another one entirely? This is indeed the really big question, touching on clinical failure rates (80-90%), on portfolio management, and on target identification and selection in general.

Where does protein structure fit in? It fits into the challenge of which compounds to make next. This is the classic domain of computationally driven approaches: modeling against a protein structure to suggest new SAR compounds although the structural data need to be of very high quality to get this to work. The energies involved are quite small. Lowe is cautious about what computation can currently offer but is optimistic about the future. Predictions of solution-phase behavior and of potential conformational changes are particularly valuable. An example is found in past HIV protease projects where it was found that there was a huge flap of the protein that could open or close. The effect was balanced on a thermodynamic knife edge that was not apparent from the crystal structure. It would also be good to know about conformational changes due to allosteric sites and protein-protein interactions, and the impact on compound binding, because induced fit is difficult to model. Protein structure has no bearing on the other two immediate problems of the medicinal chemist.

The usefulness of protein structure in terms of the medicinal chemist's larger problem concerning potency and selectivity is hard to assess. There may not be enough known about the system, especially in human disease, to make a useful prediction. Counterscreens and toxicity are a notorious jungle of unknowns which will likely be the "last frontier" for quite some time to come. In short, these larger issues get away from the importance of protein structure, and start to deal with larger cellular systems for which any computational approach will be struggling to produce an answer.

As for the medicinal chemist's biggest problem, that of which project to work on, Lowe believes that, despite some brave talk, this, the biggest question in drug discovery, is currently beyond help by computational means. We need to know much more about the biology of health and disease.

Looking back, Lowe does not believe that he has ever worked on a project where knowledge of the protein structure was a rate-limiting step, but, that said, he has been doing "traditional" drug discovery by traditional means for 30 years. If we can reach a new era where we have intimate knowledge of protein structure, especially for protein-protein interactions and complexes, complete with changes on compound binding, allosteric sites, and so on, everything could change.

Lowe does appreciate advances in protein structure prediction because getting structures is not so easy experimentally. There are many proteins and they are dynamic. Cryo-EM and X-ray methods do not establish solution behavior or explain which parts of the protein are more flexible and how they flex. This is where computational structure prediction could really shine and address real world protein behaviors that we cannot address by current analytical

experimental methods. If computational chemists could progress to that point we would be poised to open up a new world of knowledge, and we could really start to work out what happens in protein systems and perhaps see the tone of these things without having to do vast numbers of experiments. An example is the factors influencing selectivity. We could start to piece together a meaningful model of cellular activity. This summarizes Lowe's optimistic view, but a lot more work is needed. Lowe does not see that protein folding has won the race: it has merely won the race to get to the starting line.

## 21 Open access data: a cornerstone for artificial intelligence approaches to protein structure prediction

**Stephen K. Burley, University Professor and Henry Rutgers Chair, Founding Director of the Institute for Quantitative Biomedicine, and Director of the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank at Rutgers, the State University of New Jersey, New Brunswick, NJ, USA**

The full video of Burley's talk can be viewed here: https://youtu.be/73vXoH1vHG4.

The PDB was the first open access digital data resource in all of biology. It is the single global archive for protein, DNA, and RNA experimental structures, and their complexes with one another and small-molecule ligands. It was established in 1971 (with just seven X-ray structures of proteins) to archive 3D structures of biological macromolecules as a public good. As of October 2021, it provides open access to more than 180,000 structures. In recognition of its global importance, the PDB has been managed since 2003 by the Worldwide Protein Data Bank (wwPDB)[203] partnership: the RSCB PDB in the United States, PDBe in Europe, PDBJ in Japan, the Electron Microscopy Data Bank (EMDB), and the Biological Magnetic Resonance Data Bank (BMRB).

These organizations support open access to PDB, EMDB, and BMRB archives; weekly release of new 3D structures and data; expert biocuration, validation, and remediation of all the data; and PDBx and macromolecular Crystallographic Information File (mmCIF) data standards and dictionary for the PDB archive. They also support regional sharing of 3D structure and data deposition tasks (using the wwPDB deposition system, OneDep). RCSB PDB acts for the Americas and Oceania, PDB Japan for Asia and Middle East, and PDBe for Europe and Africa. Partner-hosted websites offer complementary services and views of identical PDB data. Within the wwPDB, the RCSB PDB serves as the global archive keeper for the PDB core archive.

The RCSB PDB converts global data into global knowledge.[204] Operations are supported by four interlocking services, including data deposition and biocuration (carried out in concert with the wwPDB partners); archive management and access; data exploration *via* the RCSB web portal; and outreach and education *via* the PDB101 website. PDB data are also repackaged and redistributed by more than 400 external data resources. It has been calculated that the replacement cost of PDB data would exceed 18 billion U.S. dollars.

PDB data users are working, teaching, and learning in fundamental biology, biomedicine, bioengineering, biotechnology, and energy sciences. They also represent the fields of agriculture, chemistry, physics, materials science, mathematics, statistics, computer science, and zoology, and even the social sciences. The enormous wealth of 3D structure data stored in the PDB has underpinned significant advances in our understanding of protein architecture, culminating in recent breakthroughs in protein structure prediction accelerated by artificial

intelligence approaches and machine learning methods.

As an example of the impact of PDB on fundamental biology, Burley cited the crystal structure of the nucleosome core particle (PDB ID 1AOI) first reported in 1997.[205] This is the most highly cited PDB structure (cited more than 10,000 times according to Google Scholar). Today more than 300 PDB structures explain in 3D how the DNA packaging process works and is regulated during gene expression.

The PDB has also had an impact on research in energy science. The first structure of photosystem II (PDB ID 1S5I) was deposited in 2004.[206] This is another highly cited PDB structure (cited more than 3500 times according to Google Scholar). Today more than 400 PDB structures explain in 3D how photosynthesis works and are guiding research into how it can be harnessed by energy researchers.

Discovery and development of 184 (nearly 90%) of the 210 new drugs approved by the U.S. FDA between 2010 and 2016 were facilitated by open access to 5914 PDB structures.[207] Moreover, $100 billion dollars of National Institutes of Health (NIH) funding attracted more than $600 billion dollars in biopharmaceutical company investment to bring the 210 new drugs to the market.[208] These results underscore the importance of public funding of basic research and open access underpinning the global drug discovery and development ecosystem. In Burley's opinion, this NIH funding, resulting in drugs that improve the lives of hundreds of thousands if not millions of patients worldwide, represents an excellent allocation of U.S. taxpayer monies.

The PDB has also had an impact on SARS-CoV-2 research. An unexplained illness was identified in Wuhan, China in late 2019, and the causative agent, the novel SARS-CoV-2 coronavirus, was identified in January 2020. In the same month, the first SARS-CoV-2 protein structure was deposited to PDB (Nsp5; PDB ID 6LU7). As of October 2021, there were more than 1500 3D structures in PDB related to SARS-CoV-2.

Burley also discussed the relationship between PDB and protein structure prediction. Major milestones include homology modeling, template-free methods, fragment assembly approaches, use of evolutionary information, machine learning and deep learning methods, RoseTTAFold, and Google DeepMind's AlphaFold. Enabling infrastructure encompassed open access to PDB data,[209] deep genomic sequence data, CASP (coordinated with wwPDB), the Continuous Automated Model Evaluation (CAMEO) platform (coordinated with wwPDB), peer-reviewed publications, and the sharing of computer code *via* GitHub, etc.

The future will involve more accurate small globular domain structure prediction, multidomain protein structure prediction, and better insights into intrinsically disordered protein (IDP) behavior (e.g., phase separation). Infrastructure assets include open-access PDB data, CASP, peer-reviewed publications, and sharing of computer code. In particular, there is a need for many more NMR data on IDPs to be deposited to BMRB.

Progress in protein-ligand interaction research depends on docking or pose prediction, scoring of predicted poses, accelerated medicinal chemistry optimization of lead compounds into potent and selective drug candidate molecules, and more accurate prediction of off-target binding to understand and reduce medication side effects. Infrastructure assets include open-access PDB data, Continuous Evaluation of Ligand Protein Prediction (CELPP, coordinated with wwPDB), peer-reviewed publications, and sharing of computer code. Thousands more cocrystal structures from industry ought to be deposited to PDB. Funding for

Drug Design Data Resource has ended: there is a pressing need for a Drug Design Data Resource Version 2.0.

Protein-protein interactions have to be studied by prediction of binding interfaces, prediction of interaction energetics, quantitative understanding of protein interaction networks, and accelerated protein chemistry optimization of monoclonal antibodies and biologics. Infrastructure assets include open-access PDB data, Critical Assessment of Predicted Interactions (CAPRI, coordinated with wwPDB), peer-reviewed publications, and sharing of computer code. Many more 3D electron microscopy structures need to be deposited to PDB to build up our knowledgebase of protein-protein interactions.

Sustaining open biostructure data presents challenges and opportunities. These include interoperation with additional primary data resources (e.g., correlative light microscopy); interoperation with related knowledge bases (e.g., UniProt); access to data generated within biopharmaceutical companies; and modeling of cells, tissues, organs, and populations of organisms (e.g., microbiomes). Infrastructure assets include the wwPDB partnership, open access to PDB, EMDB, and BMRB, etc., and knowledge bases such as CATH, Structural Classification of Proteins (SCOP), UniProt, and National Center for Biotechnology Information (NCBI) resources. ("CATH" stands for classification of protein domain structures: protein class (C), architecture (A), topology (T) and homologous superfamily (H)). Finally, researchers, data stewards, science funders, and industry leaders need to work together to develop sustainable funding models to ensure the next 50 years of open-access biostructure data.

## 22 Panel discussion

**In the chair: Nathan Brown (NB)**
**Panelists: Stephen Burley (SB), Charlotte Deane (CD), Jeremy Frey (JF), Derek Lowe (DL), Oscar Méndez-Lucio (OM), Tudor Oprea (TO), Chris Swain (CS)**

**NB:** Over the past few months of this seminar series we've had lots of viewpoints from different people in the field. Derek's talk considered whether the medicinal chemist cares about this. What is going to change? We've heard about recent advances such as AlphaFold 2 and we've talked about molecular dynamics, or protein dynamics. What are your views on the big advances that have been made and will be made, that are going to make big changes in drug discovery, specifically, but also in all aspects of scientific discovery?

**CD:** It's almost changing what you need people for, and what you're going to do, and it starts from the assumption that you will have a structure. If you start from that assumption, when you're thinking that you've got a target, you will have a structure. It changes lots of things. Even if you don't write any new code right now, it changes code you would run. There are lots of things you would do if you had a structure right at the start that you don't do currently, because you often don't have a structure until much, much further in. So, it will change the way we use the tools we already have. I think that's the first thing I want to say.

It also changes what people need to think about right at the beginning of a drug discovery campaign, and what they would be doing, because the minute you have a target, in theory you can have a structure of that target. So you would run everything from a structural point of view as well as all the other things you already do. And then I think the other thing is that it immediately starts asking you all the other questions. Lots of these have been raised through all the talks that people have been giving and they are questions that people were working on

anyway, but they were working on, if you like, starting from a different space. Yet if you once again start from the assumption you have a structure, it changes the way you do the question. So if you're trying to improve docking, you no longer think, "Well if I don't get a crystal structure then I'm in deep trouble because I'll have a model and it doesn't work very well". You'll think "Well, I'll have a really decent structure, so what are the things I need to do in docking that would improve that?"

It's like a fundamental shift, but it's a shift that just kind of moves lots of things earlier into the discussion. It's a bit like when we first got all the genome sequences: did that change everything? Well it did, but what it really did was it said, "Well we answer these questions but now we have this piece of information to start from, and so we'll start from there." I should be honest as well. I think there's still quite a long way to go in terms of saying "I will have a structure." We don't yet know if we'll have structures of complete complexes or be able to dock things, or the relationship between the apo and the bound structure, so I think there are lots more pieces that will come, but those are developments off that rather than trying to model that with no basic starting structure every time.

**SB:** I just wanted to point out that we did an analysis of 79 new anticancer drugs that were approved by U.S. FDA from 2010 through 2018 and looked if 3D structures of every one of the targets of those drugs had been in the PDB for more than 10 years prior to the approval. We were able to find evidence either in the literature or from companies that more than 70% of the small molecule anticancer agents that were approved in that time period were the products of structure-based drug discovery. Most of the 3D structures initially came from academe. There was one case that I know of where the first structure of the human protein came from a company, but there had been a structure of a porcine homologue that was 95% identical or so in the structure for more than a decade before they started the project. So, I think that AI is going to come in and the use of AI etc. for protein structure prediction is going to improve things, but the fact is that in areas of intense interest in biomedicine, there's already a huge number of structures in the PDB waiting for people to utilize them.

**DL:** One of the things I'd add to that is that for, especially in the oncology field, many of these approved oncology compounds are perforce kinase inhibitors, and of course there are a tremendous number of kinase enzyme structures and kinase enzymes with bound inhibitor structures in the PDB, so this makes it a really a special case. This is the area where we have probably the most to go on for structure-based drug design, and where we can confidently talk about DFG-in, DFG out and subtle variations in the binding mode. So, that is probably the best case in all of structure-based drug design if you go into the kinase area. So, if we can get that extended to more areas where we can confidently talk about minor changes in the protein and its effect on both downstream efficacy and its effect on compound binding for new SAR compounds, we will have made a real advance.

**OM:** I just wanted to follow up on what the Derek said in his talk. I think that what's coming up next for AI/ML applied to proteins is not just the structure but how the structure is moving. We need to understand more about how all these small changes are affecting the binding and how we can develop methods that help us to model the kinetics and dynamics in a cheaper and faster way.

**NB:** Yes, it's important to remember the dynamics.

**TO:** I guess I was going to suggest that perhaps the true benefit of AI, if there will be one that's distinctly clear, is if we're able to run and to enter predictions: going from the binding

mode and protein structure all the way to even clinical trials outcome, like the kind of work that Benevolent AI did with Baricitinib. That would be a type of AI work. So, what I would like to see is the possibility to have this extended in a seamless manner and then, perhaps in five to 10 years from now, we might see a fierce battle with the U.S. Patent Office as to whether AI can be an inventor. Some of you who may have followed this know that, so far, that possibility has been declined. You have to be a human to be written on a patent. That's a fascinating discussion: I was following Andrei Iancu (then USPTO Director) the other day discussing this.

There was another thing I was going to suggest. If it's possible for protein structure prediction to make an impact, I would really like to see that impact to go also into the prediction of off-targets. We struggle a lot with secondary pharmacology. We struggle a lot with understanding what causes side effects. It would be really great if you guys can contribute to that.

**NB:** Thanks, Tudor: it's really helpful. Going back to some of Derek's comments on protein structures and drug discovery projects, I've worked on structurally enabled projects and nonstructurally enabled projects in the past and I've always wondered whether there's an inflection point when you have a structure in any enabled project. This is an open question to anyone who's ever done this. I've got some opinions but I'd like to keep neutral here. Does anyone have a feel for that?

**DL:** Yes, I can speak to that. I also want to add something to the talk about dynamics. So, as part of my function of being the bucket of cold water, I was going to quote a colleague of mine a few years ago when he started seeing a lot of MD-type simulations. He said, "You know, if I don't trust a single docking pose in a static form, what do I feel about 30 frames a second of those things?" So, I would just add that when you start, if you start doing computational approaches to dynamics, it had better be good. It really had, and this is not news to anyone who's doing it, but it had better be good because you're going to rapidly go astray as small errors accumulate in your time course. So, it'll be quite interesting to see how that works out. It is ferociously computationally expensive of course but I'm hoping that the hardware and software people will come to save us there.

As far as working on projects that have a structure available and don't, I guess I've been in all four quadrants. I've been on projects that had a structure where it really did make a difference and it really did help in making predictions. I've had some where we had a structure where it didn't seem to make any difference at all because either the project already had very good compound leads from screening and the SAR by traditional medicinal chemistry made sense, so the structure predictions were either telling us things we had already found out for making compounds or were at variance with the data. And I've also been on some where not having a structure made no difference, and if you were not having a structure, you really felt the lack because you needed the guidance. So, you can end up in all four of these just depending on the project.

**CD:** I wanted to say something about what Derek just said. I was starting from what difference something like AlphaFold makes to what's happening, and obviously there's the recent paper that the Rosetta group put out as well. If I don't trust the docking results, why would I trust the interpretability? I think that is going to become a key part of this. How do you make the link between the people and the machine answers? That's the level that we're at now, which is probably convincing chemists or anyone who's working in the wet lab that the results I am giving them make sense and that they can feed back into them. So, there's a

feedback loop of what they see and can interpret what's happening in in terms of that, and the algorithms need to be able to do that.

I was involved in a discussion yesterday about how we'd love a pipeline that went all the way to the end. How do you even persuade a regulator currently to allow you to continuously update your regulations for allowing someone onto the clinical trials based on new data as they arrive because that involves changing the model that you specified the beginning of your clinical trial? At the moment that's a very complicated question and they're actively considering it. These models do require, if you're going to improve them, continuous changing of your data. So, I think interpretability and understanding is going to become one of the really key questions for this, for use of all of this technology, from the sort of atomic interactions all the way to the other end.

**DL:** So, Charlotte, does that mean we're going to see more Bayesian designs, so you explicitly decide to take advantage of that from the start.

**CD:** I think you should. I think it will come more and more because on the one hand we'd love a black box, but we're never going to trust the black box until we've spent a lot of time understanding what the black box really does and how it works and what it can do. We also have to imagine that we might be operating sometimes where the rules of the game are changing, because if we do an intervention, it changes the results that we get back. These are things to come; I hope that we will have more of them, but we will see how it develops.

**Question from Q&A box:** Would you say that a better prediction of protein posttranslational modifications, such as a glycosylation, is important to improve the prediction of protein structures?

**SB:** I'm sure it will make a difference and I'll put in a plug for the recent remediation of all the carbohydrate-containing structures in the PDB that took quite a number of years. Those of you who use the PDB to look at glycoproteins will be pleasantly surprised by the completeness with which we are now providing that information for experimental structures.

**DL:** Yes, I can try to answer. I think that that's good news. I've been seeing this with the PDB and I think it's a great initiative. And of course, Carolyn Bertozzi is going to be, you know, setting off champagne bottles for that because I think that glycosylation has been woefully undervalued and understudied, partly because it's so damn hard to study well, but clearly, because such a large percentage of the proteome is glycosylated, and glycosylated not just in simple ways, but in some fairly odd, complex, and changeable ways. It's telling us that we've got to pay attention to it along with a lot of other posttranslational modifications. Too many people who look at protein structures tend to just say, "Oh yeah, and it's decorated on the outside", like the ornaments on a Christmas tree, but these are not just ornaments. These are incredibly functional and, of course, as we know, dynamic too. These things are going to change under different conditions, so another thing that we're going to have to be able to pay attention to is what the effect of a SUMOylation, or ubiquitination, or farnesylation, or glycosylation is going to do to the protein structure and surface. So this is going to keep us all employed for a very long time.

**TO:** That's the good news. I just put a pitch in the chat section. GlyGen is an NIH Common Fund initiative dedicated to computational and informatics aspects of glycosylation. So please visit it if you want to figure out something new about your protein.

**NB:** Thank you, Tudor. Moving on a little bit, does the panel think we're heading to a place where we can start predicting the whole structural proteome using these computational methods and what would we do with it if we could?

**CD:** The answer is yes, we could. Whether it would be useful or not, is another question. The absolute answer to that one is yes. There are two parts to that. One is that we probably can do that, the other one is to be aware that even with the methods that are proposed now (and this this is going to move through time) you're still not expecting to get completely accurate results for all of that. Yes, there's enough. You look at the numbers that are currently on these papers. It's not perfect yet, so there will still be stuff that we don't know about or stuff that we're less sure about.

I'm really interested to hear what others think about what we would do with this information. I would play back to the days when they said if we solved all the sequences in the human genome, we would solve medicine and we'd be able to make drugs, and we'd know what was happening. I think we're in a similar place here if we start saying the same thing. I think this is another piece of information that will help. It's a great deal and we will use it in different ways to improve things, but I feel it's just another step. I don't want to say it's not exciting because for someone like me, it's very exciting to have access to this.

**DL:** I couldn't agree more. That is such a good point. We have to make sure not to overhype this because there are so many steps in this, but you're right. This is going to be tremendously useful, but there are problems on top of problems waiting for all of this, and this is just going to be another great sword to pick up and attack some of these problems.

**SB:** I think you're absolutely right, Derek. I recall after the genomic sequence revolution, when I was at Rockefeller University, some of my pharma colleagues complained that there were now too many targets. So once they're all structurally characterized, they're going to complain again that there are too many targets, right?

**NB:** It all comes back to target validation.

**SB:** Yes. There was a study published in in the 2000s that suggested about a third of failures in the drug industry were due to efficacy failures, meaning lack of understanding of the target. Hopefully we're going to be able to improve on that and improve on toxicity, but, as Derek pointed out, most of the toxic events that kill compounds in later stages are things that could never have been anticipated and many of the toxic events that kill compounds in earlier stages, I can tell you from personal experience, could not have been anticipated.

**DL:** Agreed, agreed. Well, this is what computational methods, AI and ML, and all these things are doing. They're doing it to organic synthesis and they're doing it to these larger questions. It's pushing us to work at higher level questions as the lower level ones get more and more susceptible to automation and to expert system and machine learning approaches. It's pushing us humans to think about the bigger, tougher questions. We're getting moved up the value chain a bit because our definition of grunt work is constantly changing. You know what it's like. There were people who made careers out of being able to sequence a single protein back in the day, or clone a single gene, and now it's done by machines, constantly, around the clock. So, we're going to see a lot of things that used to be heroic efforts turn into ordinary efforts, turn into something that no human is even going to be bothering to do, and it's pushing us to the tougher questions. I hope we're up to it.

**NB:** Can I ask a slightly philosophical question then? One thing about the prediction of protein structures is that it's a fantastic intellectual challenge and it caught people's imagination. What will be the next one then for computational and AI drug discovery? If it's in drug discovery, you're going to move towards the area where people sort of think they can do it, but they can't, which is in the area of things like docking. "Can you tell me if this will bind to this? And can you tell me how well it will bind?" To me, it's got to be a question that I can phrase that simply. So why did we like protein structure prediction? It was because it said "I'll tell you the sequence, you tell me the three-dimensional structure". So, here I'll say, "Right. Well, I know the structure, tell me if this will bind and tell me how well it will bind". It's "where" as well, and all those things, and what will its binding affinity be?

**DL:** Yes, I agree with that, and it would be tremendous if it can enable virtual screening the way that everyone's been dreaming of doing virtual screening for the last few decades. That would be a real event.

**CD:** That's also very amenable to the same kind of blind competitions that we have had for other things. And that's something that has proved to be a useful way of testing that we're actually doing what we say we're doing, but also of driving innovation, because it makes us think about the problem and actually say that we can't do this because when we're properly tested on it, we fail. Those kinds of questions tend to be the ones that catch the imagination.

**CS:** I agree with that 100%. So I guess that partially goes to what Tudor announced in his presentation about the CACHE initiative, the critical assessment, the virtual screening.

**DL:** Yes, there's one thing that that discussion brought to mind. In writing the blog, I deal with a lot of people outside of the field, and one of the things I had to explain to people over and over after the AlphaFold results was that this, to me at least, was not what you would have pictured 50 years ago. If you told people that we had done this tremendous advance in understanding protein folding, they'd have thought we'd understood a tremendous amount more about the intimate mechanics and thermodynamics of protein behavior. What this was was actually a triumph of pattern recognition enabled by the huge amount of data in the PDB and, of course, I don't have to tell any of you that, but a lot of people who don't know the field well don't realize that at all. That means that this was (I hate to sound like this) comparatively low hanging fruit because we have the enormous advantage of a PDB full of high-quality protein structures, whereas for some of these other problems there is not a similarly existing body of reliable data to do pattern recognition off, so we're either going to have to generate it and do our pattern recognition tricks, or we're going to have to get smarter.

**OM:** For me, what's valuable is that a very complex process like this folding is trained and many, many thermodynamics can be just reduced to a few patterns. Yes, it is pattern recognition. Yes, it's a lot of data, but I think it can help us to change the way we think for difficult problems. So perhaps we can reduce many problems we have in drug discovery that we think are difficult and produce something simpler that can really help us solve the problem.

**SB:** I wanted to say two things about the AlphaFold triumph, and it was a triumph, and what I'm going to say next is in no way meant to detract from their accomplishments, and the accomplishments of the protein structure prediction community writ large. This was not a solution to the protein folding problem. This was a step forward, a substantive step forward in the protein structure prediction arena, but it was touted as if we now know how proteins fold and we do not know how proteins fold based on what was done by any of the practitioners of protein structure prediction.

I also wanted to say something about the blind challenges. Charlotte correctly pointed out that the history is very clear. Having been directly involved in in the Drug Design Data Resource challenges, I can tell you that I was actually quite shocked to realize, once we got into the process with Mike Gilson and Rommie Amaro at UCSD, that this five-year project, which ran from about 2015 to 2020, showed that the docking and scoring community had made little progress since we were doing that kind of work at SGX Pharmaceuticals in the early 2000s, before we turned our attention to fragment-based screening for binding to protein targets using X-ray crystallography, along the same lines as Astex and other companies. The field went for a decade without blind challenges and really didn't make a lot of progress in substantive terms, in terms of improving the method. And I regret very much that when we went in to try to renew the Drug Design Data Resource project, the National Institute of General Medical Sciences was not interested. The fact that the need was laid bare did not persuade them that the project should be renewed. So, I hope that the AI successes will perhaps trigger a rethink at NIGMS.

**CD:** I think that the blind challenges are a key thing here and they do rely on the experimental community being happy to do this and you can't do them everywhere because there are questions that just don't fit into there. Even if I think about the work that's now ongoing in terms of generative models for molecules or trying to design better molecules at the moment, it's very difficult to benchmark what those techniques are doing and how useful they are. There are tons of data, so the pattern organization there is obviously there. You can generate good molecules, but actually, telling whether any of these techniques are truly useful is very difficult at the moment, apart from the fact that they make things similar to what have been seen before, which probably isn't the best metric for that kind of thing, so I make this plea. I think if you're sitting on the computational side, benchmarks and the blind tests and all these things are so important, partly for convincing people that these things work, but also to drive people to actually do improvements rather than prove that my model looks 1% better on a particular dataset than someone else's. That doesn't tell me enough to know whether that change in architecture is worthwhile or whether we want to develop that way.

**OM:** Just a final comment based on the challenges and doing all this benchmarking. I think for an AI/ML community, it should be very, very important to keep an eye on real projects and on real data. Something that has happened in other fields like computer vision is that they really make the models just to work for specific benchmarking data, and yes, they're better, but once you apply the same model in a real setting, a real project, it's not working the same. So I think we need to find ways to test our models and our methods in real data.

**DL:** Exactly. It's the same problem as we were talking about earlier during my talk about people running clinical trials in a way that is too protected and too perfect. You want to go to the real system as hard as you can, as quickly as you can.

**SB:** Let me just respond to the point Derek made. After my six years as Chief Scientific Officer and head of R&D at SGX Pharmaceuticals, we were acquired by Eli Lilly and Company. So I spent four years in the belly of the beast in Indianapolis, half time, and what I noticed at Lilly was that people were rewarded for success, not for failure. So, decisions were made and experiments and studies were designed in such a way as to maximize the likelihood of success, just as Derek was pointing out, *versus* trying to do the killer experiment to kill the project. And I noted that. Then I read that Google have a very different philosophy and they actually reward people for killing projects. They reward people for finding flaws and then either redirecting the project or possibly even killing the project, as opposed to punishing

anybody who fails, which is certainly what was going on at Lilly.

The logical extension of that philosophy was that every middle manager had a portfolio of projects of their own, because if you had only one project, the likelihood of failure is high. As Derek pointed out, you would fail, you would not get promoted. So that meant that there was a proliferation of drug discovery projects within the company in order that every middle manager could be responsible for multiple projects, so they could always point to one thing that was moving ahead. When a colleague from Merck joined us, he looked under the hood and was appalled at the number of projects that were being worked on just in his own therapeutic area. He remarked that at Merck the total number of projects that the whole company was working on, spending more money on R&D than Lilly, was smaller than the total number of projects that his therapeutic area was working on at Lilly. Honestly. So what were we doing at Lilly? I said this openly at the company and I was shunned for it. We were working on too many things and systematically under-resourcing all of the projects and turning potential winners into certain losers by under-resourcing them.

**DL:** I can really corroborate that experience. I have worked at organizations that claim to reward people for killing projects, but it has been a rare event when any of them actually walked that walk, when someone did kill one. Generally, when a project failed, the human nature impulse was to look for someone to blame and success always has 1000 parents and, at best, failure has a single parent. In many cases, failure was an orphan. Nobody would take responsibility; it was nobody's fault. This philosophy is still there.

**TO:** I think it's a good time to interject. Earlier I could not answer the question of what we should do next with protein structure prediction. We could use 3D structure prediction to imagine protein complexes and imagine how they are situated in a cell and literally create the living model of a cell at the molecular level. I know that there are some attempts already being made, but I'm not sure how realistic they are. Prediction coupled with enough experimental data and metadata could actually come up with a virtual cell that is realistic at the atomic level. I would be really glad to see this happen in the next decade or so.

**CS:** Getting back to the idea of blind challenges, I wonder, Jeremy, if AI3SD and RSC-CICAG should get together perhaps and think about putting together a prediction of binding affinity challenge.

**JF:** I'm sure you'll agree that the AI3SD network has been highly successful. Due to COVID, it is extended until March 2022, but we are looking beyond that and we're discussing it actively with UKRI. How and what things should be taken forward? What have been the successes and the things we should concentrate on, including actually doing more work ourselves and making more arrangements for other people? I actually think that it's been one of the clear statements today that blind challenges are extremely useful at driving forward innovation. To evaluating that work in a really good and fair way we now clearly have to think about finding the right amount of data and things to evaluate against. I think that it would be a really good idea to take that forward. The platforms necessary to do these sorts of things obviously exist, and I wouldn't want to reinvent them. I think there are always ways of doing this. Online sort of game things, Kaggle and others, provide the basis for some smaller scale ones. That's a little less controlled and for different purposes, but I think it would be a really exciting thing to do [a challenge].

**Samantha Kanza:** Definitely. We will definitely do that before March or make it part of our rebid.

**NB:** I think AI3SD did fund Matt Todd to do a study.

**Samantha Kanza:** It was more of a phenotypic type challenge and Chris wrote the report on it.

**JF:** Excellent. We will consult with you about how to take some more of these things forward. So maybe I should ask the panel, and the audience, as well as drug affinity, what other things more generally to do in terms of proteins. But perhaps even more generally, are there missing things in the evaluation space. Maybe it's only the affinities, or maybe that's the next best one.

**CS:** Sure, there must be more. One of the things that I've been following from afar is materials for photoelectric cells and things like that, but I think that people have their own ways of doing these things.

**JF:** Well, I think that's an interesting analogy. I mean, that's been the advantage of this network in covering chemistry and materials; that is you see the same thing. You can develop the new material. You still have the same fundamental questions of why you know you've got there. The next thing is you can understand why. And that leads you to hold more experiments but ultimately you want this thing to work in the device just as you want the drug to work in a person, or the insecticide to work. Actually it's designing it for the ultimate device that turns out to be much more difficult than just getting the new material.

Equally, it is really important what happens at the end. Maybe you can avoid an end. What's the circularity here? Otherwise, you end up dumping something into the environment, either the drug or the material for disposal, which is nonrecyclable, nonreusable, non-whatever. Actually, what we should do is even unclear. So designing end of life in is another thing that needs to be put into these models. So, you've got to be able to make the thing you choose, You've got to make it, it's got to survive for long enough in the device or the person or whatever, and then the disposal has got to be reasonable. This holistic thing is extremely difficult and again, hopefully, all these data are coming in. The ability of these machine-learning and traditional AI types and knowledge management systems hopefully will show up to experts in individual areas where they need to talk to other experts and try and solve some of these problems.

**NB:** Jonathan Goodman has been adding some blind challenges into the chat. I don't know if he wants to cover anything he mentioned.

**Jonathan Goodman:** "Why on earth have we got that molecule?" is a really good challenge which we're playing all the time. We just don't realize it. And so there are lots of examples of people not getting the right answer. "What is that reaction?" is also quite fun. Did that work the way we thought it did? Metal organic frameworks are a complete mystery to me. That might be because I just don't understand them, but they do sound rather fun. Boil stuff up and see what happens.

**DL:** Because this is science, but they are so much fun. I did a metal organic framework project for a little while and I don't know when I've had so much fun in the lab, but yes, there is a lot of mystery there and a huge amount of potential for catalysis, separations, and other things. That is a great field for this sort of approach, and if there's some sort of machine learning AI approach that can help make sense of it, I know people are trying, then they should go for it.

**NB:** Thank you. I think Tudor just added one in the chat as well. I don't know if we want to cover that today. So, why are some drugs better than others when affinity is similar? What is the efficacy challenge?

**TO:** I guess what I was trying to say is that the two things that really matter in drug discovery are toxicity and efficacy and we basically don't do very well at all at either of those. So I think of having good challenges that would really work.

**DL:** Yes, I agree. One of the things that catches us out is that there are a lot of things going on in the cell that we just have no clue about. There is a craze in recent years for biomolecular condensates (aka membraneless organelles): these phase-separated droplets that happen spontaneously in cells. There's a whole world of intracellular pharmacokinetics and distribution that we have not been aware of, and I'm sure it's going to turn out that there are compounds that we tested that we thought would work well and didn't, and compounds that worked a lot better than we expected in the cells that will turn out to be that way because they either got into the right condensate dropper compartments or were excluded from others, and that will probably also have effects on toxicity too.

We are profoundly ignorant about the conditions inside the cell, and I think this is where I get back to that unsexy thing: we need more knowledge about biology because with the level of knowledge we have, we cannot understand. Make gold out of straw with these things. Right now we have an awful lot of straw in cellular biology. You would not have been able to predict the existence of intracellular condensates 20 years ago in the same way that you wouldn't have been able to predict the fact that there is RNA interference or short hairpin RNA, and things like that. Twenty years from now, people are going to look at us and say, "Oh those poor people, they didn't know about X, they didn't know about Y, no wonder they couldn't figure anything out."

**NB:** That's very true. We've got 10 minutes left. I'm wondering if the panelists would like to summarize some of their take-homes from the past couple of days during the conference. We've punctured some of the hype, maybe, would that be fair?

**SW:** I think that's true, Nathan, or perhaps put things in perspective. And I guess also we're identifying what the real problems are as opposed to solving the challenge of predicting protein crystal structure; thinking ahead of what we do next.

**NB:** Dynamics is definitely a key theme that's come through, and I know we've been talking about that a lot. Any other highlights from anyone? We've got some time. So may I ask a provocative question of people who do use AI for protein structure prediction, who may be on the call?

**SB:** I have a long-time colleague who in a previous life in a pharma, before he worked for GSK, worked for an AI company. His reaction when I asked him about AlphaFold etc. was to say, "What took them so long?" I wondered if anybody else in the meeting had had a similar reaction.

**CD:** I will kick off here. I will say I did not think what took them so long. I actually thought they were slightly further ahead than they were, but I wasn't completely shocked when it happened. I think this comes back to the comments that have already been made about the fact that it was built on an enormous amount of data that existed. What if we hadn't solved all those structures and experimentalists hadn't done all that work? This isn't some magic

where AlphaFold did it from first principles. It's not like that they needed all of that. And then the other part to remember is actually most of the pieces for doing it were out there in the academic community in terms of how you solve the problem. We don't completely know what architecture has been used for some of this yet because it's not been published, but there's enough that's been drip-fed out that we have a fairly good idea.

The other part I think, in terms of what took them so long, is there is one issue (and I don't know how much this has been talked about at the conference): actually to do that would have been really difficult for an academic group. I wondered if this is something people wanted to address: the compute that they use, the type of computer, and the size of compute to train those models and to be able to run it repeatedly and test what is the best way of doing it. I have a reasonable amount of compute at my disposal but I could never have considered doing this kind of question. I didn't have those pieces. I couldn't have done it. There are bigger groups, and there are people with bigger computers in academia, but actually it's interesting to me that in this realm we have to start thinking about how we're going to work like that, because there is a real switchover here. We don't all have the access to that kind of computer, and it's not just a bigger computer, it's a particular design of computer or the specific GPUs that will run these things fast enough. Some of these questions you can't do, even if you can theoretically work out what might be interesting to do, because you can run one instance of one model, but you can't do the thousands of different versions to try and work out the best one for doing it.

To me, it's not what took them so long, it's how can I make it so that groups and academics have access to be able to start doing these things as well, or being able to contribute to that level. I reviewed Derek's blog post from the AlphaFold announcement last year and some of the comments. It's always interesting to go into the comments. Some of the comments were around why big pharma hasn't done this. And that's something to consider if they do have that kind of resource.

**CS:** Yes, I guess big pharma didn't regard it as a rate-determining step.

**DL:** It's true. As far as reading comments go, I can recommend reading the comments on the non-coronavirus posts. Don't read the comments on the ones that talk about coronavirus vaccines. It's a sewer.

**JF:** I think what Charlotte is indicating is a major problem that we have that we've seen throughout the network. It's not just the computing power. First of all, I think one of the reasons we saw that you get into this state is that you need to know science, lab science, theoretical chemistry, physics, whatever, and the mathematics and the computing, otherwise you don't appreciate what you really need to do and how difficult it is, and what the challenge to the computer sciences is, or what other things need to be added to it.

Yes, when you've got lots of data, you can kind of sort it out. So, you need a different breed of students than we're educating and bringing through. They've got to have at least enough knowledge of these areas to work with others and enough people who can be in the middle and pull it together. Otherwise, it's very difficult to solve these problems or solve the right one, and you could argue they maybe put a lot of compute power into solving the wrong one. You have to conceive of doing it to do it and then pull the right people together. The tools were all there, or probably all there, but they weren't assembled, and I think we're seeing more and more of that, and we've got to generate the right educational environment to produce more people who can think that way because that is the way forward. One of the difficulties is not

just that we don't have the compute facilities to show them this in academia.

I don't necessarily think we're always doing the cutting edge work in academia, because a lot of it's going on in industry. We don't always know about it. Fortunately, it's an area where industry is making more things open, and so we can learn from them, but it's not ideal the way that the leading groups work. I mean, they work with academia, but there is, I think, a tension here, and I think that will cause a real problem for the next generation coming through. We won't necessarily be training them or giving enough imagination to them about what could be done and that does worry me. I think we've seen that in our network when trying to bring things together.

The positive thing is that we brought a lot of industrialists into an academic network because they're interested in the academics and in what academics want to hear. So, it's been more mixed than most networks. I think that the industrial community has learned from the academic community in those areas and I think a lot of companies realize we can do this so we can do that, but what we need is both. We can't do it on our own. The push towards more open source tools from both academia and industry is helping these groups collaborate.

**CS:** I think Charlotte makes a really good point and it's followed on by Jeremy, but one other thing I observe in looking at academic groups working in these areas is that they don't often have the domain expertise. They tend to be very specialist in a certain area and if you want to have an impact on a wider area you need groups that bring in perhaps biology, chemistry, physics, statistics, whatever, all to work together, and sometimes it's a challenge to do that in an academic environment. I know of projects where they've applied for funding, and they've been told it's too chemical, or it's too physical, or too biological. There just doesn't seem to be a way of perhaps funding those sorts of projects.

**SB:** I wanted to pick up on Jeremy's point and say that it, and the point that Chris made, were all well taken and to urge a sort of critical rethink of how we actually train undergraduates. Certainly the academic institution that I am part of is extremely siloed when it comes to undergraduate training and I think that that's a mistake, and even the graduate training is quite siloed, something that we're pushing against with our own graduate program at the Institute for Quantitative Biomedicine that I established at Rutgers. People have got to be able to talk to people and working together have to be able to talk the same language, as Jeremy was saying. And, teaching, imbuing that multilingualism in science, is something that needs to be done early. I think Charlotte can talk to some of this in the CDT [Center for Doctoral Training].

**CD:** Yes, I was going to say I think I feel slightly less negative about academia than was put there, in the sense that I work in a group which has people from eight or nine different disciplines. It seems completely natural to me. It's been natural throughout my career. I head up a CDT which I've run for a very long time, which takes people from across multiple disciplines, or working in, toward an area. I think it can be done and I think it is being done.

I think if you ask me the challenge, it's difficult to imagine doing some of these questions if you don't have the facilities in your hands to be able to attempt them. So you can imagine them in a kind of abstract way such as writing a paper on what could be done, but you can't do it. I agree with things that Stephen is saying about better training people to be multidisciplinary, but let's not argue that if you go into Deep Mind, they have a highly different set of people who came through a different undergraduate degree from the ones we've got.

Yes, it's about how you use those people, how you work those teams together, and the resources you have to go alongside that, and where you choose to put those resources. You know the questions you choose to answer, so I'd like us to finish a little bit more positively and say that enormous things have been done both in industry and in academia. We've made massive leaps here, but they don't answer the question. We've not finished yet. They just give us more things to be able to do and we should keep trying to do those things and keep building off them and keep improving. We need to be more multidisciplinary, keep being more open, and keep the conversations happening and eventually we will answer some of the very hard questions and just get better.

**NB:** I came from a multidisciplinary team in Peter Willett's lab in the early 2000s, so I agree.

**JF:** Charlotte is absolutely right. There are great examples. The U.K. escience program ran for a while and generated people. What the key thing is, what the difference is, is that you have to be part of a team so you don't have to feel your expertise is being used to support somebody. It's a critical part of the project. The statisticians have probably been the people at the bad end of that: for a long time they've been brought in at the last moment, and they need to be in at the beginning and so on.

It's understanding how to run that team, and that does sometimes run a bit counter to the way universities want to run. What have you done to get your promotion? That was one of the things we faced with the escience program bringing in lots of people together. I think we learnt some of the lessons but we've forgotten some of them. The CDT and others have learned them and continued forward. I think that's the trick that's perhaps easier to do in industry, where you had the project to help. You want the people, the people perhaps at the center of the things in some of the university environments and then understanding how you all work together to carry a really important project out. Of course, when you do that and then you don't have the resources to complete it, then it's deadly. We can do this. We know how to do it, but we have to get the right people in together and we want them to work together.

**NB:** I think that's a great point to finish on for the panel discussion. I just want to thank all the panelists again for the really interesting conversation. We moved around from technology to people to how we work together, and the data. All these things are hugely valuable and that's how we ended up where we have.

## References


(1) Arnold, F. H. Design by directed evolution. *Acc. Chem. Res.* **1998**, *31* (3), 125–131, DOI: 10.1021/ar960017f.

(2) Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G. A.; Sonnhammer, E. L. L.; Tosatto, S. C. E.; Paladin, L.; Raj, S.; Richardson, L. J.; Finn, R. D.; Bateman, A. Pfam: the protein families database in 2021. *Nucleic Acids Res.* **2021**, *49* (D1), D412, DOI: 10.1093/nar/gkaa913.

(3) Price, M. N.; Wetmore, K. M.; Waters, R. J.; Callaghan, M.; Ray, J.; Liu, H.; Kuehl, J. V.; Melnyk, R. A.; Lamson, J. S.; Suh, Y.; Carlson, H. K.; Esquivel, Z.; Sadeeshkumar, H.; Chakraborty, R.; Zane, G. M.; Rubin, B. E.; Wall, J. D.; Visel, A.; Bristow, J.; Blow, M. J.; Arkin, A. P.; Deutschbauer, A. M. Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* **2018**, *557* (7706), 503–509, DOI: 10.1038/s41586-018-0124-0.

(4) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25* (17), 3389–3402, DOI: 10.1093/nar/25.17.3389.

(5) Finn, R. D.; Clements, J.; Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **2011**, *39* (Web Server), W29, DOI: 10.1093/nar/gkr367.

(6) Bileschi, M. L.; Belanger, D.; Bryant, D.; Sanderson, T.; Carter, B.; Sculley, D.; DePristo, M. A.; Colwell, L. J. Using deep learning to annotate the protein universe. *bioRxiv* **2019**, 626507, http://www.biorxiv.org/content/10.1101/626507v4.full.pdf (accessed May 8, 2021).

(7) Angermueller, C.; Dohan, D.; Belanger, D.; Deshpande, R.; Murphy, K.; Colwell, L. Model-based reinforcement learning for biological sequence design (conference paper at ICLR 2020). http://cangermueller.com/wp-content/uploads/2020/02/Angermueller-et-al.-2019-Model-based-reinforcement-learning-for-biological-.pdf (accessed May 8, 2021).

(8) Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* **2017**, http://arxiv.org/pdf/1707.06347.pdf (accessed May 8, 2021).

(9) Bryant, D. H.; Bashir, A.; Sinai, S.; Jain, N. K.; Ogden, P. J.; Riley, P. F.; Church, G. M.; Colwell, L. J.; Kelsic, E. D. Deep diversification of an AAV capsid protein by machine learning. *Nat. Biotechnol.* **2021**, *39* (6), 691–696, DOI: 10.1038/s41587-020-00793-4.

(10) Vollmar, M.; Parkhurst, J. M.; Jaques, D.; Basle, A.; Murshudov, G. N.; Waterman, D. G.; Evans, G. The predictive power of data-processing statistics. *IUCrJ* **2020**, *7* (2), 342–354, DOI: doi:10.1107/S2052252520000895.

(11) Winter, G. xia2: An expert system for macromolecular crystallography data reduction. *J. Appl. Crystallogr.* **2010**, *43* (1), 186–190, DOI: 10.1107/s0021889809045701.

(12) Winter, G.; Waterman, D. G.; Parkhurst, J. M.; Brewster, A. S.; Gildea, R. J.; Gerstel, M.; Fuentes-Montero, L.; Vollmar, M.; Michels-Clark, T.; Young, I. D.; Sauter, N. K.; Evans, G. DIALS: implementation and evaluation of a new integration package. *Acta Crystallogr., Sect. D: Struct. Biol.* **2018**, *74* (2), 85–97, DOI: 10.1107/s2059798317017235.

(13) Evans, P. R.; Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2013**, *69* (7), 1204–1214, DOI: 10.1107/s0907444913000061.

(14) Wang, H.; Yan, X.; Aigner, H.; Bracher, A.; Nguyen, N. D.; Hee, W. Y.; Long, B. M.; Price, G. D.; Hartl, F. U.; Hayer-Hartl, M. Rubisco condensate formation by CcmM in $\beta$-carboxysome biogenesis. *Nature* **2019**, *566* (7742), 131–135, DOI: 10.1038/s41586-019-0880-5.

(15) Lee, J.; Crampton, K. T.; Tallarida, N.; Apkarian, V. A. Visualizing vibrational normal modes of a single molecule with atomically confined light. *Nature* **2019**, *568* (7750), 78–82, DOI: 10.1038/s41586-019-1059-9.

(16) Zhang, Y.; Luo, Y.; Zhang, Y.; Yu, Y.-J.; Kuang, Y.-M.; Zhang, L.; Meng, Q.-S.; Luo, Y.; Yang, J.-L.; Dong, Z.-C.; Hou, J. G. Visualizing coherent intermolecular dipole-dipole coupling in real space. *Nature* **2016**, *531* (7596), 623–627, DOI: 10.1038/nature17428.

(17) Xia, R.; Kais, S. Quantum machine learning for electronic structure calculations. *Nat. Commun.* **2018**, *9* (1), 4195, DOI: 10.1038/s41467-018-06598-z.

(18) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555* (7698), 604–610, DOI: 10.1038/nature25978.

(19) Zhong, M.; Tran, K.; Min, Y.; Wang, C.; Wang, Z.; Dinh, C.-T.; De Luna, P.; Yu, Z.; Rasouli, A. S.; Brodersen, P.; Sun, S.; Voznyy, O.; Tan, C.-S.; Askerka, M.; Che, F.; Liu, M.; Seifitokaldani, A.; Pang, Y.; Lo, S.-C.; Ip, A.; Ulissi, Z.; Sargent, E. H. Accelerated discovery of $CO_2$ electrocatalysts using active machine learning. *Nature* **2020**, *581* (7807), 178–183, DOI: 10.1038/s41586-020-2242-8.

(20) Ye, S.; Hu, W.; Li, X.; Zhang, J.; Zhong, K.; Zhang, G.; Luo, Y.; Mukamel, S.; Jiang, J. A neural network protocol for electronic excitations of N-methylacetamide. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (24), 11612–11617, DOI: 10.1073/pnas.1821044116.

(21) Ye, S.; Zhong, K.; Zhang, J.; Hu, W.; Hirst, J. D.; Zhang, G.; Mukamel, S.; Jiang, J. A machine learning protocol for predicting protein infrared spectra. *J. Am. Chem. Soc.* **2020**, *142* (45), 19071–19077, DOI: 10.1021/jacs.0c06530.

(22) Hu, W.; Ye, S.; Zhang, Y.; Li, T.; Zhang, G.; Luo, Y.; Mukamel, S.; Jiang, J. Machine learning protocol for surface-enhanced Raman spectroscopy. *J. Phys. Chem. Lett.* **2019**, *10* (20), 6026–6031, DOI: 10.1021/acs.jpclett.9b02517.

(23) Jia, C.; Wang, X.; Zhong, W.; Wang, Z.; Prezhdo, O. V.; Luo, Y.; Jiang, J. Catalytic chemistry predicted by a charge polarization descriptor: synergistic $O_2$ activation and CO oxidation by Au–Cu bimetallic clusters on $TiO_2$(101). *ACS Appl. Mater. Interfaces* **2019**, *11* (9), 9629–9640, DOI: 10.1021/acsami.9b00925.

(24) Wang, X.; Ye, S.; Hu, W.; Sharman, E.; Liu, R.; Liu, Y.; Luo, Y.; Jiang, J. Electric dipole descriptor for machine learning prediction of catalyst surface-molecular adsorbate interactions. *J. Am. Chem. Soc.* **2020**, *142* (17), 7737–7743, DOI: 10.1021/jacs.0c01825.

(25) Suruzhon, M.; Bodnarchuk, M. S.; Ciancetta, A.; Viner, R.; Wall, I. D.; Essex, J. W. Sensitivity of binding free energy calculations to initial protein crystal structure. *J. Chem. Theory Comput.* **2021**, *17* (3), 1806–1821, DOI: 10.1021/acs.jctc.0c00972.

(26) Ross, G. A.; Bodnarchuk, M. S.; Essex, J. W. Water sites, networks, and free energies with grand canonical Monte Carlo. *J. Am. Chem. Soc.* **2015**, *137* (47), 14930–14943, DOI: 10.1021/jacs.5b07940.

(27) Ross, G. A.; Bruce Macdonald, H. E.; Cave-Ayland, C.; Cabedo Martinez, A. I.; Essex, J. W. Replica-exchange and standard state binding free energies with grand canonical Monte Carlo. *J. Chem. Theory Comput.* **2017**, *13* (12), 6373–6381, DOI: 10.1021/acs.jctc.7b00738.

(28) Chen, J. M.; Xu, S. L.; Wawrzak, Z.; Basarab, G. S.; Jordan, D. B. Structure-based design of potent inhibitors of scytalone dehydratase: displacement of a water molecule from the active site. *Biochemistry* **1998**, *37* (51), 17735–17744, DOI: 10.1021/bi981848r.

(29) Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. Energetics of displacing water molecules from protein binding sites: consequences for ligand optimization. *J. Am. Chem. Soc.* **2009**, *131* (42), 15403–15411, DOI: 10.1021/ja906058w.

(30) Bruce Macdonald, H. E.; Cave-Ayland, C.; Ross, G. A.; Essex, J. W. Ligand binding free energies with adaptive water networks: two-dimensional grand canonical alchemical perturbations. *J. Chem. Theory Comput.* **2018**, *14* (12), 6586–6597, DOI: 10.1021/acs.jctc.8b00614.

(31) Kim, M.-S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; Thomas, J. K.; Muthusamy, B.; Leal-Rojas, P.; Kumar, P.; Sahasrabuddhe, N. A.; Balakrishnan, L.; Advani, J.; George, B.; Renuse, S.; Selvan, L. D. N.; Patil, A. H.; Nanjappa, V.; Radhakrishnan, A.; Prasad, S.; Subbannayya, T.; Raju, R.; Kumar, M.; Sreenivasamurthy, S. K.; Marimuthu, A.; Sathe, G. J.; Chavan, S.; Datta, K. K.; Subbannayya, Y.; Sahu, A.; Yelamanchi, S. D.; Jayaram, S.; Rajagopalan, P.; Sharma, J.; Murthy, K. R.; Syed, N.; Goel, R.; Khan, A. A.; Ahmad, S.; Dey, G.; Mudgal, K.; Chatterjee, A.; Huang, T.-C.; Zhong, J.; Wu, X.; Shaw, P. G.; Freed, D.; Zahari, M. S.; Mukherjee, K. K.; Shankar, S.; Mahadevan, A.; Lam, H.; Mitchell, C. J.; Shankar, S. K.; Satishchandra, P.; Schroeder, J. T.; Sirdeshmukh, R.; Maitra, A.; Leach, S. D.; Drake, C. G.; Halushka, M. K.; Prasad, T. S. K.; Hruban, R. H.; Kerr, C. L.; Bader, G. D.; Iacobuzio-Donahue, C. A.; Gowda, H.; Pandey, A. A draft map of the human proteome. *Nature* **2014**, *509* (7502), 575–581, DOI: 10.1038/nature13302.

(32) Chen, Y.; Cong, Y.; Quan, B.; Lan, T.; Chu, X.; Ye, Z.; Hou, X.; Wang, C. Chemoproteomic profiling of targets of lipid-derived electrophiles by bioorthogonal aminooxy probe. *Redox Biol.* **2017**, *12*, 712–718, DOI: 10.1016/j.redox.2017.04.001.

(33) Chen, Y.; Liu, Y.; Lan, T.; Qin, W.; Zhu, Y.; Qin, K.; Gao, J.; Wang, H.; Hou, X.; Chen, N.; Friedmann Angeli, J. P.; Conrad, M.; Wang, C. Quantitative profiling of protein carbonylations in ferroptosis by an aniline-derived probe. *J. Am. Chem. Soc.* **2018**, *140* (13), 4712–4720, DOI: 10.1021/jacs.8b01462.

(34) Qin, W.; Lv, P.; Fan, X.; Quan, B.; Zhu, Y.; Qin, K.; Chen, Y.; Wang, C.; Chen, X. Quantitative time-resolved chemoproteomics reveals that stable O-GlcNAc regulates box C/D snoRNP biogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (33), E6749, DOI: 10.1073/pnas.1702688114.

(35) Qin, W.; Qin, K.; Fan, X.; Peng, L.; Hong, W.; Zhu, Y.; Lv, P.; Du, Y.; Huang, R.; Han, M.; Cheng, B.; Liu, Y.; Zhou, W.; Wang, C.; Chen, X. Artificial cysteine S-glycosylation induced by per-O-acetylated unnatural monosaccharides during metabolic glycan labeling. *Angew. Chem., Int. Ed.* **2018**, *57* (7), 2006, DOI: 10.1002/anie.201800116.

(36) Qin, W.; Qin, K.; Fan, X.; Peng, L.; Hong, W.; Zhu, Y.; Lv, P.; Du, Y.; Huang, R.; Han, M.; Cheng, B.; Liu, Y.; Zhou, W.; Wang, C.; Chen, X. Artificial cysteine S-glycosylation induced by per-O-acetylated unnatural monosaccharides during metabolic glycan labeling. *Angew. Chem., Int. Ed.* **2018**, *57* (7), 1817–1820, DOI: 10.1002/anie.201711710.

(37) Chen, N.; Liu, J.; Qiao, Z.; Liu, Y.; Yang, Y.; Jiang, C.; Wang, X.; Wang, C. Chemical proteomic profiling of protein N-homocysteinylation with a thioester probe. *Chem. Sci.* **2018**, *9* (10), 2826–2830, DOI: 10.1039/c8sc00221e.

(38) Chen, N.; Qiao, Z.; Wang, C. A chemoselective reaction between protein N-homocysteinylation and azides catalyzed by heme(II). *Chem. Commun.* **2019**, *55* (25), 3654–3657, DOI: 10.1039/c9cc00055k.

(39) Qin, W.; Qin, K.; Zhang, Y.; Jia, W.; Chen, Y.; Cheng, B.; Peng, L.; Chen, N.; Liu, Y.; Zhou, W.; Wang, Y.-L.; Chen, X.; Wang, C. S-glycosylation-based cysteine profiling reveals regulation of glycolysis by itaconate. *Nat. Chem. Biol.* **2019**, *15* (10), 983–991, DOI: 10.1038/s41589-019-0323-5.

(40) Qin, W.; Zhang, Y.; Tang, H.; Liu, D.; Chen, Y.; Liu, Y.; Wang, C. Chemoproteomic profiling of itaconation by bioorthogonal probes in inflammatory macrophages. *J. Am. Chem. Soc.* **2020**, *142* (25), 10894–10898, DOI: 10.1021/jacs.9b11962.

(41) Zhang, Y.; Qin, W.; Liu, D.; Liu, Y.; Wang, C. Chemoproteomic profiling of itaconations in *Salmonella*. *Chem. Sci.* **2021**, *12* (17), 6059–6063, DOI: 10.1039/d1sc00660f.

(42) Dai, J.; Liang, K.; Zhao, S.; Jia, W.; Liu, Y.; Wu, H.; Lv, J.; Cao, C.; Chen, T.; Zhuang, S.; Hou, X.; Zhou, S.; Zhang, X.; Chen, X.-W.; Huang, Y.; Xiao, R.-P.; Wang, Y.-L.; Luo, T.; Xiao, J.; Wang, C. Chemoproteomics reveals baicalin activates hepatic CPT1 to ameliorate diet-induced obesity and hepatic steatosis. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (26), E5896–E5905, DOI: 10.1073/pnas.1801745115.

(43) Zhuang, S.; Li, Q.; Cai, L.; Wang, C.; Lei, X. Chemoproteomic profiling of bile acid-interacting proteins. *ACS Cent. Sci.* **2017**, *3* (5), 501–509, DOI: 10.1021/acscentsci.7b00134.

(44) Ye, Z.; Zhang, X.; Zhu, Y.; Song, T.; Chen, X.; Lei, X.; Wang, C. Chemoproteomic profiling reveals ethacrynic acid targets adenine nucleotide translocases to impair mitochondrial function. *Mol. Pharmaceutics* **2018**, *15* (6), 2413–2422, DOI: 10.1021/acs.molpharmaceut.8b00250.

(45) Gao, J.; Yang, F.; Che, J.; Han, Y.; Wang, Y.; Chen, N.; Bak, D. W.; Lai, S.; Xie, X.; Weerapana, E.; Wang, C. Selenium-encoded isotopic signature targeted profiling. *ACS Cent. Sci.* **2018**, *4* (8), 960–970, DOI: 10.1021/acscentsci.8b00112.

(46) Wang, J.; Liu, Y.; Liu, Y.; Zheng, S.; Wang, X.; Zhao, J.; Yang, F.; Zhang, G.; Wang, C.; Chen, P. R. Time-resolved protein activation by proximal decaging in living systems. *Nature* **2019**, *569* (7757), 509–513, DOI: 10.1038/s41586-019-1188-1.

(47) Weerapana, E.; Wang, C.; Simon, G. M.; Richter, F.; Khare, S.; Dillon, M. B. D.; Bachovchin, D. A.; Mowen, K.; Baker, D.; Cravatt, B. F. Quantitative reactivity profiling predicts functional cysteines in proteomes. *Nature* **2010**, *468* (7325), 790–795, DOI: 10.1038/nature09472.

(48) Wang, C.; Weerapana, E.; Blewett, M. M.; Cravatt, B. F. A chemoproteomic platform to quantitatively map targets of lipid-derived electrophiles. *Nat. Methods* **2014**, *11* (1), 79–85, DOI: 10.1038/nmeth.2759.

(49) Wang, H.; Chen, X.; Li, C.; Liu, Y.; Yang, F.; Wang, C. Sequence-based prediction of cysteine reactivity using machine learning. *Biochemistry* **2018**, *57* (4), 451–460, DOI: 10.1021/acs.biochem.7b00897.

(50) Qiu, X.; Janson, C. A.; Konstantinidis, A. K.; Nwagwu, S.; Silverman, C.; Smith, W. W.; Khandekar, S.; Lonsdale, J.; Abdel-Meguid, S. S. Crystal structure of $\beta$-ketoacyl-acyl carrier protein synthase III. A key condensing enzyme in bacterial fatty acid biosynthesis. *J. Biol. Chem.* **1999**, *274* (51), 36465–36471, DOI: 10.1074/jbc.274.51.36465.

(51) Campbell, J. W.; Morgan-Kiss, R. M.; Cronan John E., J. A new *Escherichia coli* metabolic competency: growth on fatty acids by a novel anaerobic $\beta$-oxidation pathway. *Mol. Microbiol.* **2003**, *47* (3), 793–805, DOI: 10.1046/j.1365-2958.2003.03341.x.

(52) Balakrishnan, S.; Kamisetty, H.; Carbonell, J. G.; Lee, S.-I.; Langmead, C. J. Learning generative models for protein fold families. *Proteins: Struct., Funct., Bioinf.* **2011**, *79* (4), 1061–1078, DOI: 10.1002/prot.22934.

(53) Kamisetty, H.; Ovchinnikov, S.; Baker, D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (39), 15674–15679, DOI: 10.1073/pnas.1314045110.

(54) Kamisetty, H.; Ovchinnikov, S.; Baker, D. Correction for assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (46), 18734, DOI: 10.1073/pnas.1319550110.

(55) Ovchinnikov, S.; Park, H.; Varghese, N.; Huang, P.-S.; Pavlopoulos, G. A.; Kim, D. E.; Kamisetty, H.; Kyrpides, N. C.; Baker, D. Protein structure determination using metagenome sequence data. *Science* **2017**, *355* (6322), 294–298, DOI: 10.1126/science.aah4043.

(56) Cong, Q.; Anishchenko, I.; Ovchinnikov, S.; Baker, D. Protein interaction networks revealed by proteome coevolution. *Science* **2019**, *365* (6449), 185–189, DOI: 10.1126/science.aaw6718.

(57) Chakrabarti, S.; Panchenko, A. R. Structural and functional roles of coevolved sites in proteins. *PLoS One* **2010**, *5* (1), e8591, DOI: 10.1371/journal.pone.0008591.

(58) Fica, S. M.; Oubridge, C.; Wilkinson, M. E.; Newman, A. J.; Nagai, K. A human postcatalytic spliceosome structure reveals essential roles of metazoan factors for exon ligation. *Science* **2019**, *363* (6428), 710–714, DOI: 10.1126/science.aaw5569.

(59) Zhang, X.; Zhan, X.; Yan, C.; Zhang, W.; Liu, D.; Lei, J.; Shi, Y. Structures of the human spliceosomes before and after release of the ligated exon. *Cell Res.* **2019**, *29* (4), 274–285, DOI: 10.1038/s41422-019-0143-x.

(60) Amaro, R. E.; Mulholland, A. J. Biomolecular simulations in the time of COVID-19, and after. *Comput. Sci. Eng.* **2020**, *22* (6), 30–36, DOI: 10.1109/mcse.2020.3024155.

(61) Toelzer, C.; Gupta, K.; Yadav, S. K. N.; Borucu, U.; Davidson, A. D.; Kavanagh Williamson, M.; Shoemark, D. K.; Garzoni, F.; Staufer, O.; Milligan, R.; Capin, J.; Mulholland, A. J.; Spatz, J.; Fitzgerald, D.; Berger, I.; Schaffitzel, C. Free fatty acid binding pocket in the locked structure of SARS-CoV-2 spike protein. *Science* **2020**, *370* (6517), 725–730, DOI: 10.1126/science.abd3255.

(62) Shoemark, D. K.; Colenso, C. K.; Toelzer, C.; Gupta, K.; Sessions, R. B.; Davidson, A. D.; Berger, I.; Schaffitzel, C.; Spencer, J.; Mulholland, A. J. Molecular simulations suggest vitamins, retinoids and steroids as ligands of the free fatty acid pocket of the SARS-CoV-2 spike protein. *Angew. Chem., Int. Ed.* **2021**, *60* (13), 7098–7110, DOI: 10.1002/anie.202015639.

(63) Oliveira, A. S. F.; Ibarra, A. A.; Bermudez, I.; Casalino, L.; Gaieb, Z.; Shoemark, D. K.; Gallagher, T.; Sessions, R. B.; Amaro, R. E.; Mulholland, A. J. A potential interaction between the SARS-CoV-2 spike protein and nicotinic acetylcholine receptors. *Biophys. J.* **2021**, *120* (6), 983–993, DOI: 10.1016/j.bpj.2021.01.037.

(64) Jin, Z.; Du, X.; Xu, Y.; Deng, Y.; Liu, M.; Zhao, Y.; Zhang, B.; Li, X.; Zhang, L.; Peng, C.; Duan, Y.; Yu, J.; Wang, L.; Yang, K.; Liu, F.; Jiang, R.; Yang, X.; You, T.; Liu, X.; Yang, X.; Bai, F.; Liu, H.; Liu, X.; Guddat, L. W.; Xu, W.; Xiao, G.; Qin, C.; Shi, Z.; Jiang, H.; Rao, Z.; Yang, H. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* **2020**, *582* (7811), 289–293, DOI: 10.1038/s41586-020-2223-y.

(65) Kupferschmidt, K.; Cohen, J. Race to find COVID-19 treatments accelerates. *Science* **2020**, *367* (6485), 1412–1413, DOI: 10.1126/science.367.6485.1412.

(66) Arafet, K.; Serrano-Aparicio, N.; Lodola, A.; Mulholland, A. J.; Gonzalez, F. V.; Swiderek, K.; Moliner, V. Mechanism of inhibition of SARS-CoV-2 Mpro by N3 peptidyl Michael acceptor explained by QM/MM simulations and design of new derivatives with tunable chemical reactivity. *Chem. Sci.* **2021**, *12* (4), 1433–1444, DOI: 10.1039/d0sc06195f.

(67) Amaro, R. E.; Mulholland, A. J. A community letter regarding sharing biomolecular simulation data for COVID-19. *J. Chem. Inf. Model.* **2020**, *60* (6), 2653–2656, DOI: 10.1021/acs.jcim.0c00319.

(68) Amaro, R. E.; Mulholland, A. J. Multiscale methods in drug design bridge chemical and biological complexity in the search for cures. *Nat. Rev. Chem.* **2018**, *2* (4), 0148, DOI: 10.1038/s41570-018-0148.

(69) Warshel, A.; Karplus, M. Calculation of ground and excited state potential surfaces of conjugated molecules. I. Formulation and parametrization. *J. Am. Chem. Soc.* **1972**, *94* (16), 5612–5625, DOI: 10.1021/ja00771a014.

(70) Warshel, A.; Levitt, M. Folding and stability of helical proteins: carp myogen. *J. Mol. Biol.* **1976**, *106* (2), 421–437, DOI: 10.1016/0022-2836(76)90094-2.

(71) Field, M. J.; Bash, P. A.; Karplus, M. A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *J. Comput. Chem.* **1990**, *11* (6), 700–733, DOI: 10.1002/jcc.540110605.

(72) Senn, H. M.; Thiel, W. QM/MM methods for biomolecular systems. *Angew. Chem., Int. Ed.* **2009**, *48* (7), 1198–1229, DOI: 10.1002/anie.200802019.

(73) Van der Kamp, M. W.; Mulholland, A. J. Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology. *Biochemistry* **2013**, *52* (16), 2708–2728, DOI: 10.1021/bi400215w.

(74) Claeyssens, F.; Harvey, J. N.; Manby, F. R.; Mata, R. A.; Mulholland, A. J.; Ranaghan, K. E.; Schutz, M.; Thiel, S.; Thiel, W.; Werner, H.-J. High-accuracy computation of reaction barriers in enzymes. *Angew. Chem., Int. Ed.* **2006**, *45* (41), 6856–6859, DOI: 10.1002/anie.200602711.

(75) Bennie, S. J.; van der Kamp, M. W.; Pennifold, R. C. R.; Stella, M.; Manby, F. R.; Mulholland, A. J. A projector-embedding approach for multiscale coupled-cluster calculations applied to citrate synthase. *J. Chem. Theory Comput.* **2016**, *12* (6), 2689–2697, DOI: 10.1021/acs.jctc.6b00285.

(76) Zhang, X.; Bennie, S. J.; van der Kamp, M. W.; Glowacki, D. R.; Manby, F. R.; Mulholland, A. J. Multiscale analysis of enantioselectivity in enzyme-catalysed 'lethal synthesis' using projector-based embedding. *R. Soc. Open Sci.* **2018**, *5* (2), 171390, DOI: 10.1098/rsos.171390.

(77) Ranaghan, K. E.; Shchepanovska, D.; Bennie, S. J.; Lawan, N.; Macrae, S. J.; Zurek, J.; Manby, F. R.; Mulholland, A. J. Projector-based embedding eliminates density functional dependence for QM/MM calculations of reactions in enzymes and solution. *J. Chem. Inf. Model.* **2019**, *59* (5), 2063–2078, DOI: 10.1021/acs.jcim.8b00940.

(78) Chudyk, E. I.; Limb, M. A. L.; Jones, C.; Spencer, J.; van der Kamp, M. W.; Mulholland, A. J. QM/MM simulations as an assay for carbapenemase activity in class A $\beta$-lactamases. *Chem. Commun.* **2014**, *50* (94), 14736–14739, DOI: 10.1039/c4cc06495j.

(79) Hirvonen, V. H. A.; Hammond, K.; Chudyk, E. I.; Limb, M. A. L.; Spencer, J.; Mulholland, A. J.; van der Kamp, M. W. An efficient computational assay for $\beta$-lactam antibiotic breakdown by class A $\beta$-lactamases. *J. Chem. Inf. Model.* **2019**, *59* (8), 3365–3369, DOI: 10.1021/acs.jcim.9b00442.

(80) Hirvonen, V. H. A.; Mulholland, A. J.; Spencer, J.; van der Kamp, M. W. Small changes in hydration determine cephalosporinase activity of OXA-48 $\beta$-lactamases. *ACS Catal.* **2020**, *10* (11), 6188–6196, DOI: 10.1021/acscatal.0c00596.

(81) Oliveira, A. S. F.; Edsall, C. J.; Woods, C. J.; Bates, P.; Nunez, G. V.; Wonnacott, S.; Bermudez, I.; Ciccotti, G.; Gallagher, T.; Sessions, R. B.; Mulholland, A. J. A general mechanism for signal propagation in the nicotinic acetylcholine receptor family. *J. Am. Chem. Soc.* **2019**, *141* (51), 19953–19958, DOI: 10.1021/jacs.9b09055.

(82) Galdadas, I.; Qu, S.; Oliveira, A. S. F.; Olehnovics, E.; Mack, A. R.; Mojica, M. F.; Agarwal, P. K.; Tooke, C. L.; Gervasio, F. L.; Spencer, J.; Bonomo, R. A.; Mulholland, A. J.; Haider, S. Allosteric communication in class A $\beta$-lactamases occurs *via* cooperative coupling of loop dynamics. *Elife* **2021**, *10*, e66567, DOI: 10.7554/eLife.66567.

(83) Arcus, V. L.; Prentice, E. J.; Hobbs, J. K.; Mulholland, A. J.; Van der Kamp, M. W.; Pudney, C. R.; Parker, E. J.; Schipper, L. A. On the temperature dependence of enzyme-catalyzed rates. *Biochemistry* **2016**, *55* (12), 1681–1688, DOI: 10.1021/acs.biochem.5b01094.

(84) Arcus, V. L.; van der Kamp, M. W.; Pudney, C. R.; Mulholland, A. J. Enzyme evolution and the temperature dependence of enzyme catalysis. *Curr. Opin. Struct. Biol.* **2020**, *65*, 96–101, DOI: 10.1016/j.sbi.2020.06.001.

(85) Hobbs, J. K.; Jiao, W.; Easter, A. D.; Parker, E. J.; Schipper, L. A.; Arcus, V. L. Change in heat capacity for enzyme catalysis determines temperature dependence of enzyme catalyzed rates. *ACS Chem. Biol.* **2013**, *8* (11), 2388–2393, DOI: 10.1021/cb4005029.

(86) Arcus, V. L.; Mulholland, A. J. Temperature, dynamics, and enzyme-catalyzed reaction rates. *Annu. Rev. Biophys.* **2020**, *49*, 163–180, DOI: 10.1146/annurev-biophys-121219-081520.

(87) Van der Kamp, M. W.; Prentice, E. J.; Kraakman, K. L.; Connolly, M.; Mulholland, A. J.; Arcus, V. L. Dynamical origins of heat capacity changes in enzyme-catalysed reactions. *Nat. Commun.* **2018**, *9* (1), 1–7, DOI: 10.1038/s41467-018-03597-y.

(88) Bunzel, H. A.; Kries, H.; Marchetti, L.; Zeymer, C.; Mittl, P. R. E.; Mulholland, A. J.; Hilvert, D. Emergence of a negative activation heat capacity during evolution of a designed enzyme. *J. Am. Chem. Soc.* **2019**, *141* (30), 11745–11748, DOI: 10.1021/jacs.9b02731.

(89) Privett, H. K.; Kiss, G.; Lee, T. M.; Blomberg, R.; Chica, R. A.; Thomas, L. M.; Hilvert, D.; Houk, K. N.; Mayo, S. L. Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (10), 3790–3795, DOI: 10.1073/pnas.1118082108.

(90) Bunzel, H. A.; Anderson, J.; Hilvert, D.; Arcus, V. L.; van der Kamp, M. W.; Mulholland, A. J. Evolution of dynamical networks enhances catalysis in a designer enzyme. *Nat. Chem.* **2021**, *13* (10), 1017–1022, DOI: 0.1038/s41557-021-00763-6.

(91) O'Connor, M.; Deeks, H. M.; Dawn, E.; Metatla, O.; Roudaut, A.; Sutton, M.; Thomas, L. M.; Glowacki, B. R.; Sage, R.; Tew, P.; Wonnacott, M.; Bates, P.; Mulholland, A. J.; Glowacki, D. R. Sampling molecular conformations and dynamics in a multiuser virtual reality framework. *Sci. Adv.* **2018**, *4* (6), eaat2731, DOI: 10.1126/sciadv.aat2731.

(92) O'Connor, M. B.; Bennie, S. J.; Deeks, H. M.; Jamieson-Binnie, A.; Jones, A. J.; Shannon, R. J.; Walters, R.; Mitchell, T. J.; Mulholland, A. J.; Glowacki, D. R. Interactive molecular dynamics in virtual reality from quantum chemistry to drug binding: an open-source multi-person framework. *J. Chem. Phys.* **2019**, *150* (22), 220901, DOI: 10.1063/1.5092590.

(93) Deeks, H. M.; Walters, R. K.; Hare, S. R.; O'Connor, M. B.; Mulholland, A. J.; Glowacki, D. R. Interactive molecular dynamics in virtual reality for accurate flexible protein-ligand docking. *PLoS One* **2020**, *15* (3), e0228461, DOI: 10.1371/journal.pone.0228461.

(94) Deeks, H. M.; Walters, R. K.; Barnoud, J.; Glowacki, D. R.; Mulholland, A. J. Interactive molecular dynamics in virtual reality is an effective tool for flexible substrate and inhibitor docking to the SARS-CoV-2 main protease. *J. Chem. Inf. Model.* **2020**, *60* (12), 5803–5814, DOI: 10.1021/acs.jcim.0c01030.

(95) Bennie, S. J.; Ranaghan, K. E.; Deeks, H.; Goldsmith, H. E.; O'Connor, M. B.; Mulholland, A. J.; Glowacki, D. R. Teaching enzyme catalysis using interactive molecular dynamics in virtual reality. *J. Chem. Educ.* **2019**, *96* (11), 2488–2496, DOI: 10.1021/acs.jchemed.9b00181.

(96) Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **1973**, *181* (4096), 223–230, DOI: 10.1126/science.181.4096.223.

(97) Moult, J.; James, M. N. G. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins: Struct., Funct., Genet.* **1986**, *1* (2), 146–163, DOI: 10.1002/prot.340010207.

(98) Göbel, U.; Sander, C.; Schneider, R.; Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins: Struct., Funct., Genet.* **1994**, *18* (4), 309–317, DOI: https://doi.org/10.1002/prot.340180402.

(99) Burger, L.; van Nimwegen, E. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput. Biol.* **2010**, *6* (1), e1000633, DOI: 10.1371/journal.pcbi.1000633.

(100) Marks, D. S.; Hopf, T. A.; Sander, C. Protein structure prediction from sequence variation. *Nat. Biotechnol.* **2012**, *30* (11), 1072–1080, DOI: 10.1038/nbt.2419.

(101) Kryshtafovych, A.; Barbato, A.; Monastyrskyy, B.; Fidelis, K.; Schwede, T.; Tramontano, A. Methods of model accuracy estimation can help selecting the best models from decoy sets: assessment of model accuracy estimations in CASP11. *Proteins: Struct., Funct., Bioinf.* **2016**, *84 (Suppl 1)*, 349–369, DOI: 10.1002/prot.24919.

(102) Shrestha, R.; Fajardo, E.; Gil, N.; Fidelis, K.; Kryshtafovych, A.; Monastyrskyy, B.; Fiser, A. Assessing the accuracy of contact predictions in CASP13. *Proteins: Struct., Funct., Bioinf.* **2019**, *87* (12), 1058–1068, DOI: 10.1002/prot.25819.

(103) AlQuraishi, M. AlphaFold at CASP13. *Bioinformatics* **2019**, *35* (22), 4862–4865, DOI: 10.1093/bioinformatics/btz422.

(104) Callaway, E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature* **2020**, *588* (7837), 203–204, DOI: 10.1038/d41586-020-03348-4.

(105) Simpkin, A. J.; Sanchez Rodriguez, F.; Mesdaghi, S.; Kryshtafovych, A.; Rigden, D. J. Evaluation of model refinement in CASP14. *Proteins: Structure, Function, and Bioinformatics* **2021**, 1852–1869, DOI: 10.1002/prot.26185.

(106) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589, DOI: 10.1038/s41586-021-03819-2.

(107) Hiranuma, N.; Park, H.; Baek, M.; Anishchenko, I.; Dauparas, J.; Baker, D. Improved protein structure refinement guided by deep learning based accuracy estimation. *Nat. Commun.* **2021**, *12* (1), 1340, DOI: 10.1038/s41467-021-21511-x.

(108) Egbert, M.; Ghani, U.; Ashizawa, R.; Kotelnikov, S.; Nguyen, T.; Desta, I.; Hashemi, N.; Padhorny, D.; Kozakov, D.; Vajda, S. Assessing the binding properties of CASP14 targets and models. *Proteins: Structure, Function, and Bioinformatics* **2021**, 1922–1939, DOI: 10.1002/prot.26209.

(109) Ackloo, S.; Al-awar, R.; Amaro, R. E.; Arrowsmith, C. H.; Azevedo, H.; Batey, R. A.; Bengio, Y.; Betz, U. A. K.; Bologa, C. G.; Chodera, J. D.; Cornell, W. D.; Dunham, I.; Ecker, G. F.; Edfeldt, K.; Edwards, A. M. G.; K, M.; Gordijo, C. R.; Hessler, G.; Hillisch, A.; Hogner, A.; Irwin, J. J.; Jansen, J. M.; Kuhn, D.; Leach, A. R.; Lee, A. A.; Lessel, U.; Moult, J.; Muegge, I.; Oprea, T. I.; Perry, B. G.; Riley, P.; Saikatendu, K. S.; Santhakumar, V.; Schapira, M.; Scholten, C.; Todd, M. H.; Vedadi, M.; Volkamer, A.; Willson, T. M. CACHE (Critical assessment of computational hit-finding experiments): a public-private partnership benchmarking initiative to enable the development of computational methods for hit-finding. *ChemRxiv* **2021**, http://chemrxiv.org/engage/chemrxiv/article-details/6168ba62f718dfc39bdee0db (accessed October 26, 2021).

(110) Miller, E. B.; Murphy, R. B.; Sindhikara, D.; Borrelli, K. W.; Grisewood, M. J.; Ranalli, F.; Dixon, S. L.; Jerome, S.; Boyles, N. A.; Day, T.; Ghanakota, P.; Mondal, S.; Rafi, S. B.; Troast, D. M.; Abel, R.; Friesner, R. A. Reliable and accurate solution to the induced fit docking problem for protein–ligand binding. *J. Chem. Theory Comput.* **2021**, *17* (4), 2630–2639, DOI: 10.1021/acs.jctc.1c00136.

(111) *Coarse-grained Modeling of Biomolecules*; Papoian, G. A., Ed.; CRC Press: Boca Raton, FL, 2018.

(112) Zwanzig, R.; Szabo, A.; Bagchi, B. Levinthal's paradox. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89* (1), 20–22, DOI: 10.1073/pnas.89.1.20.

(113) Karplus, M. The Levinthal paradox: yesterday and today. *Folding Des.* **1997**, *2* (4), S69–S75, DOI: 10.1016/s1359-0278(97)00067-9.

(114) Plaxco, K. W.; Simons, K. T.; Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **1998**, *277* (4), 985–994, DOI: 10.1006/jmbi.1998.1645.

(115) Plaxco, K. W.; Simons, K. T.; Ruczinski, I.; Baker, D. Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics. *Biochemistry* **2000**, *39* (37), 11177–11183, DOI: 10.1021/bi000200n.

(116) Hopf, T. A.; Green, A. G.; Schubert, B.; Mersmann, S.; Scharfe, C. P. I.; Ingraham, J. B.; Toth-Petroczy, A.; Brock, K.; Riesselman, A. J.; Palmedo, P.; Kang, C.; Sheridan, R.; Draizen, E. J.; Dallago, C.; Sander, C.; Marks, D. S. The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* **2019**, *35* (9), 1582–1584, DOI: 10.1093/bioinformatics/bty862.

(117) Yang, J.; Anishchenko, I.; Park, H.; Peng, Z.; Ovchinnikov, S.; Baker, D. Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117* (3), 1496–1503, DOI: 10.1073/pnas.1914677117.

(118) Zhang, C.; Zheng, W.; Mortuza, S. M.; Li, Y.; Zhang, Y. DeepMSA: constructing deep multiple sequencealignment to improve contact prediction andfold-recognition for distant-homology proteins. *Bioinformatics* **2020**, *36* (7), 2105–2112, DOI: 10.1093/bioinformatics/btz863.

(119) Davtyan, A.; Schafer, N. P.; Zheng, W.; Clementi, C.; Wolynes, P. G.; Papoian, G. A. AWSEM-MD: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J. Phys. Chem. B* **2012**, *116* (29), 8494–8503, DOI: 10.1021/jp212541y.

(120) Feig, M. Local protein structure refinement via molecular dynamics simulations with locPREFMD. *J. Chem. Inf. Model.* **2016**, *56* (7), 1304–1312, DOI: 10.1021/acs.jcim.6b00222.

(121) Billings, W. M.; Hedelius, B.; Millecam, T.; Wingate, D.; Della Corte, D. ProSPr: democratized implementation of alphafold protein distance prediction network. *bioRxiv* **2019**, 830273, http://www.biorxiv.org/content/10.1101/830273v2 (accessed October 4, 2021).

(122) Adiyaman, R.; McGuffin, L. J. Methods for the refinement of protein structure 3D models. *Int. J. Mol. Sci.* **2019**, *20* (9), 2301, DOI: 10.3390/ijms20092301.

(123) Yang, X.-S., Firefly algorithm, Lévy flights and global optimization. In *Research and Development in Intelligent Systems XXVI*, Bramer, M., Ellis, R., Petridis, M., Eds.; Springer: London, 2010, pp 209–218.

(124) Yadav, A.; Vishwakarma, D. K. A comparative study on bio-inspired algorithms for sentiment analysis. *Cluster Comput.* **2020**, *23* (4), 2969–2989, DOI: 10.1007/s10586-020-03062-w, https://doi.org/10.1007/s10586-020-03062-w.

(125) Muntau, A. C.; Leandro, J.; Staudigl, M.; Mayer, F.; Gersting, S. W. Innovative strategies to treat protein misfolding in inborn errors of metabolism: pharmacological chaperones and proteostasis regulators. *J. Inherited Metab. Dis.* **2014**, *37* (4), 505–523, DOI: 10.1007/s10545-014-9701-z.

(126) Ackley, D. H. *A Connectionist Machine for Genetic Hillclimbing*; Kluwer Academic Publishers: Amsterdam, 1987.

(127) Rubiera, C. O.; Deane, C.; Nissley, D. A. Current protein structure predictors do not produce meaningful folding pathways. *bioRxiv* **2021**, 461137, http://www.biorxiv.org/content/10.1101/2021.09.20.461137v1.full.pdf (accessed October 6, 2021).

(128) Schaap, M. G.; Leij, F. J.; van Genuchten, M. T. Rosetta: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. *J. Hydrol. (Amsterdam, Neth.)* **2001**, *251* (3), 163–176, DOI: 10.1016/S0022-1694(01)00466-8.

(129) De Oliveira, S. H. P.; Law, E. C.; Shi, J.; Deane, C. M. Sequential search leads to faster, more efficient fragment-based *de novo* protein structure prediction. *Bioinformatics* **2018**, *34* (7), 1132–1140, DOI: 10.1093/bioinformatics/btx722.

(130) Marks, D. S.; Colwell, L. J.; Sheridan, R.; Hopf, T. A.; Pagnani, A.; Zecchina, R.; Sander, C. Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **2011**, *6* (12), e28766, DOI: 10.1371/journal.pone.0028766.

(131) Greener, J. G.; Kandathil, S. M.; Jones, D. T. Deep learning extends *de novo* protein modelling coverage of genomes using iteratively predicted structural constraints. *Nat. Commun.* **2019**, *10* (1), 1–13, DOI: 10.1038/s41467-019-11994-0.

(132) Kallberg, M.; Wang, h.; Wang, S.; Peng, J.; Wang, Z.; Lu, H.; Xu, J. Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.* **2012**, *7* (8), 1511–1522, DOI: 10.1038/nprot.2012.085.

(133) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millan, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373* (6557), 871–876, DOI: 10.1126/science.abj8754.

(134) Manavalan, B.; Kuwajima, K.; Lee, J. PFDB: a standardized protein folding database with temperature correction. *Sci. Rep.* **2019**, *9* (1), 1588, DOI: 10.1038/s41598-018-36992-y.

(135) Pancsa, R.; Varadi, M.; Tompa, P.; Vranken, W. F. Start2Fold: a database of hydrogen/deuterium exchange data on protein folding and stability. *Nucleic Acids Res.* **2016**, *44* (D1), D429–D434, DOI: 10.1093/nar/gkv1185.

(136) Frishman, D.; Argos, P. Knowledge-based protein secondary structure assignment. *Proteins: Struct., Funct., Bioinf.* **1995**, *23* (4), 566–579, DOI: https://doi.org/10.1002/prot.340230412.

(137) Congreve, M.; de Graaf, C.; Swain, N. A.; Tate, C. G. Impact of GPCR structures on drug discovery. *Cell* **2020**, *181* (1), 81–91, DOI: 10.1016/j.cell.2020.03.003.

(138) Congreve, M.; Andrews, S. P.; Dore, A. S.; Hollenstein, K.; Hurrell, E.; Langmead, C. J.; Mason, J. S.; Ng, I. W.; Tehan, B.; Zhukov, A.; Weir, M.; Marshall, F. H. Discovery of 1,2,4-triazine derivatives as adenosine A2A antagonists using structure based drug design. *J. Med. Chem.* **2012**, *55* (5), 1898–1903, DOI: 10.1021/jm201376w.

(139) Borodovsky, A.; Barbon, C. M.; Wang, Y.; Ye, M.; Prickett, L.; Chandra, D.; Shaw, J.; Deng, N.; Sachsenmeier, K.; Clarke, J. D.; Linghu, B.; Brown, G. A.; Brown, J.; Congreve, M.; Cheng, R. K.; Dore, A. S.; Hurrell, E.; Shao, W.; Woessner, R.; Reimer, C.; Drew, L.; Fawell, S.; Schuller, A. G.; Mele, D. A. Small molecule AZD4635 inhibitor of A(2A)R signaling rescues immune cell function including CD103(+) dendritic cells enhancing anti-tumor immunity. *J. Immunother. Cancer* **2020**, *8* (2), e000417, DOI: 10.1136/jitc-2019-000417.

(140) Vass, M.; Kooistra, A. J.; Yang, D.; Stevens, R. C.; Wang, M. W.; de Graaf, C. Chemical diversity in the G protein-coupled receptor superfamily. *Trends Pharmacol. Sci.* **2018**, *39* (5), 494–512, DOI: 10.1016/j.tips.2018.02.004.

(141) Congreve, M.; Oswald, C.; Marshall, F. H. Applying structure-based drug design approaches to allosteric modulators of GPCRs. *Trends Pharmacol. Sci.* **2017**, *38* (9), 837–847, DOI: 10.1016/j.tips.2017.05.010.

(142) Kufareva, I.; Rueda, M.; Katritch, V.; Stevens, R. C.; Abagyan, R. Status of GPCR modeling and docking as reflected by community-wide GPCR dock 2010 assessment. *Structure (Cambridge, MA, U. S.)* **2011**, *19* (8), 1108–1126, DOI: 10.1016/j.str.2011.05.012.

(143) Wong, R. S. Y.; Bodart, V.; Metz, M.; Labrecque, J.; Bridger, G.; Fricker, S. P. Comparison of the potential multiple binding modes of bicyclam, monocylam, and noncyclam small-molecule CXC chemokine receptor 4 inhibitors. *Mol. Pharmacol.* **2008**, *74* (6), 1485–1495, DOI: 10.1124/mol.108.049775.

(144) Kufareva, I.; Katritch, V.; Stevens, R. C.; Abagyan, R. Advances in GPCR modeling evaluated by the GPCR dock 2013 assessment: Meeting new challenges. *Structure (Oxford, U. K.)* **2014**, *22* (8), 1120–1139, DOI: 10.1016/j.str.2014.06.012.

(145) Michino, M.; Abola, E.; Brooks, C. L.; Dixon, J. S.; Moult, J.; Stevens, R. C. Community-wide assessment of GPCR structure modelling and ligand docking: GPCR Dock 2008. *Nat. Rev. Drug Discovery* **2009**, *8* (6), 455–463, DOI: 10.1038/nrd2877.

(146) Mason, J. S.; Bortolato, A.; Congreve, M.; Marshall, F. H. New insights from structural biology into the druggability of G protein-coupled receptors. *Trends Pharmacol. Sci.* **2012**, *33* (5), 249–260, DOI: 10.1016/j.tips.2012.02.005.

(147) Bortolato, A.; Tehan, B. G.; Smith, R. T.; Mason, J. S. Methodologies for the examination of water in GPCRs. *Methods Mol. Biol. (New York, NY)* **2018**, *1705*, 207–232, DOI: 10.1007/978-1-4939-7465-8_10.

(148) Langmead, C. J.; Andrews, S. P.; Congreve, M.; Errey, J. C.; Hurrell, E.; Marshall, F. H.; Mason, J. S.; Richardson, C. M.; Robertson, N.; Zhukov, A.; Weir, M. Identification of novel adenosine A2A receptor antagonists by virtual screening. *J. Med. Chem.* **2012**, *55* (5), 1904–1909, DOI: 10.1021/jm201455y.

(149) Rappas, M.; Ali, A. A. E.; Bennett, K. A.; Brown, J. D.; Bucknell, S. J.; Congreve, M.; Cooke, R. M.; Cseke, G.; de Graaf, C.; Dore, A. S.; Errey, J. C.; Jazayeri, A.; Marshall, F. H.; Mason, J. S.; Mould, R.; Patel, J. C.; Tehan, B. G.; Weir, M.; Christopher, J. A. Comparison of orexin 1 and orexin 2 ligand binding modes using x-ray crystallography and computational analysis. *J. Med. Chem.* **2020**, *63* (4), 1528–1543, DOI: 10.1021/acs.jmedchem.9b01787.

(150) Deflorian, F.; Perez-Benito, L.; Lenselink, E. B.; Congreve, M.; van Vlijmen, H. W. T.; Mason, J. S.; Graaf, C. d.; Tresadern, G. Accurate prediction of GPCR ligand binding affinity with free energy perturbation. *J. Chem. Inf. Model.* **2020**, *60* (11), 5563–5579, DOI: 10.1021/acs.jcim.0c00449.

(151) Robertson, N.; Rappas, M.; Dore, A. S.; Brown, J.; Bottegoni, G.; Koglin, M.; Cansfield, J.; Jazayeri, A.; Cooke, R. M.; Marshall, F. H. Structure of the complement C5a receptor bound to the extra-helical antagonist NDT9513727. *Nature* **2018**, *553* (7686), 111–114, DOI: 10.1038/nature25025.

(152) Kooistra, A. J.; Vass, M.; McGuire, R.; Leurs, R.; de Esch, I. J. P.; Vriend, G.; Verhoeven, S.; de Graaf, C. 3D-e-Chem: structural cheminformatics workflows for computer-aided drug discovery. *ChemMedChem* **2018**, *13* (6), 614–626, DOI: 10.1002/cmdc.201700754.

(153) Hollenstein, K.; Kean, J.; Bortolato, A.; Cheng, R. K. Y.; Dore, A. S.; Jazayeri, A.; Cooke, R. M.; Weir, M.; Marshall, F. H. Structure of class B GPCR corticotropin-releasing factor receptor 1. *Nature* **2013**, *499* (7459), 438–443, DOI: 10.1038/nature12357.

(154) Jazayeri, A.; Dore, A. S.; Lamb, D.; Krishnamurthy, H.; Southall, S. M.; Baig, A. H.; Bortolato, A.; Koglin, M.; Robertson, N. J.; Errey, J. C.; Andrews, S. P.; Teobald, I.; Brown, A. J. H.; Cooke, R. M.; Weir, M.; Marshall, F. H. Extra-helical binding site of a glucagon receptor antagonist. *Nature* **2016**, *533* (7602), 274–277, DOI: 10.1038/nature17414.

(155) Oswald, C.; Rappas, M.; Kean, J.; Dore, A. S.; Errey, J. C.; Bennett, K.; Deflorian, F.; Christopher, J. A.; Jazayeri, A.; Mason, J. S.; Congreve, M.; Cooke, R. M.; Marshall, F. H. Intracellular allosteric antagonism of the CCR9 receptor. *Nature* **2016**, *540* (7633), 462–465, DOI: 10.1038/nature20606.

(156) Cheng, R. K. Y.; Fiez-Vandal, C.; Schlenker, O.; Edman, K.; Aggeler, B.; Brown, D. G.; Brown, G. A.; Cooke, R. M.; Dumelin, C. E.; Dore, A. S.; Geschwindner, S.; Grebner, C.; Hermansson, N.-O.; Jazayeri, A.; Johansson, P.; Leong, L.; Prihandoko, R.; Rappas, M.; Soutter, H.; Snijder, A.; Sundstrom, L.; Tehan, B.; Thornton, P.; Troast, D.; Wiggin, G.; Zhukov, A.; Marshall, F. H.; Dekker, N. Structural insight into allosteric modulation of protease-activated receptor 2. *Nature* **2017**, *545* (7652), 112–115, DOI: 10.1038/nature22309.

(157) Vass, M.; Podlewska, S.; de Esch, I. J. P.; Bojarski, A. J.; Leurs, R.; Kooistra, A. J.; de Graaf, C. Aminergic GPCR-ligand interactions: a chemical and structural map of receptor mutation data. *J. Med. Chem.* **2019**, *62* (8), 3784–3839, DOI: 10.1021/acs.jmedchem.8b00836.

(158) Siragusa, L.; Cross, S.; Baroni, M.; Goracci, L.; Cruciani, G. BioGPS: navigating biological space to predict polypharmacology, off-targeting, and selectivity. *Proteins: Struct., Funct., Bioinf.* **2015**, *83* (3), 517–532, DOI: 10.1002/prot.24753.

(159) De Graaf, C.; Kooistra, A. J.; Vischer, H. F.; Katritch, V.; Kuijer, M.; Shiroishi, M.; Iwata, S.; Shimamura, T.; Stevens, R. C.; de Esch, I. J. P.; Leurs, R. Crystal structure-based virtual screening for fragment-like ligands of the human histamine H1 Receptor. *J. Med. Chem.* **2011**, *54* (23), 8195–8206, DOI: 10.1021/jm2011589.

(160) Kooistra, A. J.; Vischer, H. F.; McNaught-Flores, D.; Leurs, R.; de Esch, I. J. P.; de Graaf, C. Function-specific virtual screening for GPCR ligands using a combined scoring method. *Sci. Rep.* **2016**, *6*, 28288, DOI: 10.1038/srep28288.

(161) Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **2007**, *47* (1), 195–207, DOI: 10.1021/ci600342e.

(162) Kuhne, S.; Kooistra, A. J.; Bosma, R.; Bortolato, A.; Wijtmans, M.; Vischer, H. F.; Mason, J. S.; de Graaf, C.; de Esch, I. J. P.; Leurs, R. Identification of ligand binding hot spots of the histamine H1 receptor following structure-based fragment optimization. *J. Med. Chem.* **2016**, *59* (19), 9047–9061, DOI: 10.1021/acs.jmedchem.6b00981.

(163) Kooistra, A. J.; Leurs, R.; de Esch, I. J. P.; de Graaf, C. Structure-based prediction of G-protein-coupled receptor ligand function: a $\beta$-adrenoceptor case study. *J. Chem. Inf. Model.* **2015**, *55* (5), 1045–1061, DOI: 10.1021/acs.jcim.5b00066.

(164) Mason, J. S.; Bortolato, A.; Weiss, D. R.; Deflorian, F.; Tehan, B.; Marshall, F. H. High end GPCR design: crafted ligand design and druggability analysis using protein structure, lipophilic hotspots and explicit water networks. *In Silico Pharmacol.* **2013**, *1* (1), 23, DOI: 10.1186/2193-9616-1-23.

(165) Jespers, W.; Verdon, G.; Azuaje, J.; Majellaro, M.; Keraenen, H.; Garcia-Mera, X.; Congreve, M.; Deflorian, F.; de Graaf, C.; Zhukov, A.; Dore, A. S.; Mason, J. S.; Åqvist, J.; Cooke, R. M.; Sotelo, E.; Gutierrez-de-Teran, H. X-Ray crystallography and free energy calculations reveal the binding mechanism of A2A adenosine receptor antagonists. *Angew. Chem., Int. Ed.* **2020**, *59* (38), 16536–16543, DOI: 10.1002/anie.202003788.

(166) Thomas, M.; Smith, R. T.; O'Boyle, N. M.; de Graaf, C.; Bender, A. Comparison of structure- and ligand-based scoring functions for deep generative models: a GPCR case study. *J. Cheminf.* **2021**, *13* (1), 39, DOI: 10.1186/s13321-021-00516-0.

(167) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminf.* **2017**, *9*, 48, DOI: 10.1186/s13321-017-0235-x.

(168) Fisher, I. Irving Fisher. A world map on a regular icosahedron by gnomonic projection. *Geographical Review* **1943**, *33* (4), 605–619, DOI: 10.2307/209914.

(169) De Ruvo, M.; Giuliani, A.; Paci, P.; Santoni, D.; Di Paola, L. Shedding light on protein-ligand binding by graph theory: the topological nature of allostery. *Biophys. Chem.* **2012**, *165-166*, 21–29, DOI: 10.1016/j.bpc.2012.03.001.

(170) Cang, Z.; Wei, G.-W. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *International Journal for Numerical Methods in Biomedical Engineering* **2018**, *34* (2), e2914, DOI: https://doi.org/10.1002/cnm.2914.

(171) Zrimec, J.; Kokina, M.; Jonasson, S.; Zorrilla, F.; Zelezniak, A. Plastic-degrading potential across the global microbiome correlates with recent pollution trends. *bioRxiv* **2020**, 422558, http://www.biorxiv.org/content/10.1101/2020.12.13.422558v2 (accessed October 13, 2021).

(172) Sarkisyan, K. S.; Bolotin, D. A.; Meer, M. V.; Usmanova, D. R.; Mishin, A. S.; Sharonov, G. V.; Ivankov, D. N.; Bozhanova, N. G.; Baranov, M. S.; Soylemez, O.; Bogatyreva, N. S.; Vlasov, P. K.; Egorov, E. S.; Logacheva, M. D.; Kondrashov, A. S.; Chudakov, D. M.; Putintseva, E. V.; Mamedov, I. Z.; Tawfik, D. S.; Lukyanov, K. A.; Kondrashov, F. A. Local fitness landscape of the green fluorescent protein. *Nature* **2016**, *533* (7603), 397–401, DOI: 10.1038/nature17995.

(173) Repecka, D.; Jauniskis, V.; Karpus, L.; Rembeza, E.; Rokaitis, I.; Zrimec, J.; Poviloniene, S.; Laurynenas, A.; Viknander, S.; Abuajwa, W.; Savolainen, O.; Meskys, R.; Engqvist, M. K. M.; Zelezniak, A. Expanding functional protein sequence spaces using generative adversarial networks. *Nat. Mach. Intell.* **2021**, *3* (4), 324–333, DOI: 10.1038/s42256-021-00310-5.

(174) Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems* **2014**, *27*, http://arxiv.org/pdf/1406.2661.pdf (accessed October 13, 2021).

(175) Lucic, M.; Kurach, K.; Michalski, M.; Gelly, S.; Bousquet, O. Are GANS created equal? a large-scale study. *arXiv preprint arXiv:1711.10337* **2017**, http://export.arxiv.org/pdf/1711.10337 (accessed October 13, 2021).

(176) Zrimec, J.; Boerlin, C. S.; Buric, F.; Muhammad, A. S.; Chen, R.; Siewers, V.; Verendel, V.; Nielsen, J.; Toepel, M.; Zelezniak, A. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nat. Commun.* **2020**, *11* (1), 6141, DOI: 10.1038/s41467-020-19921-4.

(177) Méndez-Lucio, O.; Ahmad, M.; del Rio-Chanona, E. A.; Wegner, J. K. A geometric deep learning approach to predict binding conformations of bioactive molecules. *ChemRxiv* **2021**, http://chemrxiv.org/engage/chemrxiv/article-details/60c757b7337d6cb764e29086 (accessed October 18, 2021).

(178) Gainza, P.; Sverrisson, F.; Monti, F.; Rodola, E.; Boscaini, D.; Bronstein, M. M.; Correia, B. E. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **2020**, *17* (2), 184–192, DOI: 10.1038/s41592-019-0666-6.

(179) Li, Y.; Su, M.; Liu, Z.; Li, J.; Liu, J.; Han, L.; Wang, R. Assessing protein-ligand interaction scoring functions with the CASF-2013 benchmark. *Nat. Protoc.* **2018**, *13* (4), 666–680, DOI: 10.1038/nprot.2017.114.

(180) Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative assessment of scoring functions: the CASF-2016 update. *J. Chem. Inf. Model.* **2019**, *59* (2), 895–913, DOI: 10.1021/acs.jcim.8b00545.

(181) Lesca, G.; Eymard–Pierre, E.; Santorelli, F. M.; Cusmai, R.; Di Capua, M.; Valente, E. M.; Attia–Sobol, J.; Plauchu, H.; Leuzzi, V.; Ponzone, A.; Boespflug–Tanguy, O.; Bertini, E. Infantile ascending hereditary spastic paralysis (IAHSP). *Neurology* **2003**, *60* (4), 674, DOI: 10.1212/01.WNL.0000048207.28790.25.

(182) Sato, K.; Otomo, A.; Ueda, M. T.; Hiratsuka, Y.; Suzuki-Utsunomiya, K.; Sugiyama, J.; Murakoshi, S.; Mitsui, S.; Ono, S.; Nakagawa, S.; Shang, H.-F.; Hadano, S. Altered oligomeric states in pathogenic ALS2 variants associated with juvenile motor neuron diseases cause loss of ALS2-mediated endosomal function. *J. Biol. Chem.* **2018**, *293* (44), 17135–17153, DOI: 10.1074/jbc.ra118.003849.

(183) Soares, D. C.; Barlow, P. N.; Porteous, D. J.; Devon, R. S. An interrupted beta-propeller and protein disorder: structural bioinformatics insights into the N-terminus of alsin. *J. Mol. Model.* **2009**, *15* (2), 113–122, DOI: 10.1007/s00894-008-0381-1.

(184) Delprato, A.; Merithew, E.; Lambright, D. G. Structure, exchange determinants, and family-wide Rab specificity of the tandem helical bundle and Vps9 domains of Rabex-5. *Cell* **2004**, *118* (5), 607–617, DOI: 10.1016/j.cell.2004.08.009.

(185) Caldararu, O.; Blundell, T. L.; Kepp, K. P. Three simple properties explain protein stability change upon mutation. *J. Chem. Inf. Model.* **2021**, *61* (4), 1981–1988, DOI: 10.1021/acs.jcim.1c00201.

(186) Noid, W. G. Perspective: coarse-grained models for biomolecular systems. *J. Chem. Phys.* **2013**, *139* (9), 090901, DOI: 10.1063/1.4818908.

(187) Clementi, C.; Nymeyer, H.; Onuchic, J. N. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* **2000**, *298* (5), 937–953, DOI: 10.1006/jmbi.2000.3693.

(188) Wang, J.; Olsson, S.; Wehmeyer, C.; Perez, A.; Charron, N. E.; de Fabritiis, G.; Noe, F.; Clementi, C. Machine learning of coarse-grained molecular dynamics force fields. *ACS Cent. Sci.* **2019**, *5* (5), 755–767, DOI: 10.1021/acscentsci.8b00913.

(189) Schuett, K. T.; Sauceda, H. E.; Kindermans, P. J.; Tkatchenko, A.; Mueller, K. R. SchNet - a deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148* (24), 241722, DOI: 10.1063/1.5019779.

(190) Husic, B. E.; Charron, N. E.; Lemm, D.; Wang, J.; Perez, A.; Majewski, M.; Kraemer, A.; Chen, Y.; Olsson, S.; de Fabritiis, G.; Noe, F.; Clementi, C. Coarse graining molecular dynamics with graph neural networks. *J. Chem. Phys.* **2020**, *153* (19), 194101, DOI: 10.1063/5.0026133.

(191) Wang, J.; Charron, N.; Husic, B.; Olsson, S.; Noe, F.; Clementi, C. Multi-body effects in a coarse-grained protein force field. *J. Chem. Phys.* **2021**, *154* (16), 164113, DOI: 10.1063/5.0041022.

(192) Santos, R.; Ursu, O.; Gaulton, A.; Bento, A. P.; Donadi, R. S.; Bologa, C. G.; Karlsson, A.; Al-Lazikani, B.; Hersey, A.; Oprea, T. I.; Overington, J. P. A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discovery* **2017**, *16* (1), 19–34, DOI: 10.1038/nrd.2016.230.

(193) Oprea, T. I.; Bologa, C. G.; Brunak, S.; Campbell, A.; Gan, G. N.; Gaulton, A.; Gomez, S. M.; Guha, R.; Hersey, A.; Holmes, J.; Jadhav, A.; Jensen, L. J.; Johnson, G. L.; Karlson, A.; Leach, A. R.; Ma'ayan, A.; Malovannaya, A.; Mani, S.; Mathias, S. L.; McManus, M. T.; Meehan, T. F.; von Mering, C.; Muthas, D.; Nguyen, D.-T.; Overington, J. P.; Papadatos, G.; Qin, J.; Reich, C.; Roth, B. L.; Schurer, S. C.; Simeonov, A.; Sklar, L. A.; Southall, N.; Tomita, S.; Tudose, I.; Ursu, O.; Vidovic, D.; Waller, A.; Westergaard, D.; Yang, J. J.; Zahoranszky-Kohalmi, G. Unexplored therapeutic opportunities in the human genome. *Nat. Rev. Drug Discovery* **2018**, *17* (5), 317–332, DOI: 10.1038/nrd.2018.14.

(194) Oprea, T. I. Exploring the dark genome: implications for precision medicine. *Mamm. Genome* **2019**, *30* (7-8), 192–200, DOI: 10.1007/s00335-019-09809-0.

(195) Sheils, T. K.; Mathias, S. L.; Kelleher, K. J.; Siramshetty, V. B.; Nguyen, D.-T.; Bologa, C. G.; Jensen, L. J.; Vidović, D.; Koleti, A.; Schürer, S. C.; Waller, A.; Yang, J. J.; Holmes, J.; Bocci, G.; Southall, N.; Dharkar, P.; Mathé, E.; Simeonov, A.; Oprea, T. I. TCRD and Pharos 2021: mining the human proteome for disease biology. *Nucleic Acids Res.* **2021**, *49* (D1), D1334–D1346, DOI: 10.1093/nar/gkaa993.

(196) Pletscher-Frankild, S.; Palleja, A.; Tsafou, K.; Binder, J. X.; Jensen, L. J. DISEASES: text mining and data integration of disease-gene associations. *Methods (Amsterdam, Neth.)* **2015**, *74*, 83–89, DOI: 10.1016/j.ymeth.2014.11.020.

(197) Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; Kang, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36* (4), 1234–1240, DOI: 10.1093/bioinformatics/btz682.

(198) Tambuyzer, E.; Vandendriessche, B.; Austin, C. P.; Brooks, P. J.; Larsson, K.; Miller Needleman, K. I.; Valentine, J.; Davies, K.; Groft, S. C.; Preti, R.; Oprea, T. I.; Prunotto, M. Therapies for rare diseases: therapeutic modalities, progress and challenges ahead. *Nat. Rev. Drug Discovery* **2020**, *19* (2), 93–111, DOI: 10.1038/s41573-019-0049-9.

(199) Haendel, M.; Vasilevsky, N.; Unni, D.; Bologa, C.; Harris, N.; Rehm, H.; Hamosh, A.; Baynam, G.; Groza, T.; McMurry, J.; Dawkins, H.; Rath, A.; Thaxon, C.; Bocci, G.; Joachimiak, M. P.; Kohler, S.; Robinson, P. N.; Mungall, C.; Oprea, T. I. How many rare diseases are there? *Nat. Rev. Drug Discovery* **2020**, *19* (2), 77–78, DOI: 10.1038/d41573-019-00180-y.

(200) Avram, S.; Bologa, C. G.; Holmes, J.; Bocci, G.; Thoma; Wilson, B.; Nguyen, D.-T.; Curpan, R.; Halip, L.; Bora, A.; Yang, J. J.; Knockel, J.; Sirimulla, S.; Ursu, O.; Oprea, T. I. DrugCentral 2021 supports drug discovery and repositioning. *Nucleic Acids Res.* **2021**, *49* (D1), D1160–D1169, DOI: 10.1093/nar/gkaa997.

(201) KC, G. B.; Bocci, G.; Verma, S.; Hassan, M. M.; Holmes, J.; Yang, J. J.; Sirimulla, S.; Oprea, T. I. A machine learning platform to estimate anti-SARS-CoV-2 activities. *Nat. Mach. Intell.* **2021**, *3* (6), 527–535, DOI: 10.1038/s42256-021-00335-w.

(202) Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I., WOMBAT: world of molecular bioactivity. In *Chemoinformatics in Drug Discovery*, Mannhold, R., Kubinyi, H., Folkers, G., Oprea, T. I., Eds.; Wiley-VCH: Weinheim, Germany, 2005, pp 760–786, DOI: https://doi.org/10.1002/3527603743.ch9.

(203) Anonymous. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **2019**, *47* (D1), D520–D528, DOI: 10.1093/nar/gky949.

(204) Burley, S. K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chen, L.; Crichlow, G. V.; Christie, C. H.; Dalenberg, K.; Di Costanzo, L.; Duarte, J. M.; Dutta, S.; Feng, Z.; Ganesan, S.; Goodsell, D. S.; Ghosh, S.; Green, R. K.; Guranovic, V.; Guzenko, D.; Hudson, B. P.; Lawson, C. L.; Liang, Y.; Lowe, R.; Namkoong, H.; Peisach, E.; Persikova, I.; Randle, C.; Rose, A.; Rose, Y.; Sali, A.; Segura, J.; Sekharan, M.; Shao, C.; Tao, Y.-P.; Voigt, M.; Westbrook, J. D.; Young, J. Y.; Zardecki, C.; Zhuravleva, M. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* **2021**, *49* (D1), D437–D451, DOI: 10.1093/nar/gkaa1038.

(205) Luger, K.; Mader, A. W.; Richmond, R. K.; Sargent, D. F.; Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **1997**, *389* (6648), 251–260, DOI: 10.1038/38444.

(206) Ferreira, K. N.; Iverson, T. M.; Maghlaoui, K.; Barber, J.; Iwata, S. Architecture of the photosynthetic oxygen-evolving center. *Science* **2004**, *303* (5665), 1831–1838, DOI: 10.1126/science.1093087.

(207) Westbrook, J. D.; Burley, S. K. How structural biologists and the Protein Data Bank contributed to recent FDA new drug approvals. *Structure (Oxford, U. K.)* **2019**, *27* (2), 211–217, DOI: 10.1016/j.str.2018.11.007.

(208) Cleary, E. G.; Beierlein, J. M.; Khanuja, N. S.; McNamee, L. M.; Ledley, F. D. Contribution of NIH funding to new drug approvals 2010-2016. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (10), 2329–2334, DOI: 10.1073/pnas.1715368115.

(209) Burley, S. K.; Berman, H. M. Open-access data: a cornerstone for artificial intelligence approaches to protein structure prediction. *Structure (Oxford, U. K.)* **2021**, *29* (6), 515–520, DOI: 10.1016/j.str.2021.04.010.