J Comput Aided Mol Des (2017) 31:275–285
DOI 10.1007/s10822-016-9975-3

# Computer-aided drug design at Boehringer Ingelheim

Ingo Muegge[1] · Andreas Bergner[2] · Jan M. Kriegl[3]

Received: 15 July 2016 / Accepted: 15 September 2016 / Published online: 20 September 2016

**Abstract** Computer-Aided Drug Design (CADD) is an integral part of the drug discovery endeavor at Boehringer Ingelheim (BI). CADD contributes to the evaluation of new therapeutic concepts, identifies small molecule starting points for drug discovery, and develops strategies for optimizing hit and lead compounds. The CADD scientists at BI benefit from the global use and development of both software platforms and computational services. A number of computational techniques developed in-house have significantly changed the way early drug discovery is carried out at BI. In particular, virtual screening in vast chemical spaces, which can be accessed by combinatorial chemistry, has added a new option for the identification of hits in many projects. Recently, a new framework has been implemented allowing fast, interactive predictions of relevant on and off target endpoints and other optimization parameters. In addition to the introduction of this new framework at BI, CADD has been focusing on the enablement of medicinal chemists to independently perform an increasing amount of molecular modeling and design work. This is made possible through the deployment of MOE as a global modeling platform, allowing computational and medicinal chemists to freely share ideas and modeling results. Furthermore, a central communication layer called the computational chemistry framework provides broad access to predictive models and other computational services.

**Keywords** Computational chemistry · Molecular modeling · Predictive modeling · Chemoinformatics · Virtual screening

## Introduction

Computer-Aided Drug Design and Computational Chemistry (here termed CADD) are an integral component of drug discovery programs at multiple Boehringer Ingelheim (BI) research sites. The CADD scientists at BI work across different therapeutic areas at their sites in close proximity to the medicinal chemists (including combinatorial chemistry). The CADD groups at BI contribute to individual drug discovery projects employing a multitude of approaches extending from chemoinformatics to molecular modeling. These approaches include most aspects of structure-based and ligand-based drug design, predictive modeling, as well as the prioritization and analysis of compound selections through virtual screening, triaging of screening hit sets, and the design of combinatorial library screening decks. The application of these techniques ranges from target selection to lead discovery and optimization including toxicity predictions. In addition, providing CADD technology and encouraging the uptake of decision supporting solutions by project

✉ Ingo Muegge
ingo.mugge@boehringer-ingelheim.com

✉ Andreas Bergner
andreas.bergner@boehringer-ingelheim.com

✉ Jan M. Kriegl
jan.kriegl@boehringer-ingelheim.com

[1] Department of Small Molecule Discovery Research, Boehringer Ingelheim Pharmaceuticals, 900 Ridgebury Road, Ridgefield, CT 06877-0368, USA

[2] Department of Medicinal Chemistry, Boehringer Ingelheim RCV GmbH & Co KG, 1121 Vienna, Austria

[3] Department of Lead Identification and Optimization Support, Boehringer Ingelheim Pharma GmbH & Co. KG, 88397 Biberach, Germany

teams is a driving force for CADD activities at BI. Tasks related to bioinformatics, such as pathway or gene data analyses, are typically the remit of the computational biology groups. With very few exceptions, the CADD work at BI focuses on small molecule drug discovery, although in some cases biologics research has been supported [1, 2].

Although the CADD groups mostly support site-specific projects at the three main BI research sites in Ridgefield, Biberach and Vienna, we have also been implementing a global concept for developing key strategies, best practices, sharing of workflows, protocols and software solutions across all sites. We will illustrate the synergy that can be gained from this growing global focus using examples taken from the recently established Computational Chemistry Framework (CCFW), an in-house global virtual screening platform for designing libraries for lead identification, and a global infrastructure for deploying numerous predictive models. In addition, we will illustrate how the CADD scientists contribute to the advancement of projects, interact with medicinal chemists and develop technology that impacts project decisions.

## The roles of CADD in drug discovery

Usually, CADD scientists at BI join research project teams at the stage of hit identification. A CADD scientist fulfills different roles within a drug discovery project team, which are categorized here as project contributor, data analyst and technical enabler. In the major role as project contributor, a CADD scientist applies rational computational chemistry approaches to finding novel compounds with improved overall property profiles. To this end, a CADD scientist contributes to devising and executing hit finding plans, often including virtual screening or *de novo* design, analyzing screening hit sets, and finding and designing hit analogs, as well as target deconvolution and the identification of target-ligand associations in phenotypic screening campaigns. Hypotheses and models are generated that guide further compound optimization and trigger novel design ideas. During lead optimization, CADD influences or guides the team direction by solving project-specific problems that are often associated with target-independent parameters ranging from target selectivity to PK, tolerability and safety related issues. As more and more unprecedented targets are being pursued in drug discovery, a number of factors help to inform target selection decisions or to design research plans for target enablement. The preferred modes of action (allosteric or orthosteric target modulators), target drug ability and opportunities to identify potential tool compounds are typical examples of CADD scientists getting involved early on before hit identification. All of these tasks are part of the standard repertoire of every CADD scientist at BI.

CADD scientists collaborate very closely with scientists from other experimental disciplines (i.e., screening and profiling units or protein crystallography, for structure-enabled projects) and the medicinal chemists. BI encourages project team members to compete for generating the best design ideas. All design proposals are collected and shared. CADD scientists work very closely with individual chemists to understand their objectives, plans, strategies, timelines and synthesis conditions. Particularly in relation to structure-based projects, the CADD scientists hold regular 3D design and brainstorming sessions with the entire project team to interactively generate new synthesis suggestions from within the team. The team then comes together and decides which compounds will be synthesized next. This decision is based on a transparent and data driven analysis that is independent of the source of the idea. Open, transparent communication, sharing of ideas and team spirit are essential for entering into a productive competition of ideas capable of inspiring even better designs.

The second role of a CADD scientist is to act as a data analyst. Within the context of CADD a data analyst is specialized in transforming the relevant experimental data into hypotheses, which are in turn used to drive the discovery and optimization of compounds. This role is supported by the knowledge of the existence, availability, content, integrity [3] and architecture of internal and external data sources, as well as by the expertise in accessing, processing, and analyzing the data. Pipelining tools such as Pipeline Pilot [4] or KNIME [5] are key to extracting, combining and pre-processing data before they are subjected to more sophisticated computational analyses, such as principal component analysis, machine learning or clustering. Very often, the role of a data analyst also includes project team support by compiling the project-relevant data, e.g. for SAR analysis. Additional input to data-driven decision making within project teams is provided by analyzing cross-project data, such as the results from previous screening campaigns for the identification of potential off-targets or the triaging of hit sets [6].

The third, and increasing, role of a CADD scientist is to enable medicinal chemists to utilize the computer-aided drug design tools on their own. The medicinal chemists at BI design compounds in collaboration with computational chemists and use some CADD design tools independently. Medicinal chemists have been trained in the use of certain CADD tools and have gained an adequate understanding of the methodologies involved. This stimulates discussions about novel computational tools for compound design and simplifies the alignment between CADD and medicinal chemistry. CADD tools in the hands of the medicinal chemist enable a rapid iteration of design ideas within the context of the constraints imposed by synthetic accessibility and compound profile requirements, leading to overall faster design cycles. Enabling the medicinal chemists in this way frees up time for the CADD scientists, who can then concentrate on more demanding design or analysis tasks or new tool development.

The CADD scientists at BI are permanent or temporary project team members contributing to projects beyond their three core roles described above. Where appropriate, they can also lead the chemistry component of early drug discovery projects in the exploratory and hit identification phases. Due to the close collaboration with structural biologists and assay scientists, as well as the focus on devising hit finding plans tailored to the individual projects, experienced CADD scientists are well suited to this role.

CADD adds value to a project when it drives or influences the decisions taken by a project team, and when it facilitates faster decisions. While it is straightforward to describe the qualitative indicators for value added, a quantitative assessment is much more difficult. Different metrics for CADD performance have been discussed, including categorizing and counting the CADD contributions to individual projects [7] or quantifying the quality of CADD work (e.g., agreement of computations with experimental data [8]). Another way of measuring performance is, of course, the assessment of the customer (project team) satisfaction. At BI we collect internal feedback from project team leaders and other key partners. In our experience, a direct dialogue about the mutual expectations and the subsequent impact of CADD on a project is an appropriate way of assessing the added value. These discussions occur on an ongoing basis throughout the year to ensure an optimal alignment of CADD work and project needs within the context of the project portfolio.

In addition to contributing to drug discovery projects, CADD scientists advance the portfolio of CADD technologies that can be applied to projects. The use of sophisticated and computationally demanding CADD technology is encouraged at BI. While we agree, at least in principle, with the concept of parsimony when applying modeling approaches to projects, as postulated recently by Roche scientists [9], we are also convinced that it is worth investing in computer-intensive technologies when the results lead to new, meaningful and testable hypotheses. Molecular dynamics simulations are performed for multiple purposes at BI, including the analysis of water clusters, thermodynamic integration calculations [10] for calculating binding free energies, or simulations for assessing the stability of proposed binding modes of compounds or fragments [11]. We use GPU clusters to enable high speed MD simulations, and we have also started exploring cloud computing as an additional resource for MD simulations and other computer-intensive tasks.

In order to advance new computational chemistry methodologies into productive, value-adding applications in drug discovery projects, the CADD scientists proactively monitor the trends and new developments within the field. To keep in-house efforts associated with the implementation and maintenance of new algorithms at a minimum, we typically take advantage of new functionalities when they are added to our commercial computational chemistry software suites or robust open-source frameworks. However, very often the scheduling of new functionalities added to commercial or open-source software tools is not in-line with the in-house needs for enhancements as governed by ongoing drug discovery programs. Tapping into the full potential of new technology requires the use of in-house data sets for validation, preferably in prospective settings. Also a tight integration of new technology into internal workflows becomes necessary to harvest the full potential of technology developments. Collaborations with academic groups have proven to be an essential element in advancing the technology that is available for in-house applications. Many technologies that are now part of the productive CADD portfolio at BI were initially explored in collaboration with academic partners and summer students. These methods include SAR analyses [12], predictive modeling [13, 14], quantum-chemical calculations of H-bond strengths [15], optimal pi–pi stacking geometries [16], GPCR modeling [17], computational protocols to postulate druggable binding sites at protein–protein-interaction interfaces [18], and conformational analyses [19]. We are also engaged in research into methodologies that are still at a very early stage but that we anticipate will have a high impact on applications in the future. An example of this type of research is the decomposition of protein–ligand interaction energies employing quantum-chemical calculations [20]. Furthermore, academic collaborations have great value in stimulating discussions between academic groups and the CADD scientists at BI, as current procedures are challenged and ideas for further advancement are generated. Often, the integration of tools from academic collaborations into the in-house IT and high-performance computing environment requires additional effort. We have found that working with widely accessible toolkits or platforms such as Java, Python [21], Knime [5] or RDKit [22] helps to keep this effort to a minimum. In some cases we have implemented new algorithms from scratch [23, 24] and, more recently, we have employed crowd-sourcing as part of a community challenge to generate predictive models for mutagenicity [25, 26].

## A common platform for compound design

A common platform shared by CADD scientists, structural biologists and medicinal chemists strengthens both teamwork and collaboration in compound design. Moreover, it enhances the efficiency and transparency of decision making because hypotheses, such as proposed binding modes and ideas for compounds to be synthesized, can be shared in a common format. Molecular modeling tools such as MOE [27] are complex and powerful expert systems, with their own built-in development toolkits. These allow the development and implementation of new modules, and the design of highly customized user interfaces for incorporating in-house and external tools. In recent years, medicinal chemists have become more amenable to using such tools and to conducting, for example, structure-based compound design campaigns. However, exploitation of the full potential of modeling tools requires skills such as scripting and GUI customization, and an in-depth understanding of the individual functionalities and their underlying computational algorithms. These skills are more likely to reside with computational chemists and IT experts. At BI, a global MOE working group has been installed to coordinate the deployment of tools and new features in MOE, as well as the individual customization of the user interface at each of the three BI research sites. An efficient, world-wide deployment procedure has been implemented for new releases of MOE and the updating of BI-specific features. Computational chemists and IT experts compile one global MOE package that also allows the inclusion of site-specific settings. For example, there is a site-specific top-level menu for each site that governs the access to available tools. All site-specific customizations are setup independently to avoid dependencies. Based on a mechanism that determines from which site a MOE session is being started, the appropriate menus and libraries are loaded. MOE has been enhanced by a number of external tools that are invoked via a communication meta-layer (vide infra). These tools include, for example, various property (e.g., logP) and in silico ADME descriptors, including property profile meta-services. The services can be invoked from within the MOE system or from the MOE database viewers (MDB).

Another recent example is the introduction of a DFT-based torsional analysis tool to medicinal chemists. This tool permits an easy, color-coded assessment of optimal compound conformations (Fig. 1).

Following the interactive selection of the four atoms that define a torsion angle, the molecule along with the torsion angle specification, is submitted to a calculation service engine (CCFW, vide infra). Input molecules with fixed incremental torsion angles are constructed and subjected to QM calculations on an HPC cluster. Due to the computation times required, a synchronous service cannot be operated interactively. Therefore, the results are collated in MOE, MOE database and Excel spreadsheet formats, and are automatically sent to the user by email after the calculation has been completed. Another complex service estimates the mutagenicity potential of compounds based on ab initio calculations of nitrenium ion stability for aromatic amines [28, 29]. These examples illustrate how fairly complex and CPU intensive tasks can be transferred to the medicinal chemists at BI, provided that the tasks can be reasonably standardized or formalized as a routine workflow without the need for manual intervention by the user.

Another tool frequently used by the medicinal chemistry community is a docking utility that runs GOLD [30] and, optionally, a 2D–3D conversion step using CORINA [31] in the background. The preparation, optimization and provision of the GOLD configuration files and pre-aligned protein structures are the responsibility of the computational chemistry experts. The various configuration files that control the different docking scenarios are provided, such as constrained, unconstrained or covalent docking, depending on the individual project requirements. Medicinal chemists can then select the most appropriate docking protocol from a web-based interface that submits all input files (protein structure and ligand, configuration file) to the backend service. Docking poses are returned in SDF or MOE format ready to be used in further design cycles. The docking results can be automatically combined with property predictions that prompt the medicinal chemists to consider multi-parameter optimization criteria as part of their decision making.

An important part of the support given to medicinal chemists in computational compound design is the provision of 3D-structural data that are ready for use. We have established an automated workflow for compiling pre-aligned structural data in so-called project master files that can be customized for each project. The workflow utilizes the MOEProject framework and in-house scripts for calculating the crystallographic packing environment around ligand binding sites, splitting protein multimers into biologically relevant units, protonating the structures, and aligning defined protein chains onto a reference. In addition, structures can be annotated and grouped by chemical scaffolds, biological activity data, calculated properties and customized classifications such as, for example, the flip state of a certain side chain, point mutation, or cofactor type. This facilitates the survey of the available structural information allowing immediate utilization of the appropriate structures in compound design.

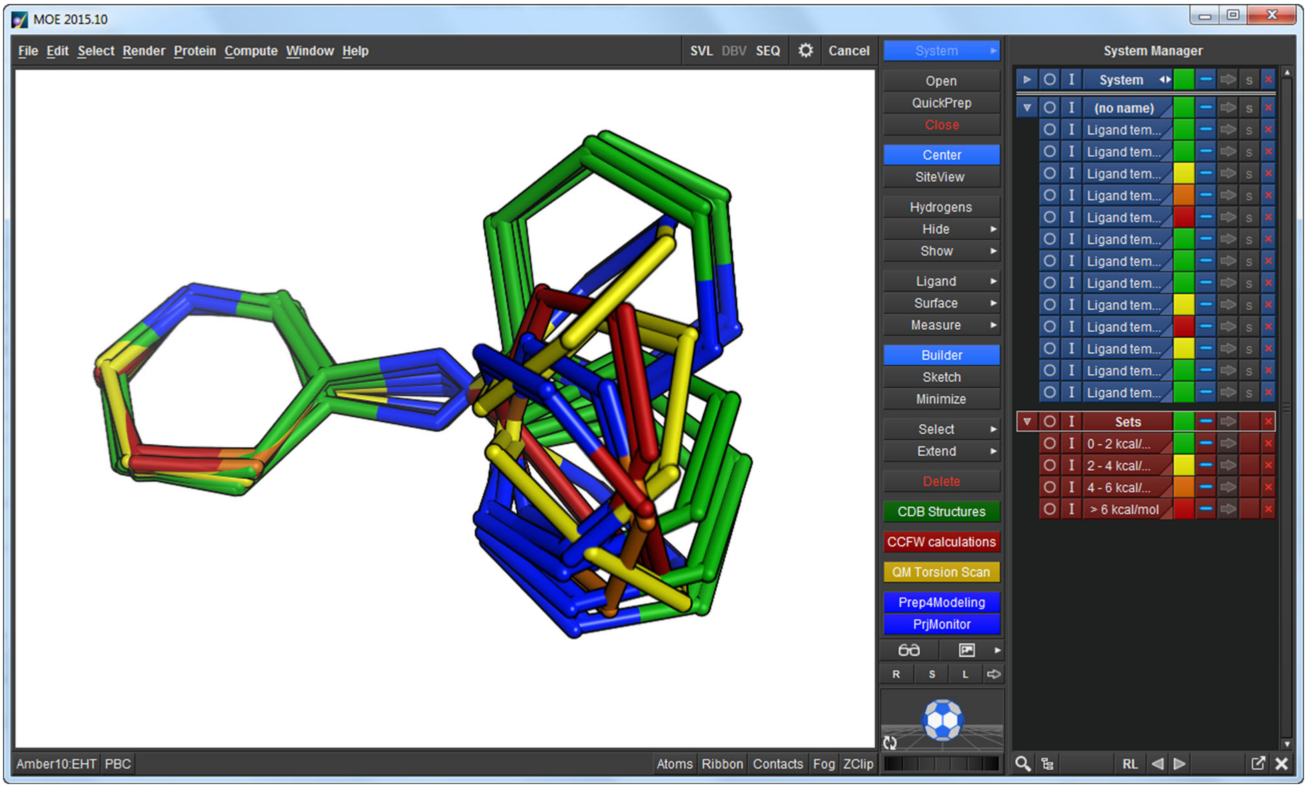

**Fig. 1** View of DFT-based torsion scan results in MOE. The conformations are color-coded by torsion strain energy. Structures with carbon atoms colored in *green* represent the low energy conformations

## Learning from data: predictive modeling and matched molecular pair (MMP) analyses

Prospectively predicting experimental parameters or the effect of molecular transformations on molecular properties significantly impacts the efficiency and shortens the design-learning cycle. Therefore, predictive modeling for ADMET endpoints has been a growing focus for computational chemists at BI [32, 33]. Unsurprisingly, commercially available models for most endpoints are not as relevant as those built from the vast repository of assay data accumulated over time at BI. We have established the following principles for building in silico models and for sharing them with the medicinal chemists:

1. Frequent, automated re-training and updating of predictive models guarantees the use of all relevant data, including the most recent data that can be particularly relevant for predictions on new analogs of actively explored compound series. After a few weeks, a deterioration in the predictive power of the predictive models can be measured [33]. Recently, we introduced a model rebuilding scheme that starts automatically as soon as new data becomes available. Automatically updating models ensure that the best prediction is available at all times. This means that predictions made today can be different from those made yesterday. The idea of giving up prediction consistency in favor of prediction accuracy has been gaining wider acceptance at BI.
2. Each prediction is returned together with a confidence estimate. Providing an easily interpretable way of managing expectations was found to be of great value when discussing prediction results with medicinal chemists. Predicted values with a confidence value below a certain threshold are not typically taken into account in the decision making process. In our experience confidence estimates derived from prediction agreements from an ensemble of different models perform particularly well (Fig. 2). However, applicability in descriptor space and confidence assessments by prediction distribution methods are used as well.
3. Early human dose predictions are used as an alternative to multi-parameter optimization scores. Currently, human dose predictions are triggered when in vitro stability and potency data are collected [34]. Initially, volume of distribution, plasma protein binding, and the efficacious dose are predicted by in silico methods [13] and are subsequently refined as experimental data becomes available.

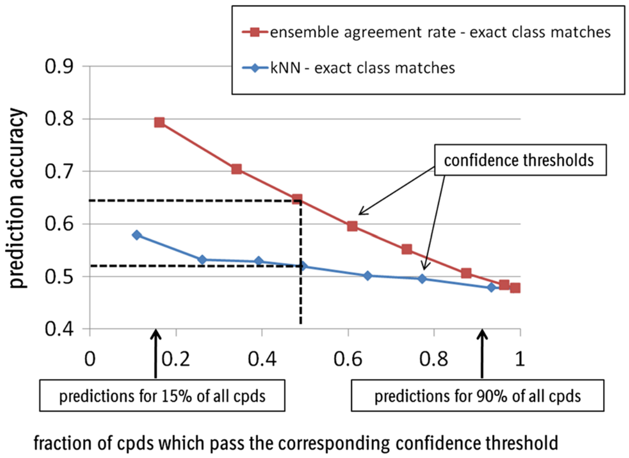

**Fig. 2** Influence of different confidence measures on the prediction accuracy of test compounds for a predictive human liver microsomal stability model. An ensemble agreement rate for 50 models used as confidence measure (shown in *red*) generates much higher prediction accuracies for a given fraction of compounds than a k-nearest neighbor (kNN) confidence measure does (shown in *blue*)

4. Seamless access to predicted properties and integration with other modeling output, such as docking results, is highly valued. Automated docking results are, for example, sometimes combined with predicted ADME properties without a specific request. The properties are then predicted on the fly when chemists draw or modify the structures in Marvin or MOE.

Although there is a trend towards building more predictive models serving a global (across sites) medicinal chemistry community at BI, many models are still built at the individual sites, primarily for local customers and often for project-specific purposes. However, we have standardized the deployment of and access to models by establishing a meta-layer (vide infra) that permits local model building and global consumption at multiple front-ends, including MOE, Marvin, Knime and Pipeline Pilot. This means that any model can easily be switched from a local to a global model and vice versa. This setup allows the combination of model predictions with the results from other tools, such as automated docking engines. For performance reasons we have moved away from workflow software such as Knime and Pipeline Pilot for in silico model building, in favor of a Python-based model building framework that is optimized for speed. It enables chemists to generate prediction results using Marvin [35] or MOE as the front-end within a few seconds of drawing or modifying a molecule. This allows predictive models to be integrated into the synthesis planning in an interactive way.

Table 1 summarizes in silico models used productively at BI. Similar to what has been reported for other pharmaceutical research efforts [36], we focus on predicting end points for which a sizable number of data points are available. We employ commonly-used machine learning methodologies such as random forests and support vector machines to train regression and classification models. Improving the predictive models for the parameters that are most relevant for human dose predictions has become a recent focus, with emphasis on in vitro clearance, volume of distribution, and plasma protein binding models. In addition, we are building target potency and efficacy models based on data automatically extracted from our compound database, and are combining them with models and data from the literature [37].

We are currently working on integrating the growing number of predictive models into the decision workflows of project teams, both for making synthesis decisions and for advancing compounds towards in depth profiling. One element is the use of in silico models, in parallel with the actual experimental assay, to advance compounds immediately onto the next level of a screening cascade, thereby reducing the time for learning cycles. Another element is the exploration of early human dose predictions as a holistic alternative to multi-parameter optimization measures, as a means of simplifying the decision making criteria for the chemists.

While these predictive models are very often black-box models that are difficult to interpret, we have recently enabled more illustrative ways of mining the wealth of in-house data to support SAR analysis and compound design by analyzing matched molecular pair MMP transformations [42–44]. Within the context of a specific project, target-related analyses are supported by displaying matched molecular series for each individual structural class. To support the improvement of optimization parameters, such as solubility or metabolic stability, we have provided statistical analyses of all molecular transformations in our corporate database and their effect on the respective parameter. These analyses have been made available to medicinal chemists who can then apply favorable in silico transformations to ongoing design campaigns. Recently, we have extended the MMP methodology to peptides [45].

## A central hub for integrating calculation engines into medicinal chemistry tools: Computational Chemistry Framework

The provision of powerful computational tools to an extended user community requires CADD and IT expertise to ensure that the workflows and applications implemented are robust and scientifically validated. Success depends on the seamless and user-friendly integration of these tools into the desktop applications that are used by medicinal chemists in their daily work for drawing molecules or performing SAR analyses, such as Marvin, Spotfire [46] or

**Table 1** *In silico* models in production at BI

| Category | Model | Number of compounds in training set | Comments* |
|---|---|---|---|
| Absorption | Caco-Efflux (PEAB, PEBA, efflux ratio) | 19,000 | Regression models; out-of-bag RMS errors for pPEAB, pPEBA, and p(PEAB/PEBA): 0.49, 0.33, 0.43; Pearson $r^2$: 0.72, 0.68, 0.67. Multiclass models: out-of-bag mean accuracy for PEAB and efflux 0.85 and 0.83, respectively |
| | MDCK efflux | 5000 | Multiclass model; out-of-bag mean accuracy 0.85 |
| Distribution | Volume of distribution (rat and human) | 8000 (rat); 670 (human from literature [38]) | Rat VDss model: Regression model for logVDss; out-of-bag RMS: 0.34; Pearson $r^2$: 0.59<br>Human VDss: regression model for logVDss; out-of-bag RMS: 0.42; Pearson $r^2$: 0.57 |
| | Human and rat plasma protein binding | 6000 | Multiclass models: out-of-bag mean accuracies 0.80 (human) and 0.82 (rat) |
| Metabolism | Microsomal stability (rat, mouse and human) | 28,000–93,000 | Multiclass models, out-of-bag mean accuracy 0.80 for human, rat and mouse |
| | CYP450 inhibition (3A4, 2D6, 2C8, 2C9, 2C19) | 12,000–32,000 | Multiclass models, out-of-bag mean accuracy 0.83 for all isoenzymes except 3A4 (0.87) |
| | In vivo CL (rat) | 6000 | Multiclass model, out-of-bag mean accuracy 0.76 |
| Toxicity | Ames mutagenicity | Ab initio method based on nitrenium ion stability [28] | Accuracy 85 %, sensitivity 91 %, specificity 72 % |
| | Phospholipidosis | Ploemen [39] and Cronin [40] models | Two class model: sensitivity 0.79, specificity 0.80, PPV 0.74, NPV 0.84 |
| | hERG inhibition | 9000 | Multiclass model, out-of-bag mean accuracy 0.82 |
| | Structural safety alerts | Collection of substructures | No statistics available |
| Physicochemical properties | Solubility at different pH values | 74,000 | Multiclass models, out-of-bag mean accuracies between 0.84 and 0.88 |
| | logP | 10,000 | kNN regression model: $r^2$: 0.69, RMSE: 0.79 |
| | pKa | MoKa2.6.4 [41] trained with in-house data | Independent test set: 87 % within 1 log unit |

* All multiclass models are 3-class models. Out-of-bag accuracies are obtained by calculating the prediction accuracies for all training set compounds not used in a bootstrap sample and then averaging over all classes and samples

D360 [47]. These applications are not part of modeling tools such as MOE, chemoinformatics packages, or the workflow tools (e.g. PipelinePilot or KNIME) that are normally used in the realm of CADD. Also, CPU-intensive workflows cannot be executed on standard PCs, requiring complex hardware infrastructure instead such as HPC clusters or clouds. To bridge the gap between the medicinal chemistry and computational chemistry software tool worlds we have developed a meta layer called the Computational Chemistry Framework (CCFW), which allows flexible connection of the front ends used by medicinal chemists with the computational chemistry calculation engines in the backend (Fig. 3). Rather than implementing a single, fully integrated system, the CCFW has been designed as a middle layer between these two worlds, allowing automated CADD tasks to be wrapped in web services, using defined parameters and standardized I/O file exchange formats. The CCFW and its backend services are completely client independent. Selected frontends can be independently enabled to trigger the CCFW services by using APIs or plugins. As a result, the CCFW services such as property calculators can be called from within MOE, Marvin or other clients without the need to develop and maintain multiple backend services for the same purpose. Since the CCFW calls and results are standardized, any amendments, bug fixes or upgrades of the backend services do not require further modification at the frontend. The concept of integrating the frontend and backend components in a modular way ensures high flexibility for developing and maintaining the CCFW backend services. In addition to the provision of automated services to the medicinal chemistry community, the CCFW also offers the

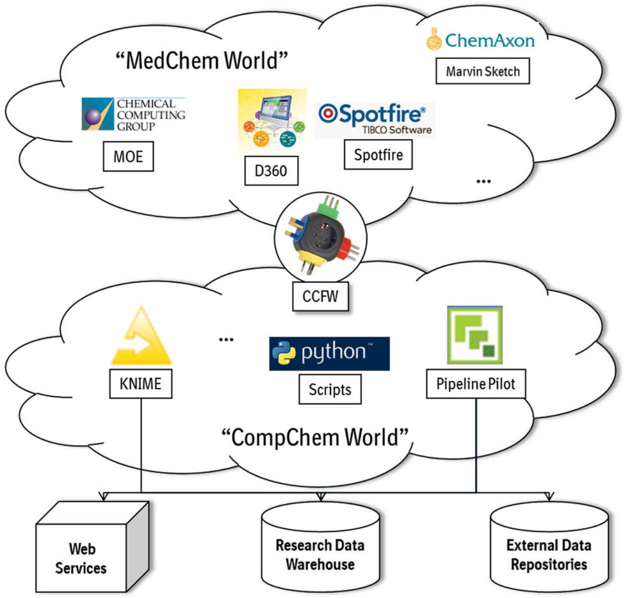

**Fig. 3** Computational Chemistry Framework (CCFW) as a bridge between the backend CADD calculation engines and user frontends. CCFW services can be invoked via client-independent web service calls from various clients. Calculations in the backend are typically performed via scripts (Python, shell) or pipelining tools

opportunity to integrate automated and standardized services in analyses that are conducted by CADD scientists, thereby increasing the efficiency.

The CADD scientists are responsible for developing the CCFW services in close cooperation with IT and the medicinal chemists. Typically used backend engines are command-line based services operated by scripting tools (BASH, PYTHON), or workflow protocols (PipelinePilot, KNIME) that can be directly invoked from within the CCFW.

## BI global development, maintenance, and application of CADD technology

Virtual screening (VS) is a key technique for hit finding at BI. Fast and robust ligand- and structure-based VS workflows enable rapid execution of tailored initial hit finding or iterative follow-up VS campaigns. Ligand-based VS and analoging workflows employing multiple complementary similarity search methods and data fusion [48] have been successful in multiple hit finding projects. VS workflows and compound decks can be flexibly adapted to available assays (biophysical or biochemical, throughput) in a project, which is essential for facilitating swift screening deck compilation for the diverse early drug discovery target portfolio at BI.

However, VS is typically restricted to the chemical space defined by the in-house and commercially available compound databases screened [49–52]. Therefore, a focus over the past 10 years has been to expand into the vast virtual chemical space accessible via combinatorial chemistry. To achieve this expansion we have created a global platform called the BI Comprehensive Library of Accessible and Innovative Molecules (BICLAIM) [53]. It consists of putative library cores and reagents that we extract computationally from our corporate compound databases—chemical space that we know or assume is accessible by mining electronic lab notebooks, as well as commercial sources. The current version of BICLAIM contains almost 90,000 cores and tens of thousands of reagents spanning a combinatorial chemistry space of more than $10^{17}$ compounds. In addition to maintaining and growing BICLAIM, we have been developing search methods that allow us to mine this compound space using various computational techniques to prioritize combinatorial libraries for synthesis. These libraries have the advantage of a reduced synthesis risk that counter-balances the higher risk associated with making these *de novo* compounds without proof of activity against the intended target. We have demonstrated that combining the power of large numbers of *de novo* compounds from libraries with the uncertainties of virtual screening [50] is an effective way of finding attractive hits.

Several 2D and 3D workflows have been developed at BI to search the BICLAIM space. If at least one template ligand with validated activity is known, 2D FeatureTree [54] searches followed by ROCS [55] 3D matches are carried out. Several examples have been reported where novel chemical matter has been identified using this approach, including GPR119 agonists [56] and CDK2 inhibitors [57].

In addition, direct 3D searches of the BICLAIM space have been enabled using the PharmShapeCC software that allows 3D pharmacophore and shape complementarity searches in the entire BICLAIM space [23]. Examples of finding novel MMP13 inhibitors, CCR1 antagonists, CXCR5 antagonists [23] and RORC inverse agonists [58] have been published. More recently, we have enabled 3D searching in partially enumerated BICLAIM subspaces using ROCS. BICLAIM is maintained as a global platform for virtual screening at BI with regular updates and access to substructure searches from within PipelinePilot and D360. After virtual screening results have been generated using 2D and 3D tools, a final selection of library cores and building blocks is made in collaboration between CADD scientists, combinatorial chemistry experts, and medicinal chemists.

BICLAIM is maintained and developed as a truly global resource at BI with input from all research sites, not only from CADD but also from the medicinal and combinatorial chemistry groups. Other examples of in-house developments that have been made accessible across sites are the *de novo* design program BiBuilder [24] which has, for example, been applied to identifying a novel CB2 agonist [59], and a Python-based model building framework which was developed at one site and has been adopted as a global platform for predictive modeling. Workflow scripts have also been exchanged on various occasions (Pipeline Pilot and Knime).

## Concluding remarks and outlook

There is currently an increased focus in pharmaceutical research towards enabling project teams to make earlier decisions in drug discovery projects governing where CADD typically invests time and resources at BI. In recent years, advances in computer power, flexibility in creating intricate workflows and the seamless access to CADD software tools connected through a meta-layer (such as the CCFW at BI) to multiple front-ends have allowed chemists easy access to fairly complex calculation engines and computational services. Even more importantly, the response times to queries made against such calculation service layers (e.g., property and model predictions) has significantly decreased to the point where the interactive use of modeling results has become feasible for both medicinal and computational chemists alike. As a consequence, the usage of novel predictive models in the context of compound profiling has increased significantly, and often guides subsequent decisions as to whether or not chemical matter should be progressed in a drug discovery project. The ease of access to computational tools has been changing the way in which CADD scientists and medicinal chemists interact in project teams. An increasing number of automatable CADD-related tasks are becoming amenable to medicinal chemists, without compromising the quality of modeling outcomes. An added benefit of medicinal chemists conducting automatable CADD tasks is that it frees up time for the CADD scientist to invest in the development of more sophisticated technology, which can then be applied to project advancement and additional design ideas in new ways, or to address aspects that traditionally may not have been within scope of CADD. A unique technology developed and broadly practiced at BI is the use of large scale combinatorial chemistry combined with virtual screening to identify hits and leads early in drug discovery projects.

As pointed out recently by scientists from Bayer [36], computational design plays a much smaller role in the pharmaceutical industry than in other industries. However, given the challenges for the pharma industry, which result in an ever increasing need for speed, efficiency and a higher share of unprecedented targets in the portfolio, we believe that CADD has the potential to influence the way in which drug design and discovery is pursued. The impact of CADD will continue to depend on translating CADD results into insights, along with tangible and reliable recommendations to medicinal chemists as to what compound should be made next. A continued investment into developing more accurate and robust predictive methods is necessary to increase the CADD impact. This will need to be supported by an increased availability of experimental data. Public data (e.g., ChEMBL [60]) are already being integrated seamlessly into in-house data sources [61]. In addition, the increased realization by many pharmaceutical companies that they need to share pre-competitive data [62] means that new opportunities for building predictive models with higher accuracy and wider applicability will be opened. We also expect that the computing resources used for CADD work will become a commodity. The increased availability of cloud computing will encourage the development of more accurate, albeit computationally-intensive, methods allowing their application on a much larger scale than is currently being carried out. Collaborations with academic groups will continue to play a key role in strengthening the portfolio of tools and the exploration of new methodologies, supplemented by crowd sourcing initiatives which provide easy access to a wealth of scientific talent for solving specific problems on demand. Modern drug discovery aims to tackle unprecedented targets that pose significant challenges to CADD. These targets, such as protein–protein interactions or RNA binding, usually have low druggability, and often require the invention of chemical matter and design modalities beyond the established small-molecule domain. There is clearly a growing need to expand the applicability domain of the CADD technology portfolio into these areas. With a computational infrastructure such as the CCFW, strong external scientific networks and the consistent execution/implementation of the three CADD scientist roles (as described in this paper) these challenges can be overcome.

**Acknowledgments** We would especially like to thank our Computational Chemistry and CADD teams in Biberach, Ridgefield and Vienna for their dedicated contribution, without whom this work would not have been possible. The support from our colleagues in IT, in particular for the implementation of the CCFW, is gratefully acknowledged. We thank Ulrike Küfner-Mühl, Darryl McConnell, Dirk Kessler, and Robert Hughes for fruitful discussions and for their continued support. We also thank Karen J. Bergner for critically reading the manuscript.

## References

1. Schiele F, van Ryn J, Litzenburger T, Ritter M, Seeliger D, Nar H (2015) Structure-guided residence time optimization of a dabigatran reversal agent. MAbs 7:871–880

2. Seeliger D, Schulz P, Litzenburger T, Spitz J, Hoerer S, Blech M, Enenkel B, Studts JM, Garidel P, Karow AR (2015) Boosting antibody developability through rational sequence optimization. MAbs 7:505–515
3. Beck B, Seeliger D, Kriegl JM (2015) The impact of data integrity on decision making in early lead discovery. J Comput Aided Mol Des 29:911–921
4. Biovia pipeline pilot: http://accelrys.com/products/collaborative-science/biovia-pipeline-pilot/. Accessed 23 Feb 2016
5. Knime: https://www.knime.org/. Accessed 23 Feb 2016
6. Beck B (2012) BioProfile—Extract knowledge from corporate databases to assess cross-reactivities of compounds. Bioorganic Med Chem 20:5428–5435
7. Loughney D, Claus BL, Johnson SR (2011) To measure is to know: an approach to CADD performance metrics. Drug Discov Today 16:548–554
8. Baldwin ET (2012) Metrics and the effective computational scientist: process, quality and communication. Drug Discov Today 17:935–941
9. Kuhn B, Guba W, Hert J, Banner D, Bissantz C, Ceccarelli S, Haap W, Korner M, Kuglstatter A, Lerner C, Mattei P, Neidhart W, Pinard E, Rudolph MG, Schulz-Gasch T, Woltering T, Stahl M (2016) A real-world perspective on molecular design. J Med Chem 59:4087–4102
10. Christ CD, Fox T (2014) Accuracy assessment and automation of free energy calculations for drug design. J Chem Inf Model 54:108–120
11. Hucke O, Coulombe R, Bonneau P, Bertrand-Laperle M, Brochu C, Gillard J, Joly MA, Landry S, Lepage O, Llinas-Brunet M, Pesant M, Poirier M, Poirier M, McKercher G, Marquis M, Kukolj G, Beaulieu PL, Stammers TA (2014) Molecular dynamics simulations and structure-based rational design lead to allosteric HCV NS5B polymerase thumb pocket 2 inhibitor with picomolar cellular replicon potency. J Med Chem 57:1932–1943
12. Wassermann AM, Haebel P, Weskamp N, Bajorath J (2012) SAR matrices: automated extraction of information-rich SAR tables from large compound data sets. J Chem Inf Model 52:1769–1776
13. Demir-Kavuk O, Bentzien J, Muegge I, Knapp EW (2011) DemQSAR: predicting human volume of distribution and clearance of drugs. J Comput Aided Mol Des 25:1121–1133
14. Kramer C, Beck B, Kriegl JM, Clark T (2008) A composite model for HERG blockade. Chem Med Chem 3:254–265
15. Nocker M, Handschuh S, Tautermann C, Liedl KR (2009) Theoretical prediction of hydrogen bond strength for use in molecular modeling. J Chem Inf Model 49:2067–2076
16. Huber RG, Margreiter MA, Fuchs JE, Von GS, Tautermann CS, Liedl KR, Fox T (2014) Heteroaromatic pi-stacking energy landscapes. J Chem Inf Model 54:1371–1379
17. Kneissl B, Leonhardt B, Hildebrandt A, Tautermann CS (2009) Revisiting automated G-protein coupled receptor modeling: the benefit of additional template structures for a neurokinin-1 receptor model. J Med Chem 52:3166–3173
18. Li H, Kasam V, Tautermann CS, Seeliger D, Vaidehi N (2014) Computational method to identify druggable binding sites that target protein-protein interactions. J Chem Inf Model 54:1391–1400
19. Hao MH, Haq O, Muegge I (2007) Torsion angle preference and energetics of small-molecule ligands bound to proteins. J Chem Inf Model 47:2242–2252
20. Phipps MJ, Fox T, Tautermann CS, Skylaris CK (2016) Energy decomposition analysis based on absolutely localized molecular orbitals for large-scale density functional theory calculations in drug design. J Chem Theory Comput 12:3135–3148
21. Python version 2.7 available at http://www.python.org. Accessed 23 Feb 2016
22. RDKit: http://www.rdkit.org. Accessed 23 Feb 2016
23. Muegge I, Zhang Q (2015) 3D virtual screening of large combinatorial spaces. Methods 71:14–20
24. Teodoro M, Muegge I (2011) BIBuilder: exhaustive Searching for De Novo Ligands. Mol Inform 30:63–75
25. Bentzien J, Muegge I, Hamner B, Thompson DC (2013) Crowd computing: using competitive dynamics to develop and refine highly predictive models. Drug Discov Today 18:472–478
26. Bentzien J, Bharadwaj R, Thompson DC (2015) Crowdsourcing in pharma: a strategic framework. Drug Discov Today 20:874–883
27. Molecular Operating Environment (MOE) (2015) 2014.09; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7
28. Bentzien J, Hickey ER, Kemper RA, Brewer ML, Dyekjaer JD, East SP, Whittaker M (2010) An in silico method for predicting Ames activities of primary aromatic amines by calculating the stabilities of nitrenium ions. J Chem Inf Model 50:274–297
29. Bentzien J, Muegge I (2014) In silico predictions of genotoxicity for aromatic amines. Front Biosci (Landmark Ed) 19:649–661
30. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267:727–748
31. Sadowski J, Gasteiger J, Klebe G (1994) Comparison of automatic three-dimensional model builders using 639 X-ray structures. J Chem Inf Comput Sci 34:1000–1008
32. Kriegl JM, Arnhold T, Beck B, Fox T (2005) A support vector machine approach to classify human cytochrome P450 3A4 inhibitors. J Comput Aided Mol Des 19:189–201
33. Muegge I, Bentzien J, Mukherjee P, Hughes RO (2016) Automatically updating predictive modeling workflows support decision making in drug design. Future Med Chem 8:1779–1796
34. Page KM (2016) Validation of early human dose prediction: a key metric for compound progression in Drug Discovery. Mol Pharm 13:609–620
35. Marvin 6.0.2, 2014, ChemAxon (http://www.chemaxon.com)
36. Hillisch A, Heinrich N, Wild H (2015) Computational chemistry in the pharmaceutical industry: from childhood to adolescence. ChemMedChem 10:1958–1962
37. Bieler M, Reutlinger M, Rodrigues T, Schneider P, Kriegl JM, Schneider G (2016) Designing multi-target compound libraries with Gaussian process models. Mol Inform 35:192–198
38. Obach RS, Lombardo F, Waters NJ (2008) Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 670 drug compounds. Drug Metab Dispos 36:1385–1405
39. Ploemen JP, Kelder J, Hafmans T, van de Sandt H, van Burgsteden JA, Saleminki PJ, Van EE (2004) Use of physicochemical calculation of pKa and CLogP to predict phospholipidosis-inducing potential: a case study with structurally related piperazines. Exp Toxicol Pathol 55:347–355
40. Przybylak KR, Alzahrani AR, Cronin MT (2014) How does the quality of phospholipidosis data influence the predictivity of structural alerts? J Chem Inf Model 54:2224–2232
41. Molecular Discovery Ltd. Moka version 1.1. http://www.moldiscovery.com/software/moka/ Accessed 23 Feb 2016)
42. Hussain J, Rea C (2010) Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. J Chem Inf Model 50:339–348
43. Griffen E, Leach AG, Robb GR, Warner DJ (2011) Matched molecular pairs as a medicinal chemistry tool. J Med Chem 54:7739–7750
44. Geppert T, Beck B (2014) Fuzzy matched pairs: a means to determine the pharmacophore impact on molecular interaction. J Chem Inf Model 54:1093–1102
45. Fuchs JE, Wellenzohn B, Weskamp N, Liedl KR (2015) Matched peptides: tuning matched molecular pair analysis for biopharmaceutical applications. J Chem Inf Model 55:2315–2323

46. Tibco Spotfire. version 6.3. http://spotfire.tibco.com/ Accessed 4 July 2016
47. Certara. D360: The Pharmaceutical Industry's Data Analytics and Scientific Informatics Platform. http://www.certara.com/software/scientific-informatics/d360 Accessed 24 Feb 2016
48. Bergner A, Parel SP (2013) Hit expansion approaches using multiple similarity methods and virtualized query structures. J Chem Inf Model 53:1057–1066
49. Muegge I, Oloff S (2006) Advances in virtual screening. Drug Discov Today Technol 3:405–411
50. Muegge I (2008) Synergies of virtual screening approaches. Mini Rev Med Chem 8:927–933
51. Muegge I, Oloff S (2010) Virtual screening. In: Abraham DJ, Rotella DP (eds) Burger's medicinal chemistry drug discovery and development, vol 2, 7th edn. Wiley, Hoboken, pp 1–46
52. Muegge I, Mukherjee P (2016) An overview of molecular fingerprint similarity search in virtual screening. Expert Opin Drug Discov 11:137–148
53. Lessel U, Wellenzohn B, Lilienthal M, Claussen H (2009) Searching fragment spaces with feature trees. J Chem Inf Model 49:270–279
54. Rarey M, Dixon JS (1998) Feature trees: a new molecular similarity measure based on tree matching. J Comput Aided Mol Des 12:471–490
55. Grant JA, Nicholls A, Stahl MT. ROCS OpenEye, 3600 Cerrillos Rd., Suite 1107, Santa Fe, NM 87507
56. Wellenzohn B, Lessel U, Beller A, Isambert T, Hoenke C, Nosse B (2012) Identification of new potent GPR119 agonists by combining virtual screening and combinatorial chemistry. J Med Chem 55:11031–11041
57. Lessel U, Wellenzohn B, Fischer JR, Rarey M (2012) Design of combinatorial libraries for the exploration of virtual hits from fragment space searches with LoFT. J Chem Inf Model 52:373–379
58. Muegge I, Collin D, Cook B, Hill-Drzewi M, Horan J, Kugler S, Labadia M, Li X, Smith L, Zhang Y (2015) Discovery of 1,3-dihydro-2,1,3-benzothiadiazole 2,2-dioxide analogs as new RORC modulators. Bioorg Med Chem Lett 25:1892–1895
59. Hickey ER, Zindell R, Cirillo PF, Wu L, Ermann M, Berry AK, Thomson DS, Albrecht C, Gemkow MJ, Riether D (2015) Selective CB2 receptor agonists. Part 1: the identification of novel ligands through computer-aided drug design (CADD) approaches. Bioorg Med Chem Lett 25:575–580
60. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, Kruger FA, Light Y, Mak L, McGlinchey S, Nowotka M, Papadatos G, Santos R, Overington JP (2014) The ChEMBL bioactivity database: an update. Nucleic Acids Res 42:D1083–D1090
61. Williams AJ, Harland L, Groth P, Pettifer S, Chichester C, Willighagen EL, Evelo CT, Blomberg N, Ecker G, Goble C, Mons B (2012) Open PHACTS: semantic interoperability for drug discovery. Drug Discov Today 17:1188–1198
62. Briggs KA (2016) Is preclinical data sharing the new norm? Drug Discov Today (in press)