Fourth RSC Artificial Intelligence in Chemistry
27th – 28th September 2021

A report by Dr Wendy A Warr, https://www.warr.com/

Introduction

The ("virtual") symposium was organised by the Royal Society of Chemistry's Biological and Medicinal Chemistry Sector (RSC BMCS) and the Royal Society of Chemistry's Chemical Information and Computer Applications Group (RSC CICAG).

AI for molecular design, past, present and future

Ola Engkvist, AstraZeneca, Gothenburg, Sweden

AI-based drug design can reduce the time to deliver a clinical candidate, by helping chemists select the most efficient synthetic route (increasing speed), and making information-rich compounds in each design, analyse, make, test (DMTA) cycle (maximising learning). This could not have been done five years ago but increased computational power, advances in neural network (NN) algorithms, and the availability of open-source software have now made it possible.

We can take advantage of progress in natural language processing (NLP) by representing molecules as SMILES. NLP can then be used in synthesis prediction and molecular optimisation, and text generation can be used in chemical space exploration. We can move from rule-based models to data-driven ones. AI-generated ideas from the whole relevant chemical space can be used for scaffold hopping and hit finding. Fast molecular optimisation is possible through AI-designed libraries. Progress in free-energy perturbation affinity prediction improves scoring of AI-generated molecules. Better prediction of synthetic routes is possible through new algorithms and there are novel and more flexible ways of predicting molecular properties.

Single-layer NNs have been used in modelling of Quantitative Structure-Activity Relationships (QSAR) for years but recent applications use more complex networks such as multi-layer, feed-forward NNs, convolutional NNs, auto-encoder NNs, and recurrent neural networks (RNNs), trained using maximum likelihood estimation to maximise the likelihood of the next character. Recurrent NNs sample the whole chemical space in hit finding and scaffold hopping. A focused chemical space can be sampled with a transformer for molecular optimisation. Generative AI in pharma is still on the ascent in the Gartner Hype Cycle for Artificial Intelligence, 2021 and will peak in 2-5 years.

The chemical space for a file of size 41 GB is 10^9 structures if it is traditionally enumerated whereas generative models can sample practically unlimited chemical space. They do not contain any explicit molecules but generate them probabilistically. An RNN learns the rules of chemistry, not the training examples. The trained RNN can then generate druglike molecules: SMILES are sampled and a probability distribution for each token (character) is used to generate a physicochemical property, or a structure (in which case, training is harder).

In a collaboration with Jean-Louis Reymond's team, Engkvist and his co-workers explored whether it is possible to show that a deep learning based molecular generator is sampling the whole relevant chemical space and only that chemical space. They trained an RNN with a subset of SMILES from the enumerated GDB-13 database of 975 million molecules. They showed that a model trained with 1 million structures reproduces 68.9% of the entire database after training, when sampling 2 billion

molecules. An analysis of the generated chemical space showed that complex molecules with many rings and heteroatoms are more difficult to sample.[1]

The teams then performed a benchmark on models trained with subsets of GDB-13 of different sizes with different SMILES variants, recurrent cell types, and hyperparameter combinations.[2] New metrics were developed that define how well a model has generalised the training set. The generated chemical space was evaluated with respect to its uniformity, closedness and completeness. The results showed that models that use long short-term memory (LSTM) cells trained with 1 million randomised SMILES are able to generalise to larger chemical spaces than the other approaches and they represent more accurately the target chemical space.

Using reinforcement learning (RL), an RNN can be tuned to target a particular section of chemical space with optimised desirable properties using a scoring function but ligands generated by some RL methods tend to have relatively low diversity, and sometimes even result in duplicate structures. Engkvist's team has developed a new method to address this issue: memory-assisted RL introduces a memory unit and a scaffold penalty assures that diverse scaffolds are identified.[3]

Mixed improvements with novel deep learning methods have been reported: there has been no "AlphaFold moment" in blind bioactivity prediction competitions. Gradient descent NNs are approximately kernel machines. Large improvements would imply a novel way of assessing molecular similarity. Pre-training can improve prediction capacity. The data used are more important factors than molecular representation and machine learning (ML) algorithms. Uncertainty quantification and interpretability have to be considered. Most models in the future are likely to be based on deep learning because of their flexibility.

The machine learning ledger orchestration for drug discovery (MELLODDY) project aims, over three years, to enhance predictive ML models on the decentralised data of 10 pharmaceutical companies, without exposing proprietary information. A multi-task approach across partners aims to improve predictive performance and applicability. Compound and activity data and assay-specific models remain locked on the server of the pharma that owns them. Lower-level model components are securely exchanged and trained over the network. Pre-agreed access arrangements are strictly enforced. In year two, a study showed that multi-partner modelling yields superior predictive models in drug discovery.

The MegaMolBART drug discovery model being developed by NVIDIA and AstraZeneca will be used in reaction prediction, molecular optimisation and *de novo* molecular generation. It is based on AstraZeneca's MolBART transformer model and is being trained on ZINC[4] using NVIDIA's Megatron framework to enable massively scaled-out training on a supercomputing infrastructure.

Engkvist's team have demonstrated the utility of a 3D shape and pharmacophore similarity scoring component in molecular design with a deep generative model trained with reinforcement learning (REINVENT).[5] Using dopamine receptor type 2 (DRD2) as an example and its antagonist haloperidol 1 as a starting point in a ligand-based design context, they have shown in a retrospective study that a 3D similarity enabled generative model can discover new leads in the absence of any other information. It can be used for scaffold hopping and generation of novel series. 3D similarity based models were compared against ones based on 2D QSAR, indicating a significant degree of orthogonality of the generated outputs, with the former having a more diverse output.[6]

A major obstacle of generative models is producing active compounds in which predictive QSAR models have been applied to enrich target activity. QSAR models are inherently limited by their applicability domains. A structure-based scoring component for REINVENT overcomes these limitations. DockStream[7] is a flexible, stand-alone molecular docking wrapper that provides access to a collection of ligand embedders and docking back-ends.

Nevertheless, AI alone cannot transform drug design. High-throughput data generation, automatisation in the DMTA cycle, and combining AI with physics (e.g., to predict physicochemical properties and estimate binding affinity) can add value to AI approaches. Combining AI with big data can transform synthesis prediction.[8] In AstraZeneca, chemists have access to data on 20 million reactions in the ReactionConnect database, from which predictive models can be built and used to automate synthesis. ReactionConnect is populated with data from AstraZeneca reaction sources and ELNs, a USPTO database, and Reaxys and Pistachio flat files.[9] AiZynthFinder can be used in retrosynthetic planning. The algorithm is based on a Monte Carlo tree search that recursively breaks down a molecule to purchasable precursors. The tree search is guided by an artificial neural network policy that suggests possible precursors by using a library of known reaction templates.[10] A "Ring Breaker" algorithm[11] improves the route-finding. It uses a data-driven approach to enable the prediction of ring-forming reactions, useful in establishing the synthetic accessibility of unprecedented ring systems. Another improvement, RAscore,[12] is an ML-based method able to classify whether a synthetic route can be identified or not for a particular compound.

Engkvist summarised the lessons that AstraZeneca has learned. The needs of workers in discovery chemistry and process chemistry are very different. Extracting and integrating reaction data is hard work. It is challenging to assess the utility of different tools such as advanced building block look-up. The impact on AI approaches on synthetic routes has mainly been from specialised tools such as Ring Breaker. Software from the Molecular AI department at AstraZeneca is openly available. iLAB is AstraZeneca's automated synthesis platform.

Engkvist is optimistic about the future of AI in drug design because of increased computational power, increased automation which provides large and consistent datasets, and advances in computational algorithms such as those that merge physics-based modelling and ML. Metrics such as time-saving cannot be used to measure success because they are the results of success not the success itself. Success can be measured by trust in the AI-designed molecules in the same way as, for instance, X-ray crystal structures are trusted. There must be trust in the predictions for individual molecules and trust that the AI generated molecules are the best molecules to take the project most efficiently to a clinical candidate.

There remain some challenges for AI driven drug design. They include scaling ML and AI solutions for drug design to a whole drug discovery project portfolio including projects with low data volume. Binding affinity and solubility predictions are major bottlenecks. The "Cambrian revolution" of new AI methods makes it difficult to assess progress. Flexibility of chemistry automation is another challenge. There are also educational, cultural and logistical challenges besides scientific ones. The bar is set high to transform drug design.

Driving lead optimisation with BRADSHAW

Ian Wall, Richard Lonsdale, David Marcus, Darren Green, Stephen Pickett, David Hirst, GlaxoSmithKline (GSK), Stevenage, UK

A *de novo* design program generates molecular structures which satisfy a set of constraints. Classic problems with *de novo* design algorithms are nonsense structures, structures with intrinsic liabilities, and structures that cannot be made. Biological Response Analysis and Design System using an Heterogenous, Automated Workflow (BRADSHAW), GSK's automated molecular design platform (Figure 1),[13] takes a dual approach, using cheminformatics methods to generate plausible structures based on what has been done before,[14-17] and deep learning algorithms trained on relevant GSK chemistry space including novel methods.[5,18]

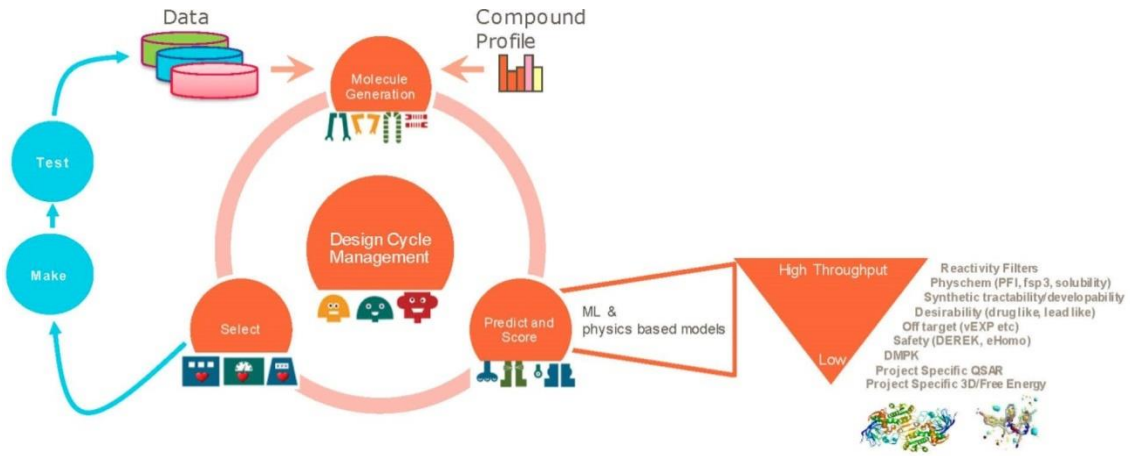


Figure 1. GSK's BRADSHAW.

A GSK team has reported[19] three Turing-inspired tests designed to evaluate the performance of three molecular generators: BioDig, a matched molecular pair-based algorithm,[16] BRICS (a fragment replacement based algorithm),[15] and RG2Smi,[18] which translates a reduced graph input to a SMILES output. BioDig performed excellently against all tests.

Currently, BRADSHAW is limited to cheminformatics and ML models. There are no 3D or docking methods, or physics-based methods such as free energy perturbation (FEP+),[20] but they can be included in a design workflow as an additional step. A multi-parameter optimisation (MPO) approach is used, in automated workflows, to design molecules with a balanced profile. MPOs can be built for predicted values and confidence in them, allowing an active learning approach with algorithmic definition of "explore and exploit".

Wall presented a case study in an active drug discovery programme. The process maximised efficiency by moving the synthetic chemistry resource between two series, allowed updating of models with new data whilst chemists moved onto alternative series, and minimised the number of compounds made without information from previous compounds. The computational chemistry workflow began with molecular generation from seed compounds , followed by building, filtering and rebuilding QSAR models, docking and scoring, and removal of undesirable compounds (by medicinal chemists). FEP calculations were then carried out and the data were collated in Spotfire

for review by medicinal chemists. More than 2 million molecules were generated, 2822 FEP calculations were made, and 38 local models were built, in over more than 30,000 GPU hours.

The technologies used in BRADSHAW are modern ML, active learning, gated recurrent unit cell recurrent neural network (GRU RNN, a new molecule generator which increases the ability to make changes at multiple positions, giving better coverage of chemists' ideas),[21] BRICS, BioDig, Matsy, and RG2smi. In addition, library enumeration, Free-Wilson analysis, pK_a prediction, MetaSite, and protein-ligand interaction fingerprints are used.

Chemists selected compounds for synthesis and viewed their profiles against a range of parameters. This technique was used in conjunction with an active learning explore-exploit plot, where with MPO score on the *x* axis and MPO confidence on the *y* axis, the top right quadrant is compounds for exploitation and the bottom right quadrant is compounds for exploration. Wall showed graphs illustrating the rapid increase in "zero-risk" compounds for the two series since BRADSHAW was introduced in February 2020. The successful outcome of this pilot project was two leads with *in vivo* activity. Wall displayed some of the interesting range of structures (core and R-groups) resulting from exploration of the chemical space, showing some simple structures but with different R-groups and complexity starting to appear.

Close interactions were needed among computational, and medicinal and synthetic chemists, including those in high throughput chemistry (HTC), to get maximum value from the technology. Many other functions were also essential. Wall outlined some pros and cons from the medicinal chemist's viewpoint. From molecule generation, interesting, novel ideas were produced, with a good synthetic success rate, but matched molecular pairs were lacking and there were incomplete enumerations. Scoring and selection were an improvement over the subjective methods used previously, but robustness of pharmacokinetic (PK) predictions and inefficiency in selection meetings were cons. Iterative cycles provided focus but lack of design input, freedom to explore, and medicinal chemistry intuition were criticised. There was an excellent working relationship between medicinal chemists and computational chemists. Unfortunately, data generation has been challenging and restrictive.

Learnings from this pilot project are driving improvements in the system. GRU RNN, improved structural filters and visualisation have been implemented. So has DISCONNECT dHTCscore, a system to identify automatically compounds that are synthesisable from available reagents and possible arrays. Medicinal chemists and computational chemists working together have learnt a huge amount about logistics, technology and ways of working.

Efficient ML strategies to explore chemical reactivity

Fernanda Duarte, University of Oxford, UK

The Duarte group have applied computational methods to design new catalysts and study reaction mechanisms. Their open source tool, cgbind can be used to generate and analyse metallocage structures.[22] Another tool, autodE is an open-source Python package capable of locating transition states and minima and delivering a full reaction energy profile from 1D (SMILES) or 2D chemical representations.[23] It combines graph theory and chemical knowledge in order to reduce the size of the chemical space required for sampling. It is compatible with multiple electronic structure packages, is broadly applicable and requires minimal user expertise.

Realistic simulations of chemical or biochemical reactions require the inclusion of the chemical environment where they occur (e.g., solvent and/or enzyme). Two main approaches have been historically used to account for these complex environments. The first is empirical reactive force fields (e.g., EVB), in combination with molecular dynamics (MD) or Monte Carlo (MC) simulations, which sample a reaction's potential energy surface but are limited in accuracy and transferability. Second are *ab initio* and quantum mechanics/molecular mechanics (QM/MM) which are accurate but computationally costly. ML force fields have the potential to revolutionise force-field based simulations, aiming to provide the best of both worlds.

Duarte's team[24] has used the Gaussian Approximation Potential (GAP)[25-27] framework with smooth overlap of atomic positions (SOAP)[25] descriptors to generate inexpensive potentials for solution phase reactions. GAPs have been applied to organic molecules,[28] and elemental materials[27,29] but this was the first example demonstrating its use to study chemical reactions.

Starting with solute and solvent structures, they developed a training strategy and devised a prospective error metric to assess the accuracy of the potentials. Active learning, where new training data are added based on the current state of the potential, is used for generating databases and accelerating the fitting process. The strategy used by Duarte's team starts from a small number of randomly selected points in the configuration space, from which active learning training of intra- and inter-molecular components of the energy and forces is carried out. The CUR algorithm[27,30] is applied.

Splitting the database into training and test sets and using a standard retrospective validation strategy is not practical in the current application so a temporal cumulative error metric was used based on the time required for the cumulative error to exceed a given threshold. This does not require *a priori* knowledge of the region of configuration space likely to be sampled during a simulation with the potential. The user can specify an acceptable margin of error. The method samples regions not accessible to direct evaluation, ensures stable dynamics, and penalises large errors resulting in instabilities.

For bespoke ML potentials to be routinely developed for molecular systems, one would hope to complete the data generation and model training, and know the accuracy of the resulting potential within a matter of hours to days. With this in mind, the team trained GAP models to simulate bulk water, aiming to minimise the number of required ground truth evaluations as well as the required human intervention, while maximising stability (measured by the new prospective error metric). Only when the relevant length and energy scales of the system are decomposed by treating intra- and inter-molecular components separately was it possible to obtain a potential that is stable for picoseconds.

The model fitted using this approach yields radial distribution functions (RDFs) in good agreement with the ground-truth method, considering both the location and intensities of the peaks corresponding to the first and second coordination shells. The real significance is in moving to more accurate ground-truth methods, for which a full MD simulation would not be straightforward: indeed, using the same method, a hybrid DFT-quality water model can be generated within a few days, which would be inaccessible with other methods. The results suggest that the training strategy (and hyperparameter selection) is suitable independent of the reference method.

To demonstrate the transferability of the models, Duarte briefly presented results of successful application to aqueous Zn (II); to metallocage dynamics;[31,32] to an S_N2 reaction in gas phase and in explicit solvent (where, in both cases, with only hundreds of evaluations of the reference method, reactive ML dynamics is possible); and to a Diels-Alder reaction in the gas phase.

Duarte concluded that Gaussian Approximation Potentials can be trained in a day for reactive molecular systems; prospective model validation is crucial; general potentials must be more than pairwise additive; accuracy beyond density functional theory (DFT) can be approached; and training can be fully automated.

ML models to support risk assessment of small molecules

Andrea Volkamer, Charité Universitätsmedizin Berlin, Germany

In the risk assessment of novel compounds, regulatory agencies require *in vivo* testing for several toxic endpoints. Alternative (*in silico*) strategies include read-across,[33] structural alerts,[34] and ML and QSAR. In this talk, Volkamer addressed computational methods for holistic risk assessment,[35] and in particular, KnowTox,[36] CalUpdate,[37] ChemBioSim,[38] and cytotoxicity maps.[39]

KnowTox, developed in collaboration with BASF, has three different approaches to allow prediction of potentially toxic effects of query compounds: ML models for 88 endpoints, alerts for 919 toxic substructures,[40] and support for read-across in the form of similarity search with RDKit Morgan fingerprints, MACCS keys and physicochemical descriptors with the Tanimoto similarity coefficient.[41]

When deriving a robust and predictive *in silico* model it is important to examine not only the statistical quality of the model but also the estimate of its predictive boundaries. Key factors are applicability, reliability and decidability.[42] Conformal prediction (CP) is a method for confidence estimation in predictions.[43] The model must be statistically valid at a given confidence level and additional calibration step is that the CP framework compares predictions to those previously seen. In a binary classification, validity is the percentage of correct classifications and efficiency is the percentage of single class predictions (SCPs). Volkamer's team built 88 CP models (using RF as the underlying ML model) in KnowTox and the ToxCast dataset of about 8000 compounds and 1000 endpoints.

They then tested, in collaboration with BASF, how the model performed on one of the company's proprietary antiandrogen activity (AA) datasets. The three datasets used were ToxCast AA (for training and testing) and two external AA datasets, from BASF[36] and Norinder *et al.*[44] Results are shown in Table 1. Firstly, the CP technique was deployed (Table 1a, where accuracy of SCPs corresponds to the ratio of correct SCPs divided by all SCPs). Secondly, to improve validity and information efficiency, two adaptations were suggested: k-nearest neighbour (*k*NN) normalisation and balancing the dataset during training (Table 1b). While, initially, valid cross-validation models were obtained, validity and accuracy dropped on in-house data. The implemented adaptions restored validity and improved accuracy at the cost of efficiency but from a toxicologist's point of view, it is better to have no prediction for a compound than a wrong one.

Table 1a. KnowTox Case Study

Dataset	Efficiency			Accuracy (SCPs)			# toxic/ non-toxic
	all	cl.1	cl.0	**all**	**cl.1**	**cl.0**	
ToxCast AA	0.87	0.89	0.87	0.78	0.80	0.78	868/5842
Norinder	0.79	0.77	0.81	0.68	0.70	0.67	160/201
BASF	0.94	0.98	0.91	**0.56**	**0.97**	**0.07**	280/254

Table 1b. After *k*NN Normalization and Balancing

Norinder	0.43	0.33	0.52	0.74	0.67	0.78	160/201
BASF	0.20	0.18	0.23	**0.75**	**0.80**	**0.71**	280/254

CalUpdate[37] (developed in conjunction with workers at University College London and the Universities of Uppsala and Stockholm) assesses model calibration and suggests strategies to update models to account for predictivity drops when training and test data do not stem from the same distribution. Here, CP is used to assess the calibration of the models. Using the chronologically released Tox21 subsets Tox21Train, Tox21Test and Tox21Score, the researchers observed that while internally valid models could be trained using cross-validation on Tox21Train, predictions on the external Tox21Score data resulted in higher error rates than expected. To improve the external predictions, a strategy exchanging the calibration set with more recent data, such as Tox21Test, was introduced. The proposed improvement strategy, exchanging the calibration data only, is convenient as it does not require retraining of the underlying model.

In ChemBioSim, workers at BASF, Örebro and Vienna Universities and in Volkamer's team have enhanced the performance of CP models for *in vivo* endpoint predictions by combining molecular descriptors (RDKit Morgan fingerprints and physicochemical properties) with predicted bioactivity ones.[38] Biological fingerprints, describing the activity profile of a molecule, are more mechanistic descriptors, independent of molecular structure. These are actual assay measurements, but since they are not necessarily available at scale and would need to be measured for new compounds as well, the researchers chose to predict them by training CP models for 373 biological assays. The method was exemplified on three *in vivo* endpoints capturing genotoxic (MNT), hepatic (DILI), and cardiological (DICC) issues. The incorporation of bioactivity descriptors increased the mean F1 scores of the MNT model from 0.61 to 0.70 and for the DICC model from 0.72 to 0.82 while the mean efficiencies increased by roughly 0.10 for both endpoints. In contrast, for the DILI endpoint, no significant improvement in model performance was observed. An analysis of the most important bioactivity features allowed detection of novel and less intuitive relationships between the predicted biological assay outcomes used as descriptors and the *in vivo* endpoints.

Finally, Volkamer's team have studied cytotoxicity prediction,[45] one of the earliest handles in drug discovery, using a deep learning approach trained on a dataset of over 34,000 compounds, fewer than 5% of which were cytotoxic. The dataset was from collaborators at the Leibniz-Forschungsinstitut für Molekulare Pharmakologie in Berlin. The encoding involved RDKit Morgan fingerprints. A deep NN with parameter optimisation, balancing and 10-fold nested cross-validation were used. The model reached a balanced accuracy of over 70%, similar to previously reported studies using RF or CP, but different underlying cytotoxicity datasets and activity shares.[46] NNs are often described as a "black boxes". To overcome this absence of interpretability, a deep Taylor decomposition method with layer-wise relevance propagation (LRP)[47] was investigated to identify

toxicophores. A forward path of the trained model is used to get a prediction score which is interpreted as relevance. A backward path of the trained model is used to get decompositions of relevance on input. Toxicophores are identified by mapping the relevance back to atom environments, namely the bits in the Morgan fingerprints. The study also introduced cytotoxicity maps which provide a visual structural interpretation of the relevance of these toxicophore substructures.

About 2.8 million laboratory animals were used in Germany in 2018; establishment of alternative methods could lead to a reduction of animal testing. To this end, Volkamer's team have used CP models and deep learning to predict compounds likely to be ineffective or toxic and exclude them *a priori* from animal testing. Holistic and combined approaches with proven applicability, reliability and interpretability, demonstrated by predictive power and prospective studies will increase acceptance by regulatory authorities.

Exploring molecular space and accelerating drug discovery with Clara Discovery and MegaMolBART

Michelle Gill, NVIDIA, Santa Clara, CA, USA

To extract scientific insights from today's massive datasets we need methods that take advantage of the complexity of the data and can scale efficiently. The increased degree of parallelism afforded by GPUs has made them ideal for the acceleration of analysis and visualisations. Such applications can be combined with methods derived from deep learning to create analysis pipelines that are both faster and more accurate than the existing state of the art. Clara Discovery is a collection of frameworks, applications, and AI models that together accelerate drug discovery, supporting research in genomics, microscopy, virtual screening, computational chemistry, visualisation, clinical imaging and natural language processing (NLP). Gill concentrated on RAPIDS and MegaMolBART.[48]

One example is an interactive clustering and visualisation workflow in which RDKit-derived Morgan fingerprints from ChEMBL (or another database) are used in principal component analysis (PCA), *k*NN clustering, and UMAP visualisation. This pipeline is implemented using cuML and can be performed in real-time due to the acceleration afforded by GPUs. The plotly interface can be customised.

MegaMolBART[48] is mentioned in Engkvist's talk earlier in this report. Pre-training is performed on a subset of ZINC15. SMILES molecules are masked and enumerated (randomised) during training. NVIDIA carries out training on a DGX SuperPOD (4-8 nodes x 8 A100 GPUs). AstraZeneca is concurrently training on Cambridge-1. The pre-trained model is wrapped into a service (Figure 2). The interactive explorer provides a framework for visualising and customising workflows. Deep learning derived features from MegaMolBART can enable analyses that previously required hours to be completed in seconds.

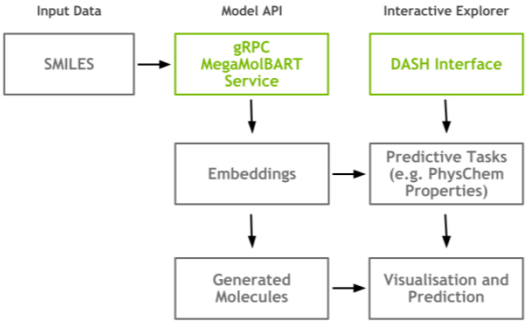


Figure 2. MegaMolBART model service.

In future, NVIDIA will investigate the limits of model size of MegaMolBART and will develop novel model architectures for improved molecule generation. Predictive tasks such as physicochemical properties, reaction prediction and retrosynthetic synthesis could be based on model embeddings. The user experience will be improved by automation of data processing, pre-training and downstream tasks.

Challenges and opportunities for machine learning in drug discovery

W. Patrick Walters, Relay Therapeutics, Cambridge, MA, USA

Over the last few years there has been a dramatic growth in the application of ML in drug discovery. It is impacting numerous areas including image analysis, organic synthesis planning, predictive models, quantum chemistry and molecule generation but there are significant challenges. AI predictions are typically treated as a "black box" which supplies no explanation, yet interpretable models could drive discovery by providing a rationale that convinces people to perform experiments, allowing scientists to gain insights that drive compound design, and enabling efficient debugging of model performance.

Matveieva and Polishchuk[49] have published benchmarks for interpretation of QSAR models. Feature attribution techniques are popular choices for explainability tools, as they can help elucidate which parts of the provided inputs used by an underlying supervised-learning method are considered relevant for a specific prediction, but Jimenez-Luna *et al.*[50] found that none of the feature attribution methods they tested generalised well when confronted with unseen examples. One interesting approach to explainability is the use of "counterfactuals".[51] They are used in credit card approval applications because the law demands that credit card denial be explained. These methods look at the small differences between two people, one whose card is declined and the other whose card is approved. Walters presented an example of the use of counterfactuals for prediction of the solubility of imatinib. He generated analogues, predicted their solubility, sorted them by similarity and evaluated the counterfactuals, looking for small differences. This method seems to work (Figure 3).

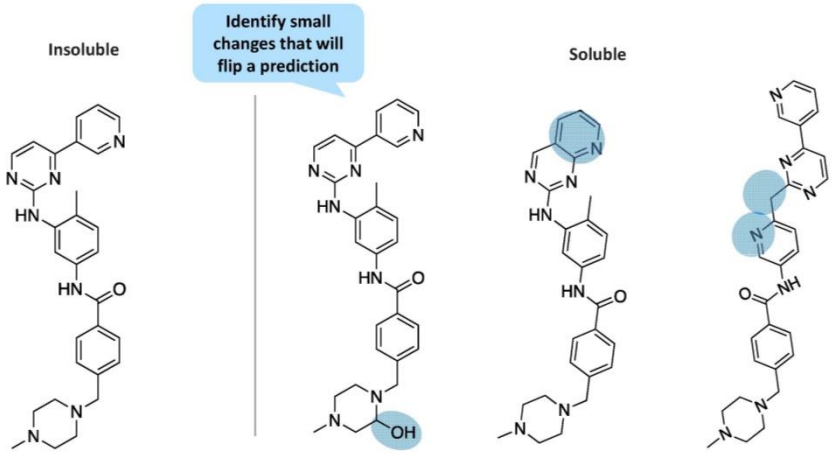


Figure 3. Predicted soluble analogues of imatinib.

Another issue is impossible molecules emerging from generative models. GuacaMol[52] benchmarking for *de novo* molecular design employs Walters' earlier metrics for compound quality but the filters do not detect a number of "chemically impossible" features such as triple bonds in aromatic rings, so Walters has written "silly walks" code.

Another issue is molecular representations. For many years, machine learning models have been constructed using standard molecular fingerprints. More recently, a number of groups have published methods that use neural networks to generate targeted molecular representations.[53,54] To determine if learned representations are better, Walters has written "Yet another ML method comparison" to compare a number of commonly used molecular representations and algorithms. In these tests a standard XGBoost method using molecular fingerprints tends to outperform the learned representations on smaller datasets (less than 2000 molecules).

Why do we not use 3D descriptors more often in ML? Traditional ML methods map one object to one label but molecules can have many 3D conformations. To tackle the relationship between multiple instances and a single label, specialised multiple instance machine learning methods must be used. Recent papers[55,56] examine whether 3D multiple-instance approaches will work. Results vary across datasets but 3D multiple-instance models do appear to be competitive with 2D ones.

Finally, Walters discussed uncertainty and model applicability. A number of methods have been tried to determine when a model is applicable but none of them is ideal. There is a pitfall in scaffold-based[57] cross-validation, training on one scaffold and testing on another. The idea feels good but why should it work? Different chemotypes often make different interactions. The model must implicitly learn these interactions. Walters found that 12 inhibitors of p38 have a very wide variety of interactions in ATP binding pockets. It is important to evaluate your model in context.

ML is impacting many aspects of drug discovery and there are many issues to address, including explainability, representation, model applicability, and multi-objective optimisation. Whilst we have made progress on parts of the puzzle, we are still far from a complete solution. To succeed we need the overlapping domains of "hacking skills", mathematics and statistics knowledge, and substantive domain expertise.

Molecular Transformer-aided biocatalysed synthesis planning

Daniel Probst, Matteo Manica, Yves Gaëtan Nana Teukam, Alessandro Castrogiovanni, Federico Paratore, Teodoro Laino, IBM Research Europe, Rüschlikon, Switzerland

Enzyme biocatalysts are an integral part of green chemistry strategies towards a more sustainable and resource-efficient chemical synthesis. Most are proteins, one third of them require one or more cofactor in the form of inorganic ions, and others require complex molecules as cofactors. Enzymes affect only the reaction rates, not the equilibria, and rate enhancements brought about by enzymes are in the range of 5-17 orders of magnitude. Enzymes have industrial uses in fermentation (e.g., in antibiotic production and brewing) and in enzyme technology (e.g., paper pre-bleaching, food processing and enantionmerically pure amino acids).

They are stereo, regio and chemoselective, highly efficient, reusable and biodegradable, and moderate temperatures and pH are required, but they are unstable at high temperatures or extreme pH, require expensive co-substrates, are potential allergens, and generate metabolic by-products. Unfortunately, a narrow substrate scope is documented in enzyme databases and synthetic chemists have difficulties in identifying patterns within enzyme classes that allow them to extend those patterns to unreported substrates. In addition, other domain-specific knowledge factors such as stereo- and regioselectivity are lacking.

Biocatalytic retrosynthesis has recently been automated by creating expert-curated reaction rules based on available literature, creating a network of molecules connected by enzymes and reaction rules, and applying the rules to arbitrary query molecules in order to find both a matching enzyme and a precursor that can be purchased. RetroBioCat[58] is an example. Unfortunately the creation of expert-curated reaction rules does not scale.

Kreutter *et al.*[59] have tackled this issue by using multi-task transfer learning to train the molecular transformer[60], a sequence-to-sequence machine learning model, with one million reactions from the USPTO database combined with 32,181 enzymatic transformations annotated with a text description of the enzyme. This translates the substrates and enzyme into products. The resulting enzymatic transformer model predicts the structure and stereochemistry of enzyme-catalysed reaction products with remarkable accuracy. The researchers combined the reaction SMILES language of only 405 atomic tokens with thousands of human language tokens describing the enzymes, such that the enzymatic transformer not only learned to interpret SMILES, but also the natural language as used by human experts to describe enzymes and their mutations.

Probst *et al.* have, in addition to the forward model, introduced a retrosynthesis model using a class token based on the Enzyme Commission (EC) number classification scheme that allows them to capture catalysis patterns among different enzymes belonging to the same hierarchical families.[61] Data sources are BRENDA, MetaNetX, PathBank, and Rhea, leading to 62,222 deduplicated, biocatalysed reactions. Probst showed TMAP[62] visualisations of the substrates and products (using MAP4 fingerprints).[63] Modified cofactors are removed from the products. The dataset is not balanced: transferases are over-represented. Tokenisation includes the first three parts of the EC number. An EC number (e.g., 2.6.1.2) has four levels: class, sub-class, sub-sub-class, and serial number (SN). The SN is not used because adding it causes a drop in performance. Performance is limited by dataset size, diversity and quality. The forward prediction model achieves a top-5

accuracy of 62.7%, while the single-step retrosynthetic model shows a top-1 round-trip accuracy of 39.6%. As regards accuracy across classes, class 2, a big class, pushes up accuracy whereas class 1 is poorer because there are too few training data.

Attention weights learned by a transformer encode atom rearrangement information between products and reactants. Attention weight analysis unboxes the forward model to understand how enzyme information is utilised. The IBM team has shown that the EC tokens relate to the centres of the enzymatic reaction and that the forward model captures enzymatic reaction rules based on the EC number. The model mimics the expert-curated reaction rules in automated retrosynthesis.

A resident chemist has tried the system out and has been able to replace a traditional reaction with an enzyme-catalysed one. Anyone can try the system for free at IBM RXN for Chemistry. Stereochemistry is included for all reactions. The enzymatic data and the trained models are available through the RXN for Chemistry network and on GitHub.

Highly accurate protein structure prediction with AlphaFold

Alexander Pritzel, DeepMind, London, UK

A central part of DeepMind's mission is to solve fundamental scientific problems with AI. Predicting the 3D structure of a protein from its amino acid sequence is one such challenge. AlphaFold[64] is DeepMind's model that aims to solve this problem. Proteins consist of chains of amino acids that fold into a 3D structure and the exact 3D shape is important for a protein's function. Experimental structure determination takes months to years; structure prediction can provide actionable information faster.

Critical Assessment of Structure Prediction (CASP) is an organisation that conducts double-blind, community-wide experiments to determine the state of the art of computational methods for modelling protein structures. The CASP assessment involves predicting recently solved structures that are not yet public. In the 14th biennial CASP (CASP14) across a wide range of difficult targets AlphaFold was the top-ranked method: assessors judged its predictions to be at an accuracy "competitive with experiment" for approximately two thirds of proteins.

A key question in the design of neural network architectures is the question of inductive bias, which controls which kind of functions are easy or hard to model. In convolutional networks (used for computer vision, for example) the data are in a regular grid and information flows to local neighbours. AlphaFold 1 used this inductive bias. In recurrent networks (e.g., for language) data are in an ordered sequence and information flows sequentially. In graph networks (e.g., for recommender systems or molecules) data are in a fixed graph structure and information flows along fixed edges. In an attention module (e.g., for language) data are in an unordered set and information flow is dynamically controlled by the network (*via* keys and queries).

High-throughput sequencing technologies have enabled the construction of a multiple sequence alignment (MSA), and accurate coevolution signals can be disentangled. Detected coevolved pairs can be used as residue-residue contact constraints in protein structure modelling and prediction of protein-protein interactions.[65,66] In AlphaFold physical and geometric insights are built into the network structure, and are not just a process around it. This is an end-to-end system directly producing a structure instead of inter-residue distances. Inductive biases reflect knowledge of

protein physics and geometry. The positions of residues in the sequence are de-emphasized. Instead, residues that are close in the folded protein need to communicate. The network iteratively learns a graph of which residues are close, while reasoning over this implicit graph as it is being built. In co-evolution, residues in contact must mutate together (mutation of a single residue breaks the contact and the organism with the mutated protein does not survive). Evolution conserves some properties such as hydrophobic and hydrophilic amino acids being on the "inside" or "outside" of a protein.

Figure 4 presents an outline of how AlphaFold works. A key input is the MSA, containing sequences evolutionarily related to the target. Related sequences are found using standard tools and public databases. The input sequence is used to create an array of representations representing all residue pairs. AlphaFold can also use template structures from the Protein Data Bank (PDB) but it often produces accurate predictions without a template. The Evoformer blocks extract information about the relationship between residues. The MSA representation can update the pair representation and *vice versa*. The Structure Module predicts a rotation and translation to place each residue. A small network predicts side chain chi angles. The final structure is run through a relaxation process. Feeding certain outputs back through the network again improves performance.

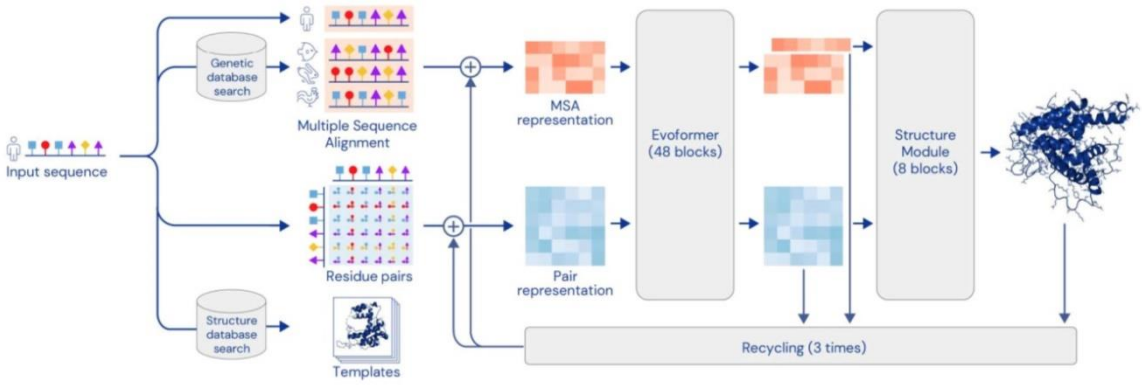


Figure 4. AlphaFold overview.

As well as a predicted structure, the Evoformer blocks produce two confidence estimates: per-residue confidence (for predicted local Distance Difference Test, plDDT)[67] and pairwise confidence (predicted aligned error, PAE). Further detail of the Evoformer architecture is given in Figure 5. In triangular attention, consider three points A, B and C. If distances AB and BC are known, the triangle inequality places a strong constraint on the distance AC. Evolution and sequence give information about relations between residues and pair embedding encodes the relations. The update for pair AC should depend on BC and AB. In the graph, edges represent pairs of residues. Since the graph is unknown it has to be inferred. There is a triplet relation in this language with cycles of a length of three in the graph. The update applied by the layer is based on all cycles involving the edge. More abstractly this can be viewed as a transitivity inductive bias that encodes the transitivity of relations (e.g., triangle inequality and loop closure).

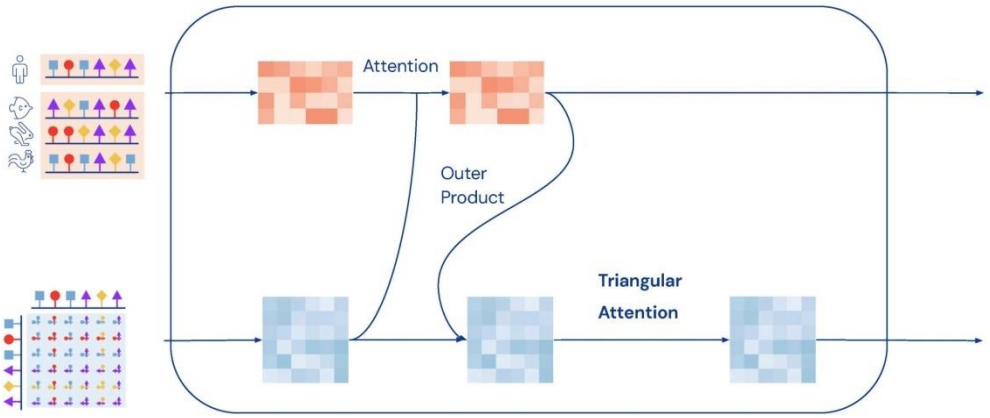
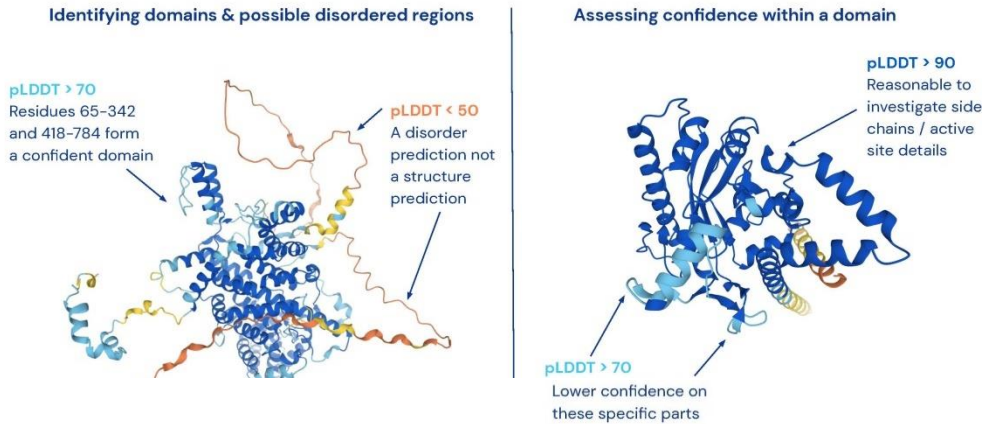


Figure 5. Evoformer.

The structure module performs end-to-end folding instead of gradient descent. Here the protein backbone is modelled as a gas of independent 3D rigid bodies. The spatial structure of the amino acid chain is not built into the model but emerges through learning. A 3D equivariant transformer architecture updates the rigid bodies modelling the backbone and also builds the side chains by predicting torsion angles. The AlphaFold architecture can be trained to high accuracy using only supervised learning on PDB data, but accuracy can be further enhanced using an approach similar to noisy student self-distillation.[68] This is the way AlphaFold makes use of unlabelled sequences. The AlphaFold model is first trained on PDB data alone. This first model is used to predict structures on a large set of unlabelled sequences and then a second model is trained where the training set is enriched by confidently predicted structures of the first model.

Computational structure prediction is typically underspecified, for example as regards oligomeric state, ligands, DNA-binding, experimental conditions, multiple conformations etc. The AlphaFold network implicitly models this missing context using a variety of physical and evolutionary information. Movies of model interpretability for SARS-CoV-2 ORF8 (T1064, one of the hardest examples in CASP14) and a RNA polymerase with over 2000 amino acids (T1044 in CASP14) were shown to illustrate how the model can be interrogated.

Predictions can be interpreted using plDDT and PAE. Roughly speaking, lDDT measures the percentage of correctly predicted interatomic distances, not how well the predicted and true structures can be superimposed. It rewards locally correct structures, and getting individual domains right. plDDT is a measure of local confidence (Figure 6) but high plDDT on all domains does not imply AlphaFold is confident of their relative positions. Assessing inter-domain confidence requires the PAE metric. This is AlphaFold's prediction of the position error at residue *x*, if the predicted and the true structures are aligned on residue *y*. PAE aims to measure confidence in the relative positions of pairs of residues. It is mainly used to assess relative domain positions, but is applicable whenever pairwise confidence is relevant. PAE is displayed as a 2D plot. If residue *y* is aligned to the true structure and the position error at residue *x* is measured, the colour at (*x*, *y*) is AlphaFold's prediction of that error.



Uniprot ID Q5VSL9 Uniprot ID Q96PD7

Figure 6. Use of plDDT.

The AlphaFold protein structure database is a website developed by DeepMind and EMBL-EBI that contains pre-run predictions for 21 model organisms. The AlphaFold colab is a website hosting a pre-written Python program to be executed on a machine in the cloud; you enter a sequence and hit "play" at each step. There are also several other community developed colabs for structure prediction. You can also download the code and run AlphaFold on your own machine. AlphaFold has been received with excitement by the biology community and incorporated in other tools. It has been used in accelerating structure determination, in docking, in predicting disorder and in finding new insights from the AlphaFold database. There is much exciting work ahead for the structural biology field: complexes, disorder and conformational change etc. DeepMind is very excited to see what others are building on top of the AlphaFold database. There is great potential in AI for science as a whole.

"Attending" to co-crystals in the Cambridge Structural Database

Aikaterini Vriza,[1] Angelos B. Canaj,[1] Rebecca Vismara,[1] Laurence J. Kershaw Cook,[1] Troy D. Manning,[1] Michael W. Gaultois,[1] Peter A. Wood,[2] Vitaliy Kurlin,[1] Neil Berry,[1] Matthew. S. Dyer,[1] Matthew J. Rosseinsky.[1] (1) University of Liverpool, UK (2) Cambridge Crystallographic Data Centre, Cambridge UK

A co-crystal is a crystalline single-phase material composed of two or more different molecular compounds in a specific stoichiometry. They are connected *via* non-covalent interactions, such as hydrogen bonding, π–π stacking, halogen bonds and charge transfer interactions. Co-crystals have been particularly useful in improving the physicochemical properties of potential drugs but the current work was focused on the design of co-crystals with electronic functionalities. Polycyclic aromatic hydrocarbons (PAHs) self-assemble *via* π–π interactions and are considered promising candidates for electronic materials. Vriza and her co-workers aimed not only to detect some weakly bound PAH co-crystals but also to understand the important factors contributing to their formation.

The aim is to find molecular pairs which are more likely to form a co-crystal. The problem is that we know which combinations can form co-crystals but we have no information for those that do not. The workflow for co-crystal prediction is a closed loop of database analysis, ML, optimisation, and experimentation. Two datasets were created starting with eight electron-rich PAHs with distinct

geometry by carrying out similarity searches and removing molecules with H-bonding. The sets contained 1722 known molecular combinations from the Cambridge Structural Database (CSD) and 21,736 possible ones from ZINC15, forming labelled (training) and unlabelled datasets respectively. Dragon descriptors were used as features of the two datasets. Each molecular pair was represented as a concatenation of the molecular descriptors.

Most co-crystal prediction research has focused on generating negative data for training binary classifiers. The current work, involving one-class classification, focuses only on the positive data and trying to define a reliable area where novel pairs can exist.[69] The aim of Deep Support Vector Data Description (DeepSVDD) is to find a data-enclosing hypersphere of minimum size, such that the normal data points will be mapped near the centre of the hypersphere whereas anomalous data are mapped further away. The objective of DeepSVDD is to learn the network parameters and minimise the volume of the hypersphere. The deep learning protocol is a two-step process. The first (pre-training) step uses a convolutional autoencoder to capture the representation of the data. During this step the centre of the hypersphere is calculated and is fixed as the mean of the network representations of the known data. During the second step, the latent dimension of the encoder is connected to a feed-forward NN to minimise the loss function (the distance from the centre of the hypersphere). In the Deep One Class method of Vriza *et al.* the convolutional autoencoder was substituted with a Set Transformer autoencoder which is capable of handling the order invariance of the molecular pairs.

The algorithms implemented for one-class classification were separated into eight traditional ones and one NN. Vriza showed the overlapping score distribution of both the labelled and unlabelled data for all the algorithms. The unlabelled data consist of both positive and negative examples in an unknown proportion. Consequently, a certain part of the unlabelled data is expected to belong to the known class (i.e., are inliers). Moreover, in the labelled data there is a small proportion of examples that significantly differs from the rest of the data and is regarded as noise of the normal class (i.e., outlier examples). The impact of the class noise is mitigated using one class classification, as a proportion of the labelled data is regarded as outliers during the hyperparameter optimisation process. A clearer and more definite separation among the two different datasets can be observed for both the Ensemble and Deep One Class methods, with Deep One Class covering a bigger range of scores and thus enabling a better separation.

Vriza showed learning curves of all the algorithms showing the performance of the models while the size of the training set increases. The validation metric used was the true positive rate (the number of correctly predicted inliers divided by the total size of the training set in each fold of the five-fold cross validation). The learning model outperforms the traditional algorithms as it has higher accuracy and low standard deviation.

Scatterplots showing the distribution of representative descriptors among the molecular pairs on the labelled dataset indicate that the deep learning model can effectively learn the trends of the labelled data and is able to score the unlabelled data based on the significant patterns of the labelled data. Focusing on the highest-ranking pairs predicted, the team tried to optimise the selection by targeting molecules with similarity to 7,7,8,8-tetracyanoquinodimethane (TCNQ) which is extensively studied for its interesting electronic properties. Pyrene:benzochromenone (CSD: EHUFIZ) and pyrene:dicyanoanthracene (CSD: EHUFEV) were identified and experimentally validated, both

containing molecules which have not previously been reported as co-formers in the CSD. These were two unlabelled inlier co-crystals lying in the densest area of the scatterplots regarding the polarity and electronic descriptors. Although shape, size and polarity are key factors, the rules that dominate co-crystal formation are far more complex than just some general properties.

The researchers then looked at molecular representations in Set Transformer and evaluated those using publicly available benchmarks. The representations were Mordred descriptors,[70] RDKit Morgan fingerprints, graph embeddings (GNN fingerprints) and representations used in natural language processing (NLP) such as Molecular Transformer.[60] Vriza *et al.* tuned the hyperparameters to reduce the reconstruction error and found that Morgan and GNN fingerprints performed best on all the validation data in terms of total accuracy (specificity, area under the receiver operating characteristic curve (ROC AUC) and recall). The two types of fingerprint also performed well in a head-to-head comparison on co-crystal screening data for 18 active pharmaceutical ingredients.

It has been said that there are some tasks for which there are simply not enough labelled data so we need to focus on ML methods that do not rely on labels. The applicability of the one class unsupervised approach to all CSD co-crystals has been validated in real case scenarios. Currently there are several ML models for co-crystal screening. The Liverpool team has provided a large amount of external validation data and carried out extensive testing against several methods. The workers focused on AI model development: permutation invariant neural networks, attention to extract relations, hyperparameter tuning and reconstruction error minimisation. They also tested several types of distinct inputs and found that Morgan and GNN fingerprints described the molecular pairs better than other inputs.

PyPEF, an integrated framework for data-driven protein design and engineering

Niklas E. Siedhoff[1], Alexander-Maurice Illig[1], Ulrich Schwaneberg,[1,2] **Mehdi D. Davari**.[3] (1) RWTH Aachen University, Aachen, Germany (2) DWI-Leibniz Institute for Interactive Materials, Aachen, Germany (3) Leibniz Institute of Plant Biochemistry, Halle, Germany

Davari's group is interested in enzymes involved in catalysis in cells. Establishing protein sequence, structure, and function relationships is a grand challenge for experiment and computation. There has been progress on structure-sequence links, on design of sequences based on function, and on prediction of function based on sequences, but the dynamics linking structure to function is still a big challenge.

Directed evolution (for which Frances H. Arnold won half a Nobel Prize in 2018) depends on generating a large gene library, needing lots of costly effort. Rational, computer-aided design techniques might never be able to sample through the entire protein sequence space and benefit from nature's full potential for the generation of better enzymes. There is a clear trend to combine the rational design and directed evolution approaches. Semi-rational design generates small, functionally rich, mutant libraries using rationally pre-selected target sites. Knowledge-driven approaches navigate sequence space intelligently. Recently, machine learning methods have been increasingly applied to find patterns in data that help predict protein structures, improve enzyme stability, solubility, and function, predict substrate specificity, and guide rational protein design.[71-74]

In evolutionary biology, fitness landscapes are used to understand the relationship between genotypes and reproductive success. It is assumed that every genotype has a well-defined replication rate (fitness). This fitness is the “height” of the landscape. Genotypes which are similar are said to be close to each other, while those that are very different are far from each other. The set of all possible genotypes, their degree of similarity, and their related fitness values is then called a fitness landscape. The size of the protein sequence space is huge and the fitness landscape is complex. Current challenges are screening throughput (leading to limited exploration, information gaps and local maxima); the combinatorial problem of epistasis (a phenomenon in which the effect of a gene mutation is dependent on the presence or absence of mutations in one or more other genes); and cost and time.

Combining next generation sequencing (high-throughput analysis of DNA and RNA sequences) with high throughput screening of 10^4-10^8 variants per day is a powerful strategy (deep mutational scanning) for comprehensively analysing sequence-function relationships.[72] ML-guided directed evolution reduces experimental effort and mutates multiple positions simultaneously, combining directed evolution and rational design (Figure 7).[71,73]

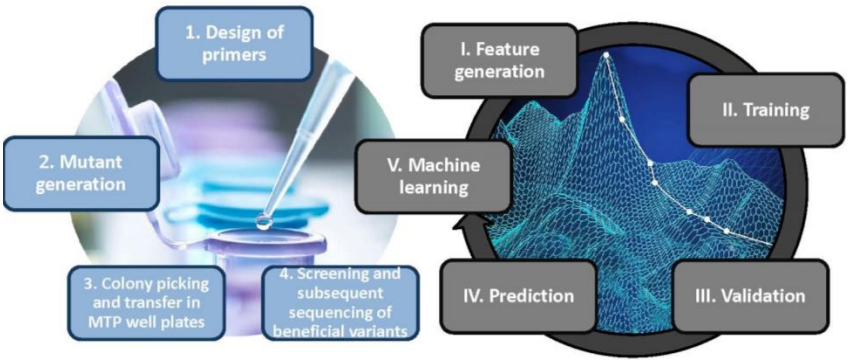


Figure 7. ML-guided directed evolution.

Pythonic Protein Engineering Framework (PyPEF, Figure 8) is a general-purpose framework for data-driven protein engineering by combining machine learning methods (partial least squares (PLS), RF, support vector regression (SVR), and multilayer perceptron (MLP) based regression) with signal processing (fast Fourier transform, FFT) and statistical physics (Metropolis-Hastings algorithm) techniques.[75] It assists in the identification and selection of beneficial proteins in the sequence space by either systematically or randomly exploring the fitness of protein variants and by sampling random evolution pathways. It applies featurisation by Fourier-transforming numerical indices, which represent physicochemical and biochemical properties for each amino acid, taken from the amino acid index (AAindex).

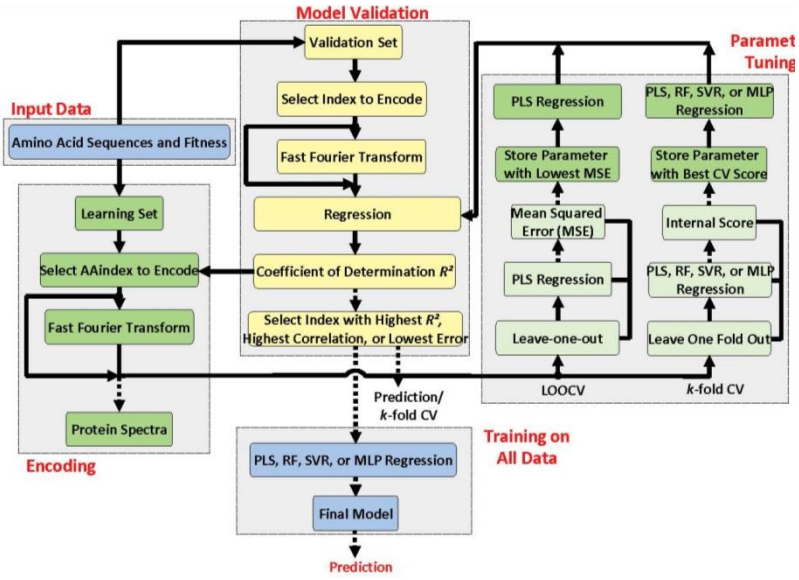


Figure 8. PyPEF framework.

The predictive accuracy and throughput performance of the framework was evaluated based on four publicly available datasets of proteins and enzymes and their properties, using common regression models. PyPEF learned on datasets of small-to-medium-size, derived by diverse evolution strategies, and demonstrated potential to generate predictive models consistently, by accounting for either additive effects only (AAindex encoding and linear models) or non-additive effects within the range of values learned during modelling (AAindex encoding and non-linear models) as well as both inside and outside the range of values learned during modelling, while providing effective *in silico* screening capabilities.

The framework could efficiently predict the fitness of protein sequences for different target properties with R^2 using PLS regression and FFT encodings ranging from 0.58 to 0.92. It enabled more than half a million protein sequences to be screened for various functions in only a few minutes on a standard PC. Data-driven models generated by PyPEF with significant accuracies on four public datasets highlighted the potential for predicting the fitness of variants with high accuracy or capturing the general trend of introduced mutations on the fitness in directed protein evolution campaigns. PyPEF code is publicly available.

Best practice for chemical language model *de novo* design of GPCR ligands: datasets, scoring functions and optimisation algorithms

Morgan Thomas[1], Noel M. O'Boyle[2], David Araripe[1], Rob T. Smith[2], Chris de Graaf,[2] Andreas Bender.[1]
(1) University of Cambridge, UK (2) Sosei Heptares, Cambridge, UK

There has been significant interest in *de novo* molecular design recently. Thomas discussed some aspects of the practical use of chemical language models (e.g., SMILES with recurrent neural networks) which are popular due to their simplicity, performance (by benchmarking works GuacaMol,[52] Molecular Sets (MOSES),[76] and Smina and Therapeutic Data Commons), code availability and support. Both structure-based and ligand-based design can be used. In the latter case prior ligand knowledge may not be available and if it is, it may bias molecule generation towards known chemotypes. Structural data are difficult to acquire (though they are increasingly available)

but structure-based design is not biased by prior ligand knowledge. G Protein Coupled Receptors (GPCRs) are a particular target class where structural data can have a significant impact.[77]

Thomas and his co-workers[78] have assessed the use of molecular docking *via* Glide (a structure-based approach) as a scoring function to guide the deep generative model REINVENT[5,79] and compare model performance and behaviour to a ligand-based scoring function. The case study involved dopamine receptor D2 (DRD2). The approach taken is depicted in Figure 9, where data sources are coloured blue and scoring functions orange. The REINVENT framework (in grey) consists of two recurrent neural networks, a prior and an agent. The main steps in the current work are (1) removing known DRD2 active molecules from the ZINC training data; (2) training the prior model on druglike molecules from ZINC; (3) initializing the agents as a copy of the prior; (4) preparing the scoring functions to evaluate *de novo* molecules; (5) iteratively training both agents *via* reinforcement learning; and (6) evaluating the structure- and ligand-based approaches with respect to different quantitative, chemical and structural aspects of the generated molecules.

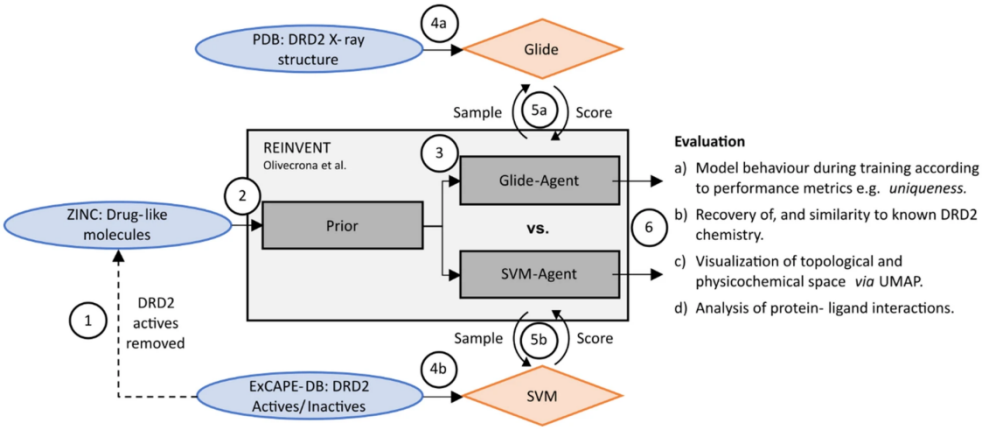


Figure 9. Comparison of structure- and ligand-based scoring functions.

The structure-based approach improved uniqueness and molecular diversity during training, produced higher similarity to the training set and provided a greater coverage of known active ligands than the ligand-based approach, despite having no prior ligand knowledge, as more clusters were shared between Glide-Agent and known actives than were shared between the SVM-Agent results and known actives. Glide-Agent generates high-scoring molecules that are more novel than the SVM-Agent ones and generates more novel areas of physicochemical space, consistent with the prior. Moreover, Glide-Agent learns to satisfy a crucial interaction with D114^{3x32} which is associated with better docking scores and is a prerequisite for experimental affinity.[80]

Unfortunately, docking score optimisation is slow (each run takes about 1 week on about 30 CPUs) and it is system dependent. Can the computational expense associated with model optimization be minimised? The REINVENT loss function (augmented likelihood) includes a value sigma used to scale up the scoring function and lower the prior contribution.[79] Comparison of REINVENT[5] with REINVENT 2[79] shows that sigma variation has a small effect on a short time scale. When rewards are sparse, loss drives agent back towards prior. This can be circumvented by using the hill-climb algorithm[81] to focus learning on the best molecules. Thomas *et al.* found that a hybrid, augmented hill-climb, is more efficient at optimising docking score and is more sensitive to sigma values and hence more tunable. Augmented hill-climb has the propensity to undergo mode collapse (drop in uniqueness).

Mode collapse can be rescued by using a diversity filter (DF)[3] to penalise non-unique or similar molecules. DF stabilises optimisation Augmented hill-climb plus DF is seven times more efficient than Glide-Agent in the original work on the short time scale, and up to 100 times more efficient on the long time scale whilst maintaining similar chemical behaviour, and runtime is reduced to about 2-5 hours compared to one week on about 30 CPUs.

DF would rescue mode collapse but it would not address the issue of generating unrealistic molecules. Benchmarking datasets are either too restrictive (as in the case of MOSES) or too broad (as in the case of GuacaMol) but the $ChEMBL_{potent}$ subset of ChEMBL provides a dataset rich with druglike properties. SMILES outperforms alternative grammars in the prior dataset. Surprisingly, DeepSMILES[82] suffer lower validity. SELFIES[83] are more diverse but fewer of them pass standard druglikeness filters. SELFIES are least like the training set and the use of them results in many more "unusual" compounds. The prior dataset is still generating relatively featureless structures compared with risperidone (a DRD2 inverse agonist), regardless of the chemical grammar.

Thomas' final topic was the effect of scoring function protocol on failures of docking and of QSAR functions. We know, for example in DockStream,[7] that different protocols lead to variable enrichment in docking and that adding constraints such as particular residue interactions increases performance[84] and can outperform ML.[85] It has been observed that ligand protonation is important in docking and that similar chemotypes can have inconsistent docked poses. There is also a trend for certain physicochemical properties to be violated. To study the effect of scoring function protocol, Thomas *et al.* chose MPO against Adenosine 2A (A2A).[86,87] They analysed a diverse range of known chemotypes.[88] They increased prior contribution by decreasing sigma from 60 to 30, protonated only the most likely states, and introduced a more difficult optimisation problem using constrained docking score, retrosynthetic accessibility score (RAscore),[12] TPSA ≥ 40 and number of rotatable bonds ≤ 6. The added constraints worsen docking optimisation but improve molecule quality. The A2A MPO recovered more of, and a wider range of known A2A chemotypes.[88] The added constraints avoid full occupation of the cavity.

As for failures of QSAR functions, we know that generative models can overfit QSAR functions[89] and that QSAR models with similar performance select different prospective candidates in virtual screening.[90] Thomas *et al.* compared the molecules designed *de novo* for three targets using several different molecular representations, QSAR models and generative models. Both descriptor and QSAR method have a significant impact on generative model behaviour, such as molecular diversity and similarity to the training set. The vast majority of molecules are unique to a particular replicate and a particular method. The best way to incorporate synthesizability has not been considered. Work is ongoing on prospective validation, interaction fingerprints and alternative scoring functions.

Machine learning models for predicting human *in vivo* PK parameters using chemical structure and dose

Olga Obrezanova[1], Filip Miljković[2], Anton Martinsson[2], Beth Williamson[1], Martin Johnson[1], Andy Sykes[1], Andreas Bender[1], Nigel Greene[3] (1) AstraZeneca, Cambridge, UK (2) AstraZeneca, Gothenburg, Sweden (3) AstraZeneca, Waltham, MA, USA

Animal and human pharmacokinetic (PK) data are routinely used in drug discovery to understand absorption, disposition, metabolism, and excretion (ADME) of candidate drugs. AstraZeneca has a

suite of over 40 global ADME and safety models to guide virtual compound generation, enable compound selection and prioritisation, design compounds with good ADME and safety profiles, improve speed and efficiency in the DMTA cycle and reduce the number of *in vitro* experiments. The ultimate goal is to enable human PK prediction at the point of design.

Prediction of rat PK is a stepping stone towards modelling human PK. An AI model predicts rat PK parameters from chemical structure and measured *in vitro* ADME properties. The chemical structure is encoded by a graph convolutional neural network (GCN). Properties used as input features are solubility, Caco2 (colorectal adenocarcinoma cell) intrinsic permeability and efflux, intrinsic clearance (CL_{int}) in human liver microsomes (HLM), rat hepatocytes, intrinsic clearance and fraction unbound, and rat and human plasma protein binding (PPB). Properties predicted are clearance (CL), bioavailability (%F, the fraction of an oral administered drug that reaches systemic circulation), C_{max} (the maximum serum concentration that a drug achieves in a specified test area of the body after the drug has been administrated and before the administration of a second dose), $t_{1/2}$ (elimination half-life, the time required for the concentration of the drug in the plasma to reach half of its original value) and V_{ss} (volume of distribution at steady-state).

The method uses message passing NNs for molecular property prediction ("chemprop") from the Machine Learning for Pharmaceutical Discovery and Synthesis (MLPDS) consortium, a collaboration between industry and the Massachusetts Institute of Technology. The rat PK model achieved good accuracy on key PK parameters (Table 2). CL was predicted within 2-fold error for 75% of compounds and within 3-fold for 90% of compounds.

Table 2. Rat PK model: test set performance.

	R^2	RMSE	Experimental variability
CL	0.57	0.28	0.18
%F	0.48	0.72	0.55
V_{ss}	0.50	0.28	0.21

RMSE = root mean square error

The use of *in vitro in vivo* extrapolation (IVIVe) from human hepatocyte and HLM stability assays, typically the "well stirred model" (WSM)[91] is a widely accepted predictive methodology for human metabolic clearance. The rat PK CL prediction results were compared with those of WSM IVIVe. The *in vivo* rat CL model has higher accuracy (RMSE= 0.28, R^2= 0.57 as opposed to RMSE= 0.43, R^2= -0.11) and is not limited by liver blood flow (LBF). It provides insight into potential additional routes of elimination when compared to WSM (which is restricted by LBF). To test if predictions could be made for virtual compounds purely from chemical structure, the researchers went back to "old" *in silico* models for ADME properties for a training set. The test set was made by a 10% temporal split. RMSE results proved that the rat PK models are useful at the point of design.

To build a human PK model,[92] PK data were extracted from PharmaPendium and curated based on expert opinion. The final dataset contained 1001 SMILES and 4491 compound-dose combinations for 12 PK parameters. For each compound−dose combination median values were calculated per PK parameter. Levels of data completeness for each PK parameter varied from 3.5% to 67%. The data are biased towards optimised compounds with good PK profiles. Doses covered a wide range.

Obrezanova *et al.* built on the rat PK model to predict human PK. The feature set consisted of dose, chemical structure, predicted *in vitro* ADME data and *in vivo* rat PK data. Random forest was used as modelling technique. The split into validation and test sets was random by compound with dose stratification. Varying data distributions and data availability had an impact on the ability to model endpoints. Three endpoints had satisfactory models: oral area under the plasma time−concentration curve (AUC PO, R^2_{test}= 0.63; $RMSE_{test}$ = 0.76), maximum plasma drug concentration *per* oral (C_{max} PO, R^2_{test} = 0.68; $RMSE_{test}$ = 0.62), and volume of distribution intravenous (V_d IV, R^2_{test} = 0.47; $RMSE_{test}$ = 0.50).[92] Dose is one of the most important features to model AUC PO and C_{max} PO. Predictions of *in vivo* rat PK parameters and *in vitro* ADME properties are also important.

Performance of the models was additionally investigated using an internal AstraZeneca compendium of first-time-in-human measurements in the 2000−2020 period.[93] In addition, drug metabolism and pharmacokinetics (DMPK) prediction values are provided allowing for a side-by-side performance comparison with machine learning models. Despite the different sample sizes and chemical composition of the hold-out test set and internal clinical candidates, the model performance was comparable for both datasets. The accuracy of the ML models was lower than that of pre-clinical prediction (DMPK). Nevertheless, the ML models are fit-for-purpose to be used in early drug discovery and are complementary to current pre-clinical predictions.

The *in vivo* rat and human PK models increase the efficiency of the DMTA cycle allowing scientists to design compounds with better safety and PK properties early in the drug discovery process. The models can drive prioritisation for *in vivo* testing and reduction in animal experiments and guide *de novo* generative models to build in good PK. They can also inform safety-related models of the therapeutic window: predicted human C_{max} can be used to enable safety risk assessment at earlier stages. In future Obrezanova and her colleagues will expand the in-house and commercial datasets, use dog and rat PK models built on larger datasets to improve the human model, and explore transfer learning and multitask learning deep learning architectures.

References

(1) Arús-Pous, J.; Blaschke, T.; Ulander, S.; Reymond, J.-L.; Chen, H.; Engkvist, O. Exploring the GDB-13 chemical space using deep generative models. *J. Cheminf.* **2019,** *11*, 20.
(2) Arús-Pous, J.; Johansson, S. V.; Prykhodko, O.; Bjerrum, E. J.; Tyrchan, C.; Reymond, J.-L.; Chen, H.; Engkvist, O. Randomized SMILES strings improve the quality of molecular generative models. *J. Cheminf.* **2019,** *11*, 71.
(3) Blaschke, T.; Engkvist, O.; Bajorath, J.; Chen, H. Memory-assisted reinforcement learning for diverse molecular de novo design. *J. Cheminf.* **2020,** *12* (1), 68.
(4) Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* **2020,** *60* (12), 6065-6073.
(5) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminf.* **2017,** *9*, 48.
(6) Papadopoulos, K.; Giblin, K. A.; Janet, J. P.; Patronov, A.; Engkvist, O. De novo design with deep generative models based on 3D similarity scoring. *Bioorg. Med. Chem.* **2021,** *44*, 116308.
(7) Guo, J.; Janet, J. P.; Bauer, M. R.; Nittinger, E.; Giblin, K. A.; Papadopoulos, K.; Voronov, A.; Patronov, A.; Engkvist, O.; Margreittera, C. DockStream: A Docking Wrapper to Enhance De Novo Molecular Design. http://chemrxiv.org/engage/chemrxiv/article-details/6107fc3340c8bd01539a36f4 (accessed November 5, 2021).

(8) 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.
(9) Thakkar, A.; Kogej, T.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chem. Sci.* **2020,** *11* (1), 154-168.
(10) Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J. Cheminf.* **2020,** *12* (1), 70.
(11) Thakkar, A.; Selmi, N.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. "Ring Breaker": Neural Network Driven Synthesis Prediction of the Ring System Chemical Space. *J. Med. Chem.* **2020,** *63,* (16), 8791-8808.
(12) Thakkar, A.; Chadimová, V.; Bjerrum, E. J.; Engkvist, O.; Reymond, J.-L. Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chem. Sci.* **2021,** *12*, 3339-3349.
(13) Green, D. V. S.; Pickett, S.; Luscombe, C.; Senger, S.; Marcus, D.; Meslamani, J.; Brett, D.; Powell, A.; Masson, J. BRADSHAW: a system for automated molecular design. *J. Comput.-Aided Mol. Des.* **2020,** *34* (7), 747-765.
(14) Lewell, X. Q.; Judd, D.; Watson, S.; Hann, M. RECAP-Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998,** *38* (3), 511-522.
(15) Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem* **2008,** *3* (10), 1503-1507.
(16) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010,** *50* (3), 339-348.
(17) Free, S. M., Jr.; Wilson, J. W. A mathematical contribution to structure-activity studies. *J. Med. Chem.* **1964,** *7* (4), 395-399.
(18) Pogany, P.; Arad, N.; Genway, S.; Pickett, S. D. De Novo Molecule Design by Translating from Reduced Graphs to SMILES. *J. Chem. Inf. Model.* **2019,** *59* (3), 1136-1146.
(19) Bush, J. T.; Pogany, P.; Pickett, S. D.; Barker, M.; Baxter, A.; Campos, S.; Cooper, A. W. J.; Hirst, D.; Inglis, G.; Nadin, A.; Patel, V. K.; Poole, D.; Pritchard, J.; Washio, Y.; White, G.; Green, D. V. S. A Turing Test for Molecular Generators. *J. Med. Chem.* **2020,** *63* (20), 11964-11971.
(20) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015,** *137* (7), 2695-2703.
(21) Amabilino, S.; Pogany, P.; Pickett, S. D.; Green, D. V. S. Guidelines for Recurrent Neural Network Transfer Learning-Based Molecular Generation of Focused Libraries. *J. Chem. Inf. Model.* **2020,** *60* (12), 5699-5713.
(22) Young, T. A.; Gheorghe, R.; Duarte, F. cgbind: A Python Module and Web App for Automated Metallocage Construction and Host-Guest Characterization. *J. Chem. Inf. Model.* **2020,** *60* (7), 3546-3557.
(23) Young, T. A.; Silcock, J. J.; Sterling, A. J.; Duarte, F. autodE: Automated Calculation of Reaction Energy Profiles- Application to Organic and Organometallic Reactions. *Angew. Chem., Int. Ed.* **2021,** *60* (8), 4266-4274.
(24) Young, T. A.; Johnston-Wood, T.; Deringer, V. L.; Duarte, F. A transferable active-learning strategy for reactive molecular force fields. *Chem. Sci.* **2021,** *12* (32), 10944-10955.
(25) Bartok, A. P.; Csanyi, G. Gaussian approximation potentials: A brief tutorial introduction. *Int. J. Quantum Chem.* **2015,** *115* (16), 1051-1057.

(26) Bartok, A. P.; Payne, M. C.; Kondor, R.; Csanyi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010,** *104* (13), 136403.
(27) Bernstein, N.; Csányi, G.; Deringer, V. L. De novo exploration and self-guided learning of potential-energy surfaces. *npj Comput. Mater.* **2019,** *5* (1), 99.
(28) Cole, D. J.; Mones, L.; Csanyi, G. A machine learning based intramolecular potential for a flexible organic molecule. *Faraday Discuss.* **2020,** *224*, 247-264.
(29) Deringer, V. L.; Proserpio, D. M.; Csanyi, G.; Pickard, C. J. Data-driven learning and prediction of inorganic crystal structures. *Faraday Discuss.* **2018,** *211*, 45-59.
(30) Mahoney, M. W.; Drineas, P. CUR matrix decompositions for improved data analysis. *Proc. Natl. Acad. Sci. U. S. A.* **2009,** *106* (3), 697-702.
(31) Young, T. A.; Marti-Centelles, V.; Wang, J.; Lusby, P. J.; Duarte, F. Rationalizing the Activity of an "Artificial Diels-Alderase": Establishing Efficient and Accurate Protocols for Calculating Supramolecular Catalysis. *J. Am. Chem. Soc.* **2020,** *142* (3), 1300-1310.
(32) Spicer, R. L.; Stergiou, A. D.; Young, T. A.; Duarte, F.; Symes, M. D.; Lusby, P. J. Host-Guest-Induced Electron Transfer Triggers Radical-Cation Catalysis. *J. Am. Chem. Soc.* **2020,** *142* (5), 2134-2139.
(33) Raies, A. B.; Bajic, V. B. In silico toxicology: comprehensive benchmarking of multi-label classification methods applied to chemical toxicity data. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2018,** *8* (3), e1352.
(34) Sushko, I.; Salmina, E.; Potemkin, V. A.; Poda, G.; Tetko, I. V. ToxAlerts: A Web Server of Structural Alerts for Toxic Chemicals and Compounds with Potential Adverse Reactions. *J. Chem. Inf. Model.* **2012,** *52* (8), 2310-2316.
(35) Lang, A.; Volkamer, A.; Behm, L.; Röblitz, S.; Ehrig, R.; Schneider, M.; Geris, L.; Wichard, J.; Buttgereit, F. In silico methods - Computational alternatives to animal testing. *ALTEX* **2018,** *35* (1), 124-126.
(36) Morger, A.; Mathea, M.; Achenbach, J. H.; Wolf, A.; Buesen, R.; Schleifer, K.-J.; Landsiedel, R.; Volkamer, A. KnowTox: pipeline and case study for confident prediction of potential toxic effects of compounds in early phases of development. *J. Cheminf.* **2020,** *12* (1), 24.
(37) Morger, A.; Svensson, F.; Arvidsson McShane, S.; Gauraha, N.; Norinder, U.; Spjuth, O.; Volkamer, A. Assessing the calibration in toxicological in vitro models with conformal prediction. *J. Cheminf.* **2021,** *13* (1), 35.
(38) Garcia de Lomana, M.; Morger, A.; Norinder, U.; Buesen, R.; Landsiedel, R.; Volkamer, A.; Kirchmair, J.; Mathea, M. ChemBioSim: Enhancing Conformal Prediction of In Vivo Toxicity by Use of Predicted Bioactivities. *J. Chem. Inf. Model.* **2021,** *61* (7), 3255-3272.
(39) Webel, H. E.; Kimber, T. B.; Radetzki, S.; Neuenschwander, M.; Nazare, M.; Volkamer, A. Revealing cytotoxic substructures in molecules using deep learning. *J. Comput.-Aided Mol. Des.* **2020,** *34* (7), 731-746.
(40) Ji, C.; Svensson, F.; Zoufir, A.; Bender, A. eMolTox: prediction of molecular toxicity with confidence. *Bioinformatics* **2018,** *34* (14), 2508-2509.
(41) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998,** *38* (6), 983-996.
(42) Hanser, T.; Barber, C.; Marchaland, J. F.; Werner, S. Applicability domain: towards a more formal definition dol. *SAR QSAR Environ. Res.* **2016,** *27* (11), 865-881.
(43) Norinder, U.; Carlsson, L.; Boyer, S.; Eklund, M. Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination. *J. Chem. Inf. Model.* **2014,** *54* (6), 1596-1603.
(44) Norinder, U.; Rybacka, A.; Andersson, P. L. Conformal prediction to define applicability domain - A case study on predicting ER and AR binding. *SAR QSAR Environ. Res.* **2016,** *27* (4), 303-316.
(45) Svensson, F.; Norinder, U.; Bender, A. Modelling compound cytotoxicity using conformal prediction and PubChem HTS data. *Toxicol. Res. (Cambridge, U. K.)* **2017,** *6* (1), 73-80.

(46) Mervin, L. H.; Cao, Q.; Barrett, I. P.; Firth, M. A.; Murray, D.; McWilliams, L.; Haddrick, M.; Wigglesworth, M.; Engkvist, O.; Bender, A. Understanding Cytotoxicity and Cytostaticity in a High-Throughput Screening Collection. *ACS Chem. Biol.* **2016,** *11* (11), 3007-3023.
(47) Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Mueller, K.-R.; Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One* **2015,** *10* (7), e0130140/1.
(48) Irwin, R.; Dimitriadis, S.; He, J.; Bjerrum, E. Chemformer: A Pre-Trained Transformer for Computational Chemistry. http://chemrxiv.org/engage/chemrxiv/article-details/60ee8a3eb95bdd06d062074b (accessed November 17, 2021).
(49) Matveieva, M.; Polishchuk, P. Benchmarks for interpretation of QSAR models. *J. Cheminf.* **2021,** *13* (1), 41.
(50) Jimenez-Luna, J.; Skalic, M.; Weskamp, N. Benchmarking molecular feature attribution methods with activity cliffs. http://chemrxiv.org/engage/chemrxiv/article-details/613b21fe27d906d4c183cfc1 (accessed November 25, 2021).
(51) Wellawatte, G. P.; Seshadri, A.; White, A. D. Model agnostic generation of counterfactual explanations for molecules. http://chemrxiv.org/engage/chemrxiv/article-details/613268f0d5f0803706ba0c79 (accessed November 25, 2021).
(52) Brown, N.; Fiscato, M.; Segler, M. H. S.; Vaucher, A. C. GuacaMol: Benchmarking Models for de Novo Molecular Design. *J. Chem. Inf. Model.* **2019,** *59* (3), 1096-1108.
(53) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput.-Aided Mol. Des.* **2016,** *30* (8), 595-608.
(54) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019,** *59* (8), 3370-3388.
(55) Zankov, D. V.; Matveieva, M.; Nikonenko, A. V.; Nugmanov, R. I.; Baskin, I. I.; Varnek, A.; Polishchuk, P.; Madzhidov, T. I. QSAR Modeling Based on Conformation Ensembles Using a Multi-Instance Learning Approach. *J. Chem. Inf. Model.* **2021,** *61* (10), 4913-4923.
(56) Nikonenko, A.; Zankov, D.; Baskin, I.; Madzhidov, T.; Polishchuk, P. Multiple Conformer Descriptors for QSAR Modeling. *Mol. Inf.* **2021,** *40* (11), 2060030.
(57) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996,** *39* (15), 2887-2893.
(58) Finnigan, W.; Hepworth, L. J.; Flitsch, S. L.; Turner, N. J. RetroBioCat as a computer-aided synthesis planning tool for biocatalytic reactions and cascades. *Nat. Catal.* **2021,** *4* (2), 98-104.
(59) Kreutter, D.; Schwaller, P.; Reymond, J.-L. Predicting enzymatic reactions with a molecular transformer. *Chem. Sci.* **2021,** *12* (25), 8648-8659.
(60) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019,** *5* (9), 1572-1583.
(61) Probst, D.; Manica, M.; Teukam, Y. G. N.; Castrogiovanni, A.; Paratore, F.; Laino, T. Molecular transformer-aided biocatalysed synthesis planning. http://chemrxiv.org/engage/chemrxiv/article-details/60c75919842e6599a7db4990 (accessed November 26, 2021).
(62) Probst, D.; Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminf.* **2020,** *12*, 12.
(63) Capecchi, A.; Probst, D.; Reymond, J.-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J. Cheminf.* **2020,** *12* (1), 43.
(64) 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.

(65) 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.
(66) Cong, Q.; Anishchenko, I.; Ovchinnikov, S.; Baker, D. Protein interaction networks revealed by proteome coevolution. *Science* **2019,** *365* (6449), 185-189.
(67) Mariani, V.; Biasini, M.; Barbato, A.; Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **2013,** *29* (21), 2722-2728.
(68) Xie, Q.; Luong, M. T.; Hovy, E.; Le, Q. V. Self-Training With Noisy Student Improves ImageNet Classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: 2020; pp 10684-10695.
(69) Vriza, A.; Canaj, A. B.; Vismara, R.; Kershaw Cook, L. J.; Manning, T. D.; Gaultois, M. W.; Wood, P. A.; Kurlin, V.; Berry, N.; Dyer, M. S.; Rosseinsky, M. J. One class classification as a practical approach for accelerating π-π co-crystal discovery. *Chem. Sci.* **2021,** *12* (5), 1702-1719.
(70) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminf.* **2018,** *10*, 4.
(71) Yang, K. K.; Wu, Z.; Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **2019,** *16* (8), 687-694.
(72) Mazurenko, S.; Prokop, Z.; Damborsky, J. Machine Learning in Enzyme Engineering. *ACS Catal.* **2020,** *10* (2), 1210-1223.
(73) Siedhoff, N. E.; Schwaneberg, U.; Davari, M. D. Machine learning-assisted enzyme engineering. *Methods Enzymol.* **2020,** *643*, 281-315.
(74) Wittmann, B. J.; Johnston, K. E.; Wu, Z.; Arnold, F. H. Advances in machine learning for directed evolution. *Curr. Opin. Struct. Biol.* **2021,** *69*, 11-18.
(75) Siedhoff, N. E.; Illig, A.-M.; Schwaneberg, U.; Davari, M. D. PyPEF-An Integrated Framework for Data-Driven Protein Engineering. *J. Chem. Inf. Model.* **2021,** *61* (7), 3463-3476.
(76) Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Johansson, S.; Chen, H.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. 2018, arXiv e-print archive. http://arxiv.org/abs/1811.12823 (accessed January 20, 2021).
(77) Congreve, M.; de Graaf, C.; Swain, N. A.; Tate, C. G. Impact of GPCR Structures on Drug Discovery. *Cell* **2020,** *181* (1), 81-91.
(78) 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.
(79) Blaschke, T.; Arus-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. REINVENT 2.0: An AI Tool for De Novo Drug Design. *J. Chem. Inf. Model.* **2020,** *60* (12), 5918-5922.
(80) Kaczor, A. A.; Silva, A. G.; Loza, M. I.; Kolb, P.; Castro, M.; Poso, A. Structure-Based Virtual Screening for Dopamine D2 Receptor Ligands as Potential Antipsychotics. *ChemMedChem* **2016,** *11* (7), 718-729.
(81) Neil, D.; Segler, M.; Guasch , L.; Ahmed, M.; Plumbley, D.; Sellwood, M.; Brown, N. Exploring deep recurrent models with reinforcement learning for molecule design. http://openreview.net/pdf?id=Bk0xiI1Dz (accessed December 1, 2021).
(82) O'Boyle, N. M.; Dalke, A. DeepSMILES: an adaptation of SMILES for use in machine-learning of chemical structures. http://chemrxiv.org/engage/chemrxiv/article-details/60c73ed6567dfe7e5fec388d (accessed November 30, 2021).
(83) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn.: Sci. Technol.* **2020,** *1* (4), 045024.

(84) 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.
(85) Tran-Nguyen, V.-K.; Bret, G.; Rognan, D. True Accuracy of Fast Scoring Functions to Predict High-Throughput Screening Data from Docking Poses: The Simpler the Better. *J. Chem. Inf. Model.* **2021,** *61* (6), 2788-2797.
(86) 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.
(87) 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 A2AR signaling rescues immune cell function including CD103+ dendritic cells enhancing anti-tumor immunity. *J. Immunother. Cancer* **2020,** *8* (2), e000417.
(88) Weiss, D. R.; Bortolato, A.; Tehan, B.; Mason, J. S. GPCR-Bench: A Benchmarking Set and Practitioners' Guide for G Protein-Coupled Receptor Docking. *J. Chem. Inf. Model.* **2016,** *56* (4), 642-651.
(89) Renz, P.; Van Rompaey, D.; Wegner, J. K.; Hochreiter, S.; Klambauer, G. On failure modes in molecule generation and optimization. *Drug Discovery Today: Technol.* **2019,** *32-33*, 55-63.
(90) Jiang, D.; Wu, Z.; Hsieh, C.-Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T. Could graph neural networks learn better molecular representation for drug discovery A comparison study of descriptor-based and graph-based models. *J. Cheminf.* **2021,** *13* (1), 12.
(91) Yang, J.; Jamei, M.; Yeo, K. R.; Rostami-Hodjegan, A.; Tucker, G. T. Misuse of the well-stirred model of hepatic drug clearance. *Drug Metab. Dispos.* **2007,** *35* (3), 501-502.
(92) Miljkovic, F.; Martinsson, A.; Obrezanova, O.; Williamson, B.; Johnson, M.; Sykes, A.; Bender, A.; Greene, N. Machine Learning Models for Human In Vivo Pharmacokinetic Parameters with In-House Validation. *Mol. Pharmaceutics* **2021,** *18* (12), 4520-4530.
(93) Davies, M.; Jones, R. D. O.; Grime, K.; Jansson-Lofmark, R.; Fretland, A. J.; Winiwarter, S.; Morgan, P.; McGinnity, D. F. Improving the Accuracy of Predicted Human Pharmacokinetics: Lessons Learned from the AstraZeneca Drug Pipeline Over Two Decades. *Trends Pharmacol. Sci.* **2020,** *41* (6), 390-408.