to the formulation and empirical testing of *best modeling practices*. This latter endeavor is especially critical in light of the growing interest of regulatory agencies to developing the most reliable and predictive models for environmental risk assessment[83] and placing such models in the public domain.

PREDICTIVE QSAR MODELING WORKFLOW AND ITS APPLICATION TO VIRTUAL SCREENING

Our experience in QSAR model development and validation has led us to establishing a complex strategy that is summarized in Figure 10.2. It describes the predictive QSAR modeling workflow focused on delivering validated models and, ultimately, computational hits confirmed for the experimental validation. We start by randomly selecting a fraction of compounds (typically, 10–15%) as an external validation set. The remaining compounds are then divided rationally (using the sphere exclusion protocol implemented in our laboratory[46]) into multiple training and test sets that are used for model development and validation, respectively, using criteria discussed in more detail below. We employ multiple QSAR techniques based on the combinatorial exploration of all possible pairs of descriptor sets coupled with various statistical data-mining techniques (combi-QSAR) and select models characterized by high accuracy in predicting both training and test sets data. Validated models are finally tested using the evaluation set. The critical step of the external validation is the use of applicability domains. If external validation demonstrates the significant predictive power of the models we use all such models for virtual screening of available chemical databases (e.g., ZINC[84]) to identify putative active compounds and work with collaborators who could validate such hits experimentally. The entire approach is described in detail in several recent articles and reviews.[18,19,40]

In our recent studies we were fortunate to recruit experimental collaborators who have validated computational hits identified through our modeling of anticonvulsants,[59] HIV-1 reverse transcriptase inhibitors,[85] D1 antagonists,[37] antitumor compounds,[86] and β-lactamase inhibitors.[88] Thus, models resulting from this workflow could be used to prioritize the selection of chemicals for the experimental validation. However, because we cannot generally guarantee that every prediction resulting from our modeling effort will be validated experimentally we cannot include the experimental validation step as a mandatory part of the workflow in Figure 10.2, which is why we used the dotted line for this component. We note that our approach shifts the emphasis on ensuring good (best) statistics for the model that fits known experimental data toward generating testable hypothesis about purported bioactive compounds. Thus, the output of the modeling has exactly the same format as the input [i.e., chemical structures and (predicted) activities making model interpretation and utilization completely seamless for medicinal chemists].

The development of truly validated and predictive QSAR models affords their growing application in chemical data mining and combinatorial library design.[88,89] For example, 3D stereoelectronic pharmacophore based on QSAR modeling was used recently to search the National Cancer Institute Repository of Small Molecules[11] to find new leads for inhibiting HIV type 1 reverse transcriptase at the nonnucleoside binding site.[90]

Our studies have shown that QSAR models could be used successfully as virtual screening tools to discover compounds with the desired biological activity in chemical databases or virtual libraries.[18,37,59,86,87,91] The discovery of novel bioactive chemical entities is the primary goal of computational drug discovery, and the development of validated and predictive QSAR models is critical to achieve this goal. We present several examples of these studies below to illustrate the use of QSAR models as virtual screening tools for lead identification.

EXAMPLES OF APPLICATIONS OF THE PREDICTIVE QSAR MODELING WORKFLOW FOR LEAD OPTIMIZATION

To illustrate the power of validated QSAR models as virtual screening tools we shall discuss the examples of studies that resulted in experimentally confirmed hits. We note that such studies could only be done if there is sufficient data available for a series of tested compounds such that robust validated models could be developing using the workflow described in Figure 10.2.

The first example is *anticonvulsant compounds*. In the first phase of modeling, we have applied *k*NN[17] and simulated annealing-partial least squares (SA-PLS)[89]

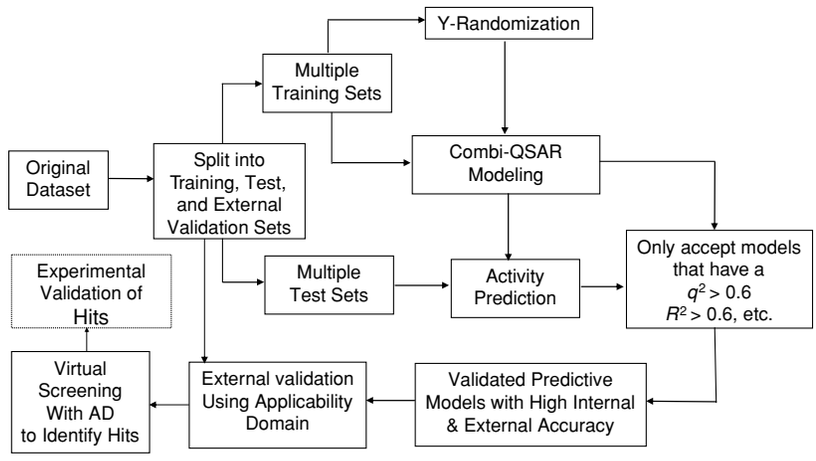

Figure 10.2. Flowchart of predictive QSAR modeling framework based on the validated combi-QSAR models.