ELSEVIER

feature

Capturing and applying knowledge to guide compound optimisation

Matthew Segall, matt.segall@optibrium.com**, Tamsin Mansley, Peter Hunt and Edmund Champness**

Successful drug discovery requires knowledge and experience across many disciplines, and no current 'artificial intelligence' (AI) method can replace expert scientists. However, computers can recall more information than any individual or team and facilitate the transfer of knowledge across disciplines. Here, we discuss how knowledge relating to chemistry and the biological and physicochemical properties required for a successful compound can be captured. Furthermore, we illustrate how, by combining and applying this knowledge computationally, a broader range of optimisation strategies can be rigorously explored, and the results presented in an intuitive way for consideration by the experts.

Introduction

Drug discovery is a knowledge-driven process. Its success depends on leveraging the knowledge of a multidisciplinary team of scientists, covering fields such as medicinal chemistry, pharmacology, drug metabolism and pharmacokinetics, and toxicology.

Each discipline brings a different viewpoint on the direction of a project. The transfer of this knowledge between scientists is key to balancing the many requirements of an optimal drug candidate; not only activity against therapeutic target(s), but also appropriate absorption, distribution, metabolism, and excretion (ADMET), selectivity over off-targets, and avoiding nonspecific toxicity. Focussing on compounds with a good balance of properties, early in a project, helps to reduce the time taken in lead optimisation and improves the chance of success [1].

Nevertheless, even within our own field, our experience is limited and our memories are finite. There is a wide body of knowledge available outside of our immediate experience and drawing upon this can suggest new optimisation strategies, highlight potential risks that we might not be aware of, or identify additional experiments that would provide valuable information on which to base decisions.

Many databases are available that capture data on a range of chemistry and biological targets. These include public domain sources, such as ChEMBL [2,3], PubChem [4], and the Protein Data Bank (PDB) [5]; and commercial platforms, such as SciFinder [6] and Reaxys [7]. Careful searches of these databases can reveal a wealth of valuable information for drug discovery.

However, to gain the most from these data sources a user is required who knows how best to construct a search to unearth the most relevant data and analyse the results to extract the knowledge these data reveal. This process is also limited by the speed with which we can compose, run and analyse the results of a search.

For this reason, computational methods that can draw on these databases, analyse the underlying data, and apply the resulting knowledge automatically can provide a huge advantage. Systematic analysis of the raw data, at scale, can extract patterns and trends that, in turn, can be applied much faster than a 'manual' process.

Computational methods are also unbiased in their application of knowledge, sometimes revealing examples that do not agree with our intuition and can help to explore 'outside the box'. Like all people, scientists have well-established 'cognitive biases' that can subconsciously influence the decisions we make [8], leading to missed opportunities or wasted effort. For example, 'confirmation bias' leads us to search for data that confirm our hypotheses and neglect those that do not conform, inappropriately limiting the search before choosing a lead series or candidate.

New computational methods for capturing and sharing knowledge can supplement the

https://doi.org/10.1016/j.drudis.2019.02.004

hard-earned experience of teams to guide the direction of research and the optimisation of compounds. This combination of speed and objectivity enables a more rigorous exploration of potential options, subject to a scientific team's strategic oversight of a project.

Here, we review methods for capturing and applying knowledge in chemistry, structure–activity relationships (SAR), and the therapeutic objectives of a drug discovery project. We also present an example illustrating an application to generate new compound ideas for potent, selective compounds with good physicochemical properties.

Capturing and applying knowledge

Chemistry

From a chemistry perspective, the knowledge that comes with experience provides insight into what compounds we might be able to synthesise and how this could be achieved.

For example, a database of chemical building blocks, whether in-house or from a commercial supplier, can provide input to enumerate a virtual library of compounds and explore hypotheses for optimisation. Although limiting the design scope based on the availability of building blocks might offer short-term benefits, other approaches to compound design can offer a broader search of potential optimisation strategies that might be outside of an individual chemist's, or even a project team's, experience, leading to greater potential for long-term success for the project.

One such approach is based on generating new, relevant compound structures by applying structural transformations to existing compounds of interest [9,10]. These transformations represent optimisation steps that have been previously applied in chemistry projects and might be more general than a single reaction step, enabling a broader exploration of chemical space. Simple transformations can be generated by matched molecular pair analysis (MMPA) [11] of existing compound collections, whereas larger transformations can be curated in databases, such as BIOSTER [12], from practical examples published in the chemistry literature.

A recent development in AI methods has resulted in approaches that learn what a drug molecule 'looks like' from large databases of compounds that have been considered in medicinal chemistry projects. These 'generative adversarial networks', such as ORGANIC [13] and RANC [14], can propose new structures for consideration. These approaches are new, and anecdotal evidence suggests that many of the structures generated may not be 'desirable' from a medicinal chemistry perspective; however, the field is developing quickly, and this is likely to improve.

Whenever proposing novel compounds, the first question is whether they are likely to be synthetically tractable and, if so, what routes might be considered. Synthetic tractability is subjective and chemists' opinions and the routes they propose will depend on the reaction schemes with which they are most familiar. This, itself, is a form of cognitive bias, known as

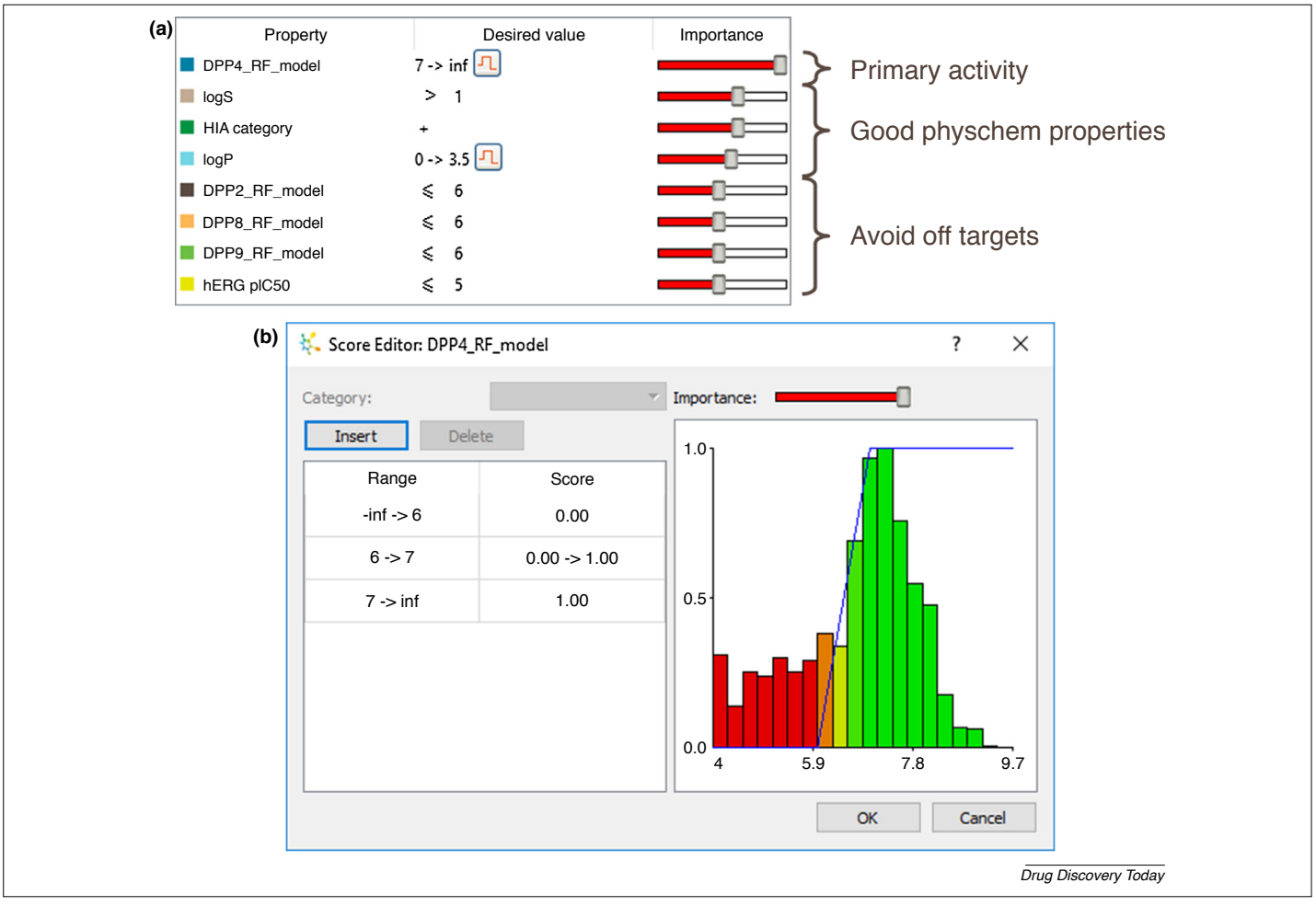

FIGURE 1

Example multi-parameter optimisation scoring profile. (a) Example of a multiparameter optimisation scoring profile that balances optimising primary activity for dipeptidyl peptidase-4 (DPP-4) with selectivity at off-target receptors and desirable physicochemical properties. (b) Desirability function for DPP-4 activity.

familiarity bias. However, computational methods can also capture chemistry knowledge about reactions from the broad body of chemistry literature or from electronic laboratory notebooks and propose synthetic routes for consideration [15]. Expert systems, such as IC *SYNTH* [16,17] and ChemPlanner [18], apply reaction rules to perform a retrosynthetic analysis of a query compound, rank multiple potential synthetic routes, and provide references to support the pathways proposed. Here, again, developments in AI might also have a role to 'learn' from reaction data to predict synthetic routes for novel compounds [19].

Other approaches to propose novel compound suggestions combine the generation of the structure with a potential synthetic route. Software, such as Design of Genuine Structures (DOGS) [20], works by exploring a space of available reagents and reactions and then combining the reagents in multiple steps to generate new compounds. This approach limits the space of ideas that can be searched to those reactions and reagents that are readily available, similar to library enumeration. However, such a computational approach can explore a larger space of possibilities than any semimanual library enumeration, and the suggestion of a plausible synthetic scheme is a benefit, even if it cannot be guaranteed to be successful.

Structure–activity relationships

Of course, any of these methods for compound idea generation can quickly create an overwhelmingly large number of potential compounds, so it is essential to prioritise those that are most likely to have the desired properties. This involves capturing another form of knowledge, namely the relationship between the structure of a compound and its activities.

Drug discovery scientists often have a good understanding of SAR within their series and the practice of generating quantitative SAR (QSAR) models is well established [21,22]. QSAR models capture statistical relationships between the structural features of compounds and their biological activities or ADMET properties using machine-learning algorithms. Thus captured, the SAR can be readily applied to new compounds to predict these endpoints, guiding the design and selection of compounds.

It is also important to recognise the limits of QSAR models; they can only make confident predictions for new compounds that are similar to those used to train the models. Outside of this so-called 'domain of applicability', the predictions of a model should be used with caution, because they are not supported by the data with which the model was built. However, recent developments in 'active learning' use machine-learning methods to propose the most informative compounds for experimental investigation, such that the resulting data can be used to iteratively retrain a QSAR model to optimally improve its accuracy and expand its domain of applicability [23,24]. Active learning strategies balance exploration (selection of compounds with the highest information content) with exploitation (selection of compounds predicted to be best) to most efficiently identify the best compounds in a large chemical space [25,26].

SAR can also be captured through statistical analyses of existing compounds and data by MMPA or matched series analysis. These approaches look for consistent correlations in activity or property values between compounds that differ in only one small substitution at a single location. Although these analyses tend to be qualitative, such consistent changes can indicate new substitutions that are likely to offer improvements [27,28].

Of course, where structural information is available for therapeutic and off-target proteins, this can also be used to guide the design of new compounds, using docking or other structure-based design methods, such as free energy perturbation [29].

Project objectives

The domain knowledge of multidisciplinary experts in a project team guides the optimisation

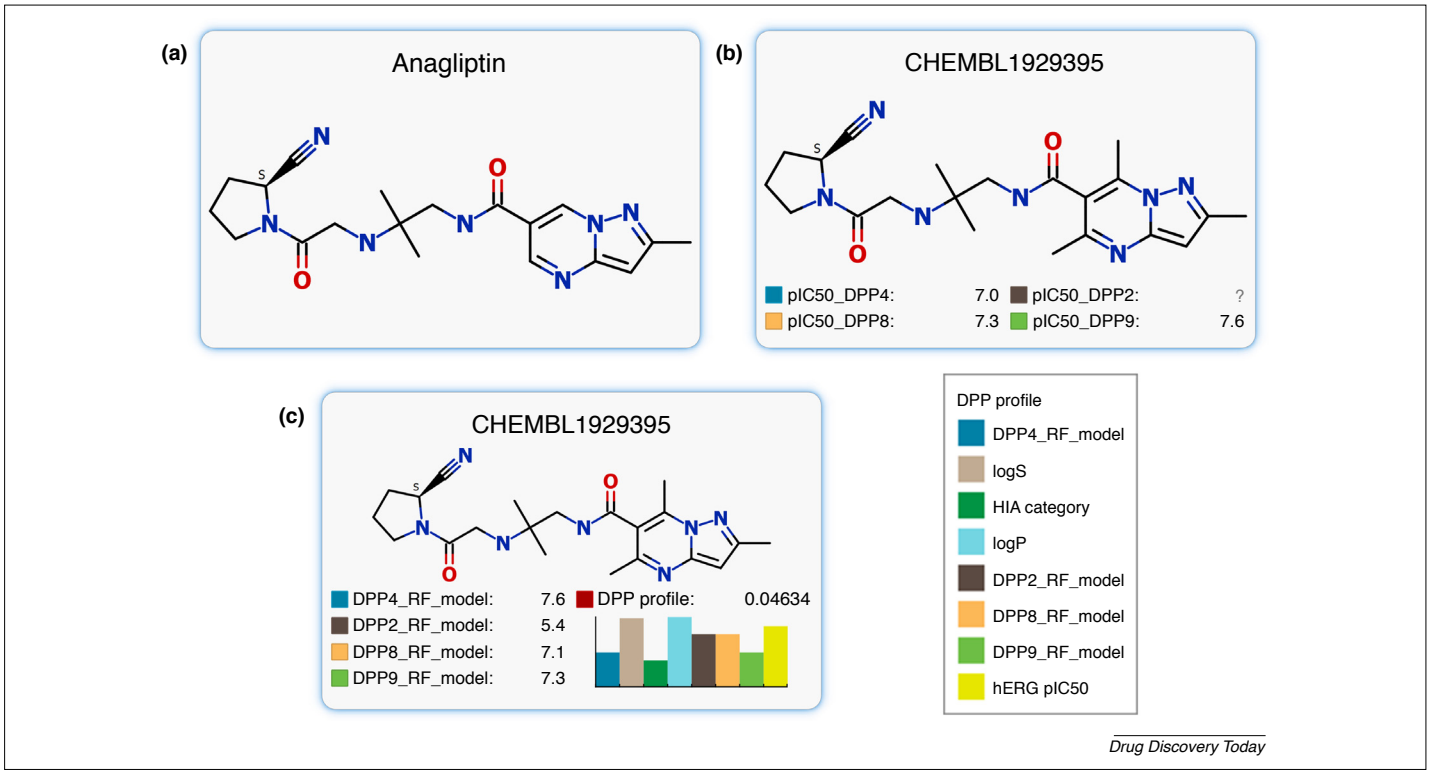

FIGURE 2

Structure of anagliptin and structure and data for an analogue CHEMBL12929395. (a) Structure of anagliptin. (b) ChEMBL activity data for CHEMBL1929395. (c) Predicted activities and multiparameter optimisation score (0.046) for CHEMBL1929395.

of high-quality compounds; what are the requirements for the activity, selectivity, and ADMET properties of an ideal compound? These can be captured as a multiparameter profile of property criteria, as illustrated in the simple example in Fig. 1a, which combines activity against a primary target with selectivity against off-targets and suitable physicochemical properties.

When defining such a profile, it is important to recognise that all criteria might not be equally important; it might be appropriate to trade off one less-important property to achieve a good outcome for a crucial factor, such as primary activity. Furthermore, hard property cut-offs might draw inappropriately harsh distinctions between similar compounds; for example, is the risk associated with a logP of 5.1 really different from that for 4.9? Both the importance of a property and more subtle relationships between property values and their associated risks can be captured through the use of 'desirability functions', as also illustrated in Fig. 1b [30].

Our ability to discriminate between compounds is further limited by the uncertainties in all data we obtain in drug discovery, because of experimental variability or statistical errors in predictive models. When applying multiparameter optimisation (MPO) to prioritise compounds, it is important to take these uncertainties into account to avoid inappropriately rejecting compounds and missing opportunities to identify a high-quality candidate [31].

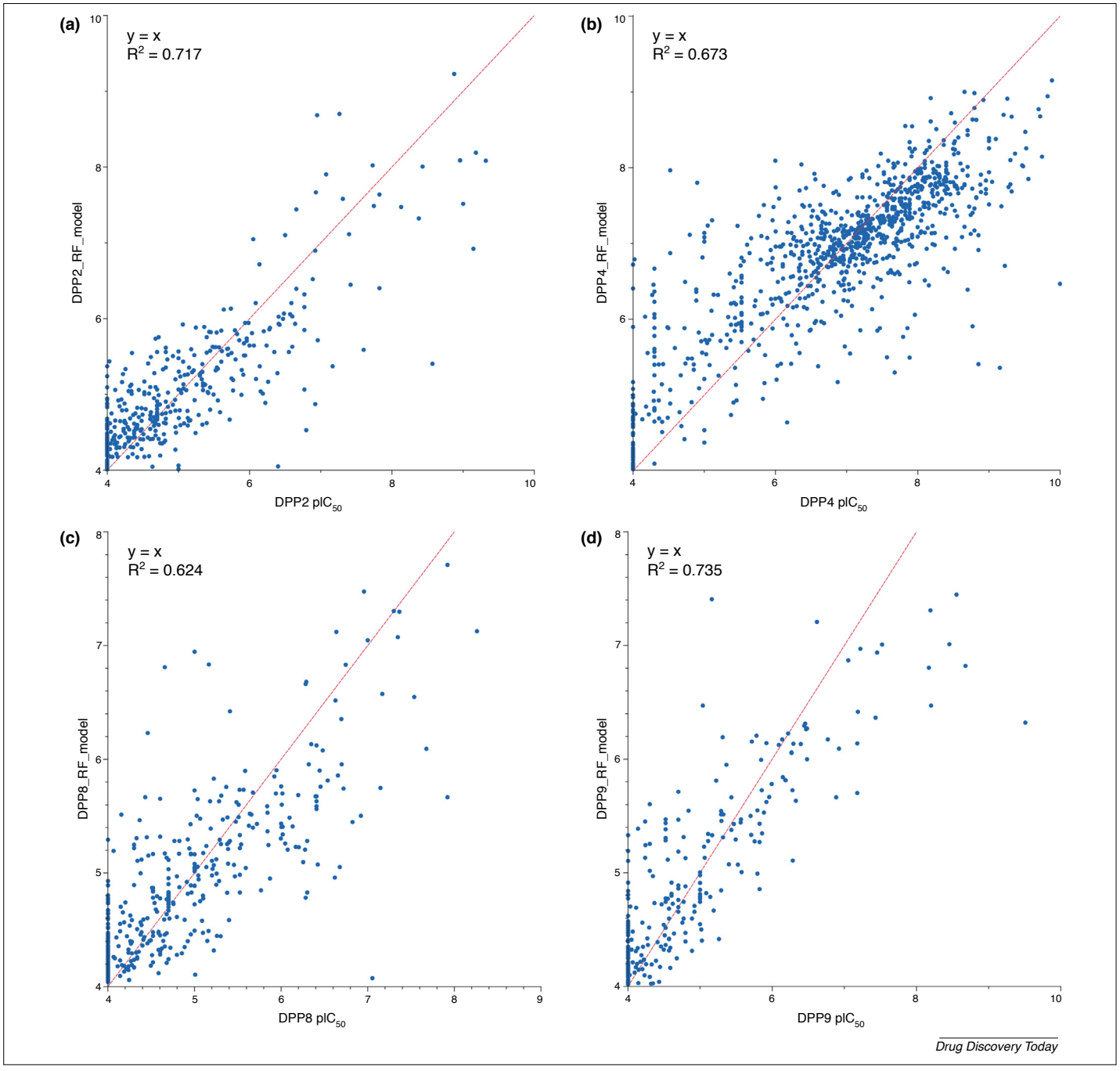

FIGURE 3

Validation of independent test sets of Random Forests quantitative structure–activity relationship (QSAR) models built from pIC_{50} data in ChEMBL to predict activity of compounds for dipeptidyl peptidase (DPP)-2 (a), DPP-4 (b), DPP-8 (c), and DPP-9 (d) receptors.

The property criteria for selection of compounds are, to some extent, subjective and different experts' opinions can vary, particularly given their alternative perspectives. Therefore, it is also useful to consider the sensitivity of the choice of compounds to the selection criteria and their importance [32]. Sometimes, the selection of compounds can change significantly with only a small change in a property criterion. In these cases, it is important to carefully consider the choice of criterion because an inappropriate value might, again, result in missed opportunities.

Putting it all together: example application to the optimisation of high-quality, selective anagliptin analogues

Computational methods to capture and apply knowledge about chemistry, SAR, and the objectives of a project can be combined to rigorously explore optimisation strategies and suggest new compound ideas for expert consideration. We illustrate this with an example application to the *in silico* optimisation of analogues of anagliptin, by applying medicinal chemistry transformations [10], guided by QSAR models and MPO using the Probabilistic Scoring method [31].

Anagliptin (Fig. 2a) is a member of the 'gliptin' family of dipeptidyl peptidase-4 (DPP-4) inhibitors. It has been approved in Japan since 2012 [33] for the treatment of type 2 diabetes mellitus. However, it was withdrawn from other markets because of observed animal toxicity [34], raising concerns, that have since been disproved, about selectivity over the DPP-2, DPP-8, and DPP-9 receptors.

Examination of anagliptin co-crystallised in the DPP-4 binding site [35] shows the cyanomethyl pyrrolidine moiety binding deep in the binding pocket. Lack of differentiation between the DPP-4, DPP-2, DPP-8, and DPP-9 binding pockets suggests that a structure-based approach to optimising selectivity would be challenging and, hence, a ligand-based approach using QSAR modelling might provide an alternative.

Capturing SAR for DPP activities

To prioritise molecules suggested by *in silico de novo* design, we need to predict potencies of the virtual compound ideas that are generated. Searching the ChEMBL database [2,3] enables data sets containing structurally diverse compounds with measured pIC_{50} data to be prepared, which in turn can be used to build QSAR models for the DPP-2, DPP-4, DPP-8, and DPP-9 receptors.

Applying multiple machine-learning approaches available in StarDrop's Auto-Modeller™ module [36], QSAR models were built and validated using a variety of 2D SMARTS descriptors in addition to a range of whole-molecule properties. Figure 3 illustrates the validation of the models for the four receptors, which were built with the random forests method.

Capturing the MPO objectives

As discussed earlier, a successful drug candidate will have a balance of properties in addition to potency. The goals for this project include: (i) good potency for DPP-4 ($pIC_{50} > 7$); (ii) a balance of desirable physicochemical properties typically exhibited by an orally bioavailable drug; and (iii) avoidance of off-target effects, including potential cardiotoxicity because of hERG inhibition, and selectivity over the other DPP receptors.

The Probabilistic Scoring method [31] was used to score and prioritise new compound ideas from the *de novo* design process against the profile of property criteria shown in Fig. 1a, to identify those with the highest chance of success for meeting the goals of the project.

CHEMBL1929395 (Fig. 2b) is a close structural analogue of anagliptin, but exhibits no selectivity for DPP-4 over DPP-8 and DPP-9. In addition to being nonselective, its predicted ADMET properties do not meet the desired criteria. The resulting low MPO score (0.046) makes it an ideal starting point to explore simultaneous, MPO of potency, selectivity, and ADMET.

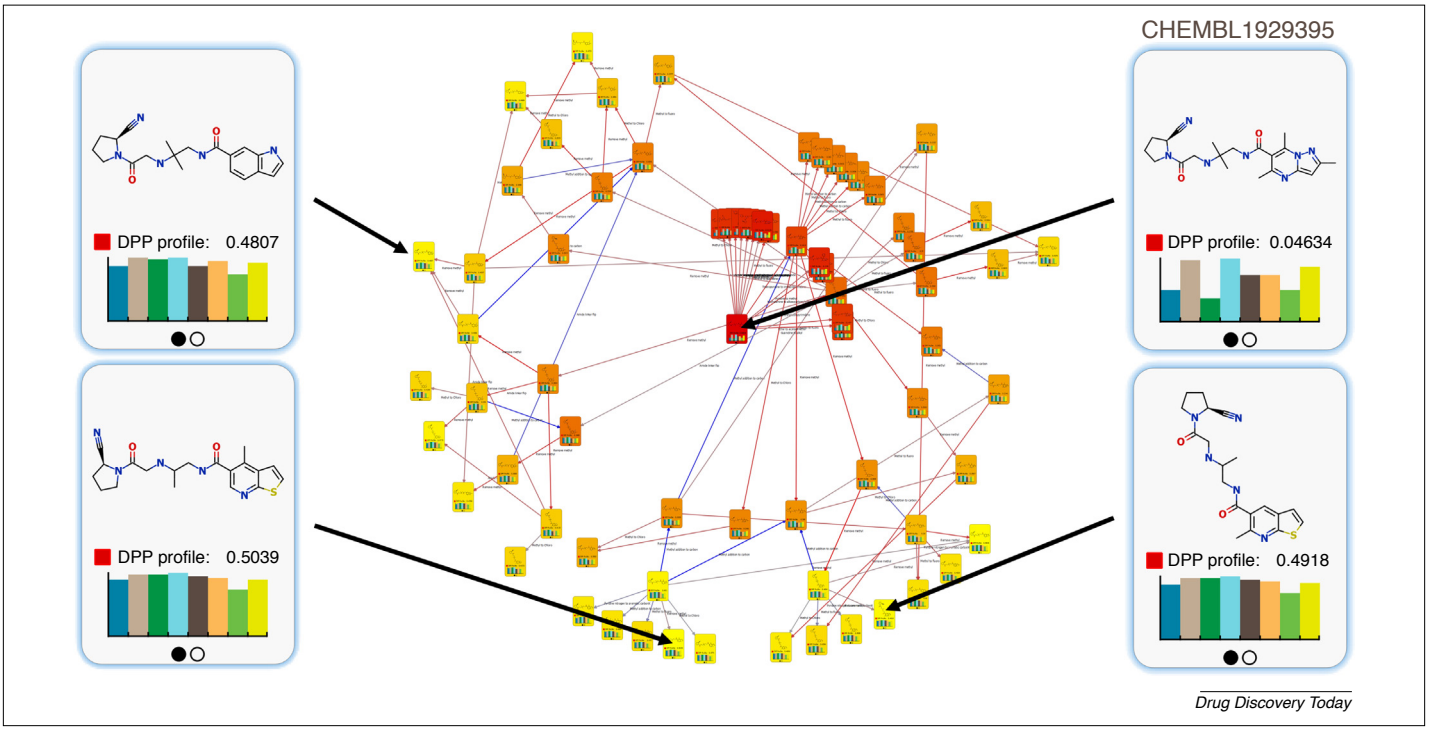

FIGURE 4

Output from the *de novo* design and multiparameter optimisation process. Compounds are coloured by score, with low-scoring compounds coloured red and high-scoring coloured yellow. CHEMBL1929395 and 3 representative high-scoring ideas are highlighted.

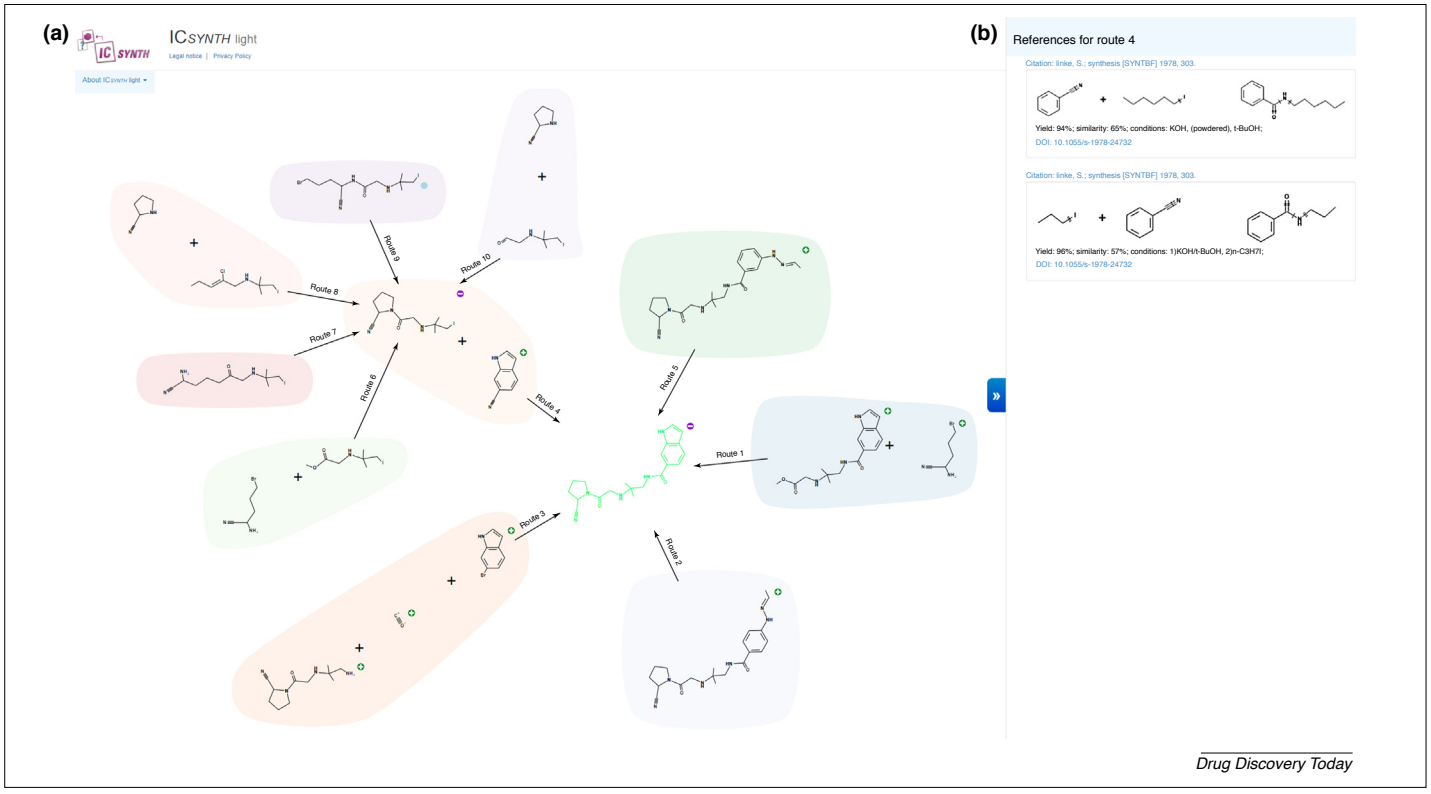

FIGURE 5

Five suggested synthetic strategies towards one of the dipeptidyl peptidase-4 (DPP-4) virtual idea compounds, as output from IC*SYNTH* (a). Each 'route' might require additional steps, such as use of protecting groups, and is supported by references to the primary literature (b).

De novo design: guided optimisation by applying captured knowledge

Starting from CHEMBL1929395 as an initial seed and applying the transformation-based method [10] in StarDrop's Nova™ module, 239 functional group and framework transformations were applied over four generations. During this process, the cyanopyrrolidine moiety of CHEMBL 1929395, known to interact favourably with the protein-binding site, was conserved. To prevent an exponential explosion in the number of ideas generated, the process was focussed by selecting the 20 highest-scoring compounds in each generation as the basis for the next.

Results

In total, over 16 000 compound ideas were evaluated by this process and Fig. 4 illustrates the progression of the *de novo* design process from the low-scoring (0.046, coloured red) initial compound in the centre, CHEMBL1929395, to higher scoring ideas (~0.50, coloured yellow) in the fourth generation. The higher bars in the scoring histograms for the designed compounds indicate that the requirements of each property in the scoring profile are likely to be satisfied.

Three representative examples of the designed compounds are highlighted to illustrate how replacing the pyrazolo(1,5-a)pyrimidine heterocycle and modifying the linker can produce compounds that are predicted to be active at DPP-4, selective over DPP-2, DPP-8, and DPP-9, and to have the desired profile of physicochemical and ADMET properties. Some of the suggested compounds explore areas of chemical space occupied by other known DPP-4 inhibitors, whereas others lie in areas close to CHEMBL1929395 that were previously unexplored by compounds reported in ChEMBL.

The compounds generated by this *de novo* design process were submitted to IC*SYNTH* [16], InfoChem's synthesis planning tool that suggests potential routes for synthesis of a compound, together with the literature precedence for each putative route. Figure 5 illustrates proposed synthetic pathways for one of the compounds highlighted above, providing the confidence that the compounds suggested by the *de novo* design process are reasonable structures and likely to be synthesisable. The logical next step would be to synthesise these idea compounds to further validate the hypothesis, as has been carried out successfully in similar studies; one example being that by Hopkins *et al.* [37] for the optimisation of D4 receptor activity starting from the acetylcholinesterase inhibitor donepezil.

Concluding remarks

Here, we have discussed a how a variety of computational approaches can capture a range of knowledge, relevant to chemistry optimisation projects, and apply this to guide the design of new compounds. We have illustrated how this can be seamlessly integrated to aid the exploration of a range of optimisation strategies, focussing quickly on high-quality, synthetically accessible compounds for expert consideration.

For the foreseeable future, computational algorithms will not replace an expert medicinal chemist. However, an expert's ability can be supplemented by the latest computational methods if these are made accessible in an intuitive way, to offer the best outcome. The expert can define and guide the overall strategy, whereas computational methods enable a rigorous tactical analysis of the available options.

References

1 Segall, M. *et al.* (2006) Focus on success: using in silico optimisation to achieve an optimal balance of properties. *Expert Opin. Drug Metab. Toxicol.* 2, 325–337
2 Gaulton, A. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, D1100–D1107
3 Bento, A. *et al.* (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* 42, 1083–1090

4 Kim, S. *et al.* (2016) PubChem substance and compound databases. *Nucleic Acids Res.* 44, D1202–1213
5 Berman, H. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242
6 Scifinder. https://www.cas.org/products/scifinder [Accessed 14 February 2019]
7 Reaxys. www.elsevier.com/solutions/reaxys [Accessed 14 February 2019]
8 Chadwick, A. and Segall, M. (2010) Overcoming psychological barriers to good discovery decisions. *Drug Discov. Today* 15, 561–569
9 Stewart, K. *et al.* (2006) Drug Guru: a computer software program for drug design using medicinal chemistry rules. *Bioorg. Med. Chem.* 14, 7011–7022
10 Segall, M. *et al.* (2011) Applying medicinal chemistry transformations to guide the search for high quality leads and candidates. *J. Chem. Inf. Model.* 51, 2967–2976
11 Dossetter, A. *et al.* (2013) Matched molecular pair analysis in drug discovery. *Drug Discov. Today* 18, 724–731
12 Ujváry, I. and Hayward, J. (2012) BIOSTER: a database of bioisosteres and bioanalogues. In *Bioisosteres in Medicinal Chemistry* (Brown, N., ed.), pp. 55–74, Wiley-VCH
13 Putin, E. *et al.* (2018) Reinforced adversarial neural computer for de novo molecular design. *J. Chem. Inf. Model.* 56, 1194–1204
14 Popova, M. *et al.* (2018) Deep reinforcement learning for de novo drug design. *Sci. Adv.* 4, eaap7885
15 Szymkuć, S. *et al.* (2016) Computer-assisted synthetic planning: the end of the beginning. *Angew. Chem. Int. Ed.* 55, 5904–5937
16 Bøgevig, A. *et al.* (2015) Route design in the 21st century: the ICSYNTH software tool as an idea generator for synthesis prediction. *Org. Process Res. Dev.* 19, 357–368
17 IC*SYNTH*. www.infochem.de/products/software/icsynth.shtml [Accessed 14 February 2019]
18 ChemPlanner. www.cas.org/products/scifinder-n/chemplanner [Accessed 14 February 2019]
19 Segler, M. *et al.* (2018) Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* 555, 604–610
20 Hartenfeller, M. *et al.* (2012) DOGS: reaction-driven de novo design of bioactive compounds. *PLoS Comput. Biol.* 8, e1002380
21 Tropsha, A. (2010) Best practices for QSAR model development, validation, and exploitation. *Mol. Inform.* 29, 476–488
22 Gola, J. *et al.* (2006) ADMET property prediction: The state of the art and current challenges. *QSAR Comb. Sci.* 25, 1127–1129
23 Lang, T. *et al.* (2016) Feasibility of active machine learning for multiclass compound classification. *J. Chem. Inf. Model.* 56, 12–20
24 Reker, D. and Schneider, G. (2015) Active-learning strategies in computer-assisted drug discovery. *Drug Discov. Today* 20, 458–465
25 Reker, D. *et al.* (2017) Active learning for computational chemogenomics. *Future Med. Chem.* 9, 381–402
26 Pyzer-Knapp, E. (2018) Bayesian optimization for accelerated drug discovery. *IBM J. Res. Dev.* Published online November 19, 2018. https://doi.org/10.1147/JRD.2018.2881731
27 O'Boyle, N. *et al.* (2014) Using matched molecular series as a predictive tool to optimize biologial activity. *J. Med. Chem.* 57, 2944–2951
28 Kramer, C. *et al.* (2018) Learning medicinal chemistry absorption, distribution, metabolism, excretion, and toxicity (ADMET) rules from cross-company matched molecular pairs analysis (MMPA). *J. Med. Chem.* 61, 3277–3292
29 Wang, L. *et al.* (2017) 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.* 137, 2695–2703
30 Wager, T. *et al.* (2010) Moving beyond rules: the development of a central nervous system multiparameter optimization (CNS MPO) approach to enable alignment of druglike properties. *ACS Chem. Neurosci.* 1, 435–449
31 Segall, M. (2012) Multi-parameter optimization: identifying high quality compounds with a balance of properties. *Curr. Pharm. Des.* 18, 1292–1310
32 Segall, M. *et al.* (2016) Avoiding missed opportunities by analyzing the sensitivity of our decisions. *J. Med. Chem.* 59, 4267–4277
33 Graul, A. *et al.* (2013) 2012 in review – part 1, the year's new drugs and biologics. *Drugs Today* 49, 33–68
34 Kushwaha, R. and Wahajul, H. (2014) Sixteen-years of clinically relevant dipeptidyl peptidase-IV (DPP-IV) inhibitors for treatment of type-2 diabetes: a perspective. *Curr. Med. Chem.* 21, 4013–4045
35 Watanabe, Y. *et al.* (2015) Anagliptin, a potent dipeptidyl peptidase IV inhibitor: its single-crystal structure and enzyme interactions. *J. Enzyme Inhib. Med. Chem.* 30, 981–988
36 Obrezanova, O. *et al.* (2008) Automatic QSAR modeling of ADME properties: blood–brain barrier penetration and aqueous solubility. *J. Comput. Aided Mol. Des.* 22, 431–440
37 Hopkins, A. *et al.* (2012) Automated design of ligands to polypharmacological profiles. *Nature* 492, 215–220

Matthew Segall*
Tamsin Mansley
Peter Hunt
Edmund Champness

Optibrium Limited, F5-6 Blenheim House, Cambridge Innovation Park, Denny End Road, Cambridge, CB25 9PB, UK

**Corresponding author.*