



Universiteit  
Leiden  
The Netherlands

## Application of portfolio optimization to drug discovery

Yevseyeva, I.; Lenselink, E.B.; Vries, A. de; IJzerman, A.P.; Deutz, A.H.; Emmerich, M.T.M.

### Citation

Yevseyeva, I., Lenselink, E. B., Vries, A. de, IJzerman, A. P., Deutz, A. H., & Emmerich, M. T. M. (2018). Application of portfolio optimization to drug discovery. *Information Sciences*, 475, 29-43. doi:10.1016/j.ins.2018.09.049

Version: Not Applicable (or Unknown)

License: [Leiden University Non-exclusive license](#)

Downloaded from: <https://hdl.handle.net/1887/81942>

**Note:** To cite this publication please use the final published version (if applicable).

# 1 Application of portfolio optimization to drug discovery

2 Iryna Yevseyeva<sup>1,3</sup>, Eelke B. Lenselink<sup>2</sup>, Alice de Vries<sup>3</sup>,  
3 Adriaan P. IJzerman<sup>2</sup>, André H. Deutz<sup>3</sup> and Michael T.M. Emmerich<sup>3</sup>

4 <sup>1</sup>*School of Computer Science and Informatics, Faculty of Technology, De Montfort*  
5 *University, LE1 9BH, Leicester, UK*

6 <sup>2</sup>*Medicinal Chemistry/Leiden Academic Centre for Drug Research, Leiden University,*  
7 *2333-CA Leiden, The Netherlands*

8 <sup>3</sup>*Leiden Institute of Advanced Computer Science, 2333-CA Leiden, The Netherlands*

---

## 9 **Abstract**

In this work, a problem of selecting a subset of molecules, which are potential lead candidates for drug discovery, is considered. Such molecule subset selection problem is formulated as a portfolio optimization, well known and studied in financial management. The financial return, more precisely the return rate, is interpreted as return rate from a potential lead and calculated as a product of gain and probability of success (probability that a selected molecule becomes a lead), which is related to performance of the molecule, in particular, its (bio-)activity. The risk is associated with not finding active molecules and is related to the level of diversity of the molecules selected in portfolio. It is due to potential of some molecules to contribute to the diversity of the set of molecules selected in portfolio and hence decreasing risk of portfolio as a whole. Even though such molecules considered in isolation look inefficient, they are located in sparsely sampled regions of chemical space and are different from more promising molecules. One way of computing diversity

of a set is associated with a covariance matrix, and here it is represented by the Solow-Polasky measure. Several formulations of molecule portfolio optimization are considered taking into account the limited budget provided for buying molecules and the fixed size of the portfolio. The proposed approach is tested in experimental settings for three molecules datasets using exact and/or evolutionary approaches. The results obtained for these datasets look promising and encouraging for application of the proposed portfolio-based approach for molecule subset selection in real settings.

10 *Keywords:* Portfolio Approach, Multicriteria optimization, Decision  
11 Support, Drug Discovery

---

## 12 **1. Introduction**

13 When searching for the most promising drug like molecules for a drug  
14 discovery project, usually, de novo drug discovery uses *in vitro* experiments  
15 (colloquially called “test-tube experiments”). For this, circa 100 promising  
16 molecules are selected from a database and typically only circa 1 percent of  
17 the molecules are tested successfully *in vitro*, that is they become so called  
18 *lead* molecules [4]. High-throughput screening (HTS) allows for testing a  
19 large number of molecules by robotized machines using advanced laboratory  
20 equipment. However, testing *in vitro* is an expensive process and cannot al-  
21 ways be applied to all possible projects, even though in industry millions of  
22 molecules can be screened if the target is interesting enough. To reduce costs,  
23 HTS can be complemented by preliminary *in silico* (performed via computer

24 simulation) virtual screening (VS). VS approaches [27] are used to pre-select  
25 molecules from virtual libraries or large databases of commercially available  
26 molecules (e.g. ZINC [13]) based on their chemical properties. Selection is  
27 typically done based on the assessment of the success probability of candidate  
28 molecules using either a compound-based method or a target-based method  
29 or a combination of both. Typically, no explicit economical information is  
30 taken into account and simple methods like clustering are applied. The suc-  
31 cess probability is not given directly, but in the form of a score corresponding  
32 to (bio-)activity that is proportional to it.

33 A typical scenario in a pharmaceutical research laboratory is that a  
34 chemist selects a subset from a large vendor database of molecules (e.g.  
35 ZINC) and orders these. Each molecule has a price, and the budget of the  
36 chemist is limited, but it has to be allocated for buying molecules. That  
37 is, money not spent cannot be used for another purpose (and, thus, will be  
38 lost). Note that here we do not look into experimental planning and drug  
39 production, see e.g. [1], which is a separate subject of research.

40 Classical approaches for selecting promising molecules are based on pre-  
41 dicted activity score. However, selecting molecules based on their perfor-  
42 mance (success probabilities / activity scores) only is not enough and even  
43 risky. It is due to a high probability of selecting well-performing, but similar  
44 molecules, which all might be unsuccessful for the same reason.

45 An alternative approach is to take into account diversity of selected sub-  
46 sets of molecules. However, selection purely based on diversity will neglect

47 the information about activity given to the chemists by VS models. More-  
48 over, price might also play a role in the choice as it influences the number  
49 of molecules that can be bought given a limited budget. Hence, existing,  
50 just clustering- or just scoring-based selection models are not sufficient for  
51 handling these problems.

52 In this work, we consider the first stage of drug discovery of identifying  
53 lead candidates with an approach which takes into account performance of  
54 molecules according to their predicted (bio-) activity and diversity of selected  
55 molecules simultaneously. Similar approach was taken in [17], where activity  
56 score and diversity were maximized at the same time. Here, the molecule  
57 subset selection problem is considered and modeled by analogy with a well-  
58 known financial portfolio selection problem, see e.g. [5]. A similar binary  
59 problem of finding an optimal combination of items subject to constraints  
60 is known in operations research as the knapsack problem, see e.g. [16]. Ac-  
61 cording to the portfolio optimization approach when selecting a subset of  
62 molecules to be tested *in vitro*, in addition to choosing molecules with high-  
63 est performance values (and maximizing the average quality of the selected  
64 subset of molecules), the molecules with the most dissimilar structures should  
65 be considered. The former aspect contributes to maximizing the quality of  
66 the selected subset of molecules and the latter one corresponds to maximizing  
67 the diversity of such a subset.

68 The financial portfolio return, more precisely the return rate, is inter-  
69 preted as the return rate from a potential lead and calculated as a prod-

uct of the gain and the probability of success (probability that a selected  
molecule becomes a drug in the end), which is related to the performance  
of the molecule, in particular, its (bio-)activity. The risk is associated with  
not finding active molecules when choosing a portfolio and is related to the  
level of diversity of the molecules in the portfolio. The diversity can be ex-  
pressed as a covariance matrix used by Solow and Polasky [23] for measuring  
diversity of a biological population. Interestingly, as an example of a utili-  
tarian approach to the biological diversity preservation, Solow and Polasky  
indicated the potential utility in future from one of the preserved species as  
a cure of some yet unknown disease (see [23]).

Some molecules, when considered in isolation look inefficient, but as part  
of a portfolio may contribute to the decreasing risk of a portfolio as a whole  
and may be included in a portfolio as they are located in sparsely sampled  
regions of chemical space and are different from more promising molecules.  
In addition, the limited budget provided for buying molecules and the fixed  
size of the portfolio are taken into account in the introduced drug portfolio  
model as constraints.

This article is structured as follows: In the next section 2, we consider  
the general (multiobjective) formulation of the (financial) portfolio selection  
problem and, then, in section 3, we model the lead subset selection problem as  
portfolio optimization. In section 4, we propose algorithms to solve portfolio  
selection formulations, and in section 5, we discuss results obtained for three  
molecule datasets. Finally, in section 6, we draw conclusions and indicate

93 directions for future research.

## 94 **2. Related Work**

### 95 *2.1. Portfolio selection as a multi-objective optimization problem*

96 The most-widely used formulation of portfolio selection problem was de-  
97 veloped by Markowitz early in the 50s [15]. It addresses a way of selecting  
98 a combination of several assets called *portfolio* that collectively would be of  
99 the best quality and be as diverse as possible. Hence, portfolio optimiza-  
100 tion should simultaneously satisfy two conflicting goals, minimizing risk and  
101 maximizing expected return of the portfolio, that is formally:

$$\begin{aligned} \min \sigma^2(\mathbf{x}) &= \sum_{i=1}^{N_{Total}} \sum_{j=1}^{N_{Total}} q_{ij} x_i x_j = \mathbf{x}^\top \mathbf{Q} \mathbf{x}; & (1) \\ \max E(\mathbf{x}) &= \sum_{i=1}^{N_{Total}} r_i x_i = \mathbf{r}^\top \mathbf{x}; \\ \text{s.t.} \quad & \sum_{i=1}^{N_{Total}} x_i = 1; \\ & x_i \in [0, 1], i = 1, \dots, N_{Total}, \end{aligned}$$

102 where  $N_{Total}$  is the number of assets;  $x_i$  is the proportion of money invested  
103 in the asset  $i$ ;  $r_i$  is the expected return (per period) of the asset  $i$ ; and  $q_{ij}$  is  
104 the real-valued covariance of expected returns of the assets  $i$  and  $j$ .

105 As a result of optimizing this problem not a single portfolio but a set  
106 of portfolios are selected that are optimal with respect to the two specified  
107 objectives.

108 For this problem the search space of portfolios  $\mathcal{S}$  is  $[0, 1]^{N_{Total}}$ . The set of  
 109 feasible portfolios  $\mathcal{F}$  is the subset of portfolios in  $\mathcal{S}$  with  $\sum_{i=1}^{N_{Total}} x_i = 1$ . We  
 110 consider two real valued objective functions defined on  $\mathcal{S}$ ,  $\sigma^2(\mathbf{x}) = \mathbf{x}^\top \mathbf{Q} \mathbf{x}$  and  
 111  $E(\mathbf{x}) = \mathbf{r}^\top \mathbf{x}$ . Each portfolio  $\mathbf{x}$  is associated with a 2-dimensional evaluation  
 112 vector in the objective space,  $(\sigma^2(\mathbf{x}), E(\mathbf{x}))^\top$ , where the risk objective is to  
 113 be minimized and the return objective is to be maximized.

114 Optimizing two or more conflicting objectives simultaneously is referred  
 115 to as Multiobjective Optimization (MOO). The portfolio selection problem  
 116 formulated as in (1) is bi-objective: Minimizing the risk and maximizing  
 117 the expected return should be taken into consideration and optimized at the  
 118 same time. These objectives are generally in conflict with each other and  
 119 finding a portfolio with minimal risk and maximal return simultaneously is  
 120 infeasible. Hence, decreasing risk for a portfolio can be obtained at the cost  
 121 of lowering its return only.

122 Interestingly, including some assets, which look inefficient when consid-  
 123 ered in isolation, may benefit the portfolio as a whole, since they contribute  
 124 to decreasing the risk of a portfolio when considered in combination with  
 125 other assets. This is due to their location in sparsely sampled regions of  
 126 search space and their difference from more promising assets. Cost of assets  
 127 may also be taken into account as a separate objective, but we included it  
 128 in the return (which is reduced by the costs invested in initial assets) and in  
 129 the budget constraint.

130 Recently, the principles of portfolio optimization have been successfully



131 applied not only for optimizing financial portfolio selection [24], but also in  
 132 other domains, such as strategic decision making [14] (for instance, team  
 133 management), projects selection [11], IT project portfolio management [3],  
 134 and evolutionary algorithms selection [29]. For instance, for evolutionary  
 135 algorithms selection it is important to keep good, but different individuals,  
 136 which should avoid fast convergence of the population to a single individual  
 137 or few similar individuals. Hence, the selection procedure should simultane-  
 138 ously optimize quality and diversity of population. In [29], a multiobjective  
 139 evolutionary algorithm based on the portfolio selection idea was introduced  
 140 and results comparable to the results of the state-of-the-art algorithms were  
 141 obtained.

## 142 *2.2. A posteriori Markowitz model*

143 The general idea of the a posteriori approach to solving MOO problems  
 144 rephrased in terms of portfolios is: first, to compute the set of efficient (or  
 145 non-dominated) portfolios and, then, to select a single portfolio from it. The  
 146 selection of a final portfolio can be done by the decision maker or expert,  
 147 e.g. with the help of multi-criteria decision aiding approaches, see e.g. [2].  
 148 Given two objective functions, in our case  $\sigma^2(\mathbf{x}) = \mathbf{x}^\top \mathbf{Q} \mathbf{x}$  and  $E(\mathbf{x}) = \mathbf{r}^\top \mathbf{x}$ ,  
 149 one can associate to each solution  $\mathbf{x}$  a 2-dimensional evaluation vector in the  
 150 objective space,  $(\sigma^2(\mathbf{x}), E(\mathbf{x}))^\top$ , where the risk objective is to be minimized  
 151 and the return objective is to be maximized;  $\mathbf{r}$  and  $\mathbf{Q}$  are defined as before  
 152 in (1).

153 A portfolio  $\mathbf{x}^{(1)}$  dominates a portfolio  $\mathbf{x}^{(2)}$  (in symbols  $\mathbf{x}^{(1)} \prec \mathbf{x}^{(2)}$ ), if  
 154 and only if  $E(\mathbf{x}^{(1)}) \geq E(\mathbf{x}^{(2)})$  and ( $\sigma^2(\mathbf{x}^{(1)}) < \sigma^2(\mathbf{x}^{(2)})$  or  $E(\mathbf{x}^{(1)}) > E(\mathbf{x}^{(2)})$ )  
 155 and  $\sigma^2(\mathbf{x}^{(1)}) \leq \sigma^2(\mathbf{x}^{(2)})$ . The efficient set  $X_E$  (of portfolios) is given by the  
 156 portfolios that are not dominated by any other portfolio. The image of this  
 157 set is called the Pareto front  $PF$ , i. e.

$$PF = \{(y_1, y_2)^\top \in \mathbb{R}^2 \mid \exists \mathbf{x} \in X_E : y_1 = \sigma^2(\mathbf{x}) \text{ and } y_2 = E(\mathbf{x})\}.$$

158 An example of Pareto fronts of optimal portfolios can be seen in Figure  
 159 4. Note that we chose the first coordinate ( $y_1$ ) for the risk objective (or  
 160 variance), and the second coordinate ( $y_2$ ) for the expected return objective,  
 161 thereby following the convention in portfolio optimization.

162 It should be noted that here the formulation (1) is adapted from the  
 163 continuous version to a discrete, in particular an *integer* one. In integer  
 164 adaptation an asset is either taken or not at a fixed price. The search space  
 165 of the problem  $\mathcal{S}$  is  $\{0, 1\}^{N_{Total}}$ .

166 The Pareto front will be obtained at the upper left boundary of the  
 167 set of attainable solutions  $Y = \{(y_1, y_2)^\top \in \mathbb{R}^2 \mid \exists \mathbf{x} \in \{0, 1\}^{N_{Total}} : y_1 =$   
 168  $\sigma^2(\mathbf{x}) \text{ and } y_2 = E(\mathbf{x})\}$ . It can, for instance, be obtained by a series of con-  
 169 strained single objective optimization problems. Moreover, a fixed budget  
 170  $B$  is allocated for buying assets of a portfolio, which in research projects is  
 171 lost if not spent. Hence, the *integer adaptation of the Markowitz model* is as

172 follows:

$$\begin{aligned} E(\mathbf{x}) &\rightarrow \max; \\ \sigma^2(\mathbf{x}) &\rightarrow \min; \\ \text{s.t. } \sum_{i=1}^{N_{Total}} c_i x_i = \mathbf{x}^\top \cdot \mathbf{c} &\leq B; \\ x_i &\in \{0, 1\}, i = 1, \dots, N_{Total}, \end{aligned} \tag{2}$$

173 where  $c_i$  refers to the cost of an asset, which can be different for different  
174 assets.

175 The set of feasible portfolios  $\mathcal{F}$  is now the subset of portfolios in  $\mathcal{S}$  with  
176  $\sum_{i=1}^{N_{Total}} c_i x_i = \mathbf{x}^\top \cdot \mathbf{c} \leq B$ .

177 Earlier VS for drug discovery was formulated as a multiobjective opti-  
178 mization problem in [17], where both activity and diversity were maximized  
179 simultaneously. Our portfolio-based formulation is similar, however, different  
180 diversity measure based on Solow-Polasky diversity [23] is used, see section  
181 3.2, and expected return based on activity is computed instead of activity  
182 score maximization.

### 183 2.3. *A priori Sharpe ratio model*

184 All portfolios belonging to the efficient set present tradeoffs between re-  
185 turn and risk. Eventually however, from the set of efficient portfolios a single  
186 one should be chosen. Instead of letting the decision maker make a subjec-  
187 tive decision by viewing solutions on the Pareto front (a posteriori decision

188 making) one could also establish beforehand a criterion by which the best  
189 solution on the Pareto front is selected (a priori decision making).

190 The investment management suggests a large number of measures to eval-  
191 uate return-to-risk ratios of portfolios, relatively to time period (e. g., stan-  
192 dard deviation), to market behavior (e. g., beta ratio), to benchmark asset  
193 (e. g., tracking error, excess return, Sharpe ratio). The Sharpe ratio, also  
194 called reward-to-volatility ratio, is the most widely used risk-adjusted per-  
195 formance index [5] and will be used here.

196 The Sharpe ratio can be defined with the help of the capital allocation line  
197 (CAL). It is a straight line on the return-risk graph (see Figure 1) that shows  
198 all possible combinations of risky portfolios with the risk-free asset  $r_f \geq 0$ .  
199 The risk-free asset,  $r_f$ , has a return that is smaller than the minimal expected  
200 return of an efficient portfolio  $r_f < r_{min}$ , and it assumes risk-free investment.  
201 The optimal CAL corresponds to the portfolios with lowest risk for any given  
202 value of return  $r > r_f$ . The slope of the optimal CAL is a sub-derivative of  
203 the function that defines the Pareto front of efficient portfolios. The point  
204 at which the CAL touches the front of efficient portfolios corresponds to the  
205 Sharpe ratio that provides an optimal risky portfolio.

Here, the risk free investment is chosen to be  $r_f = -B$ , as this will be  
the exact return if we do not invest in the research. Then the Sharpe ratio  
is defined as

$$Sh(\mathbf{x}) = \frac{E(\mathbf{x}) - r_f}{\sigma(\mathbf{x})}.$$

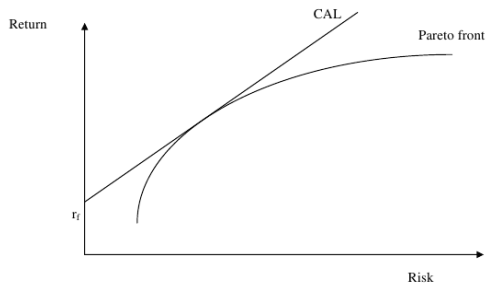


Figure 1: Sharpe ratio on intersection of CAL and Pareto front

206 The Sharpe ratio characterizes how well the return of a portfolio compensates  
 207 the risk taken, and it measures excess of return per unit of risk. When  
 208 comparing two portfolios, the one with the higher Sharpe ratio gives more  
 209 return per risk. Finding the portfolio with maximal *Sharpe ratio* yields the  
 210 following nonlinear integer programming problem:

$$\begin{aligned}
 \frac{E(\mathbf{x}) - r_f}{\sigma(\mathbf{x})} &\rightarrow \max; & (3) \\
 \text{s.t. } \mathbf{x}^\top \cdot \mathbf{c} &\leq B; \\
 x_i &\in \{0, 1\}, i = 1, \dots, N_{Total},
 \end{aligned}$$

211 where  $B$  refers to the budget, which in research projects if not spent is lost.

#### 212 2.4. Portfolios with fixed size

213 The problem with the Markowitz (2) and optimal Sharpe ratio (3) for-  
 214 mulations is that they both favor selection of empty portfolios as they may  
 215 be best at minimizing risk of any losses. One way to neutralize this effect  
 216 is to require a fixed number of assets to be selected into the portfolio. This

217 problem formulation is referred to as *fixed size portfolio selection* and it as-  
218 sumes that the number of assets to be selected is limited to a specific number  
219  $N_{Portfolio}$ . Then, in addition to the formulation (2) or (3), a constraint of the  
220 following form is assumed:

$$\mathbf{x}^\top \cdot \mathbf{e} = N_{Portfolio},$$

221 where  $\mathbf{e}$  is in  $\{0, 1\}^{N_{Total}}$ ; each coordinate is either 0 or 1, summing up to  
222 portfolio of  $N_{Portfolio}$  size (with  $N_{Portfolio} \ll N_{Total}$  not all molecules being  
223 selected in portfolio out of  $N_{Total}$ ).

224 This formulation is equivalent to the 0-1 quadratic knapsack problem.  
225 Problems of this form were intensively studied in the literature due to their  
226 simple and practical formulation, but there are difficulties in finding exact  
227 solutions for them (as indicated in [16] and [20]).

### 228 **3. Drug subset selection as portfolio optimization**

229 Several formulations from the previous section can be used for selecting  
230 portfolio of molecules that are potential drugs. For formulating such prob-  
231 lems the following model variables are considered:

- 232 1. A fixed budget  $B$  is available and has to be spent. Money that is not  
233 used will be lost.
- 234 2. Each successful molecule is associated with a gain  $G$ , which is the value  
235 (expressed in monetary units) gained if the molecule becomes a drug.

236 The gain is the same for each successful molecule  $G_i = G$  and is zero  
237 for unsuccessful molecules.

238 3. For each available molecule  $i = 1, \dots, N_{Total}$  a probability of success  $p_i$   
239 is given or obtained a priori.

240 4. For each candidate molecule  $i = 1, \dots, N_{Total}$  the cost  $c_i$  for buying  
241 and testing it is known. The cost of different molecules may vary  
242 significantly, and this cost does not involve indirect costs, e. g. costs of  
243 the *in vitro* testing.

244 5. From a given set of  $N_{Total}$  candidates, a subset of  $N_{Portfolio}$  molecules  
245 is selected such that

246 (a) the budget  $B$  is not exceeded.

247 (b) The expected return  $E$  is to be maximized, where the expected  
248 return is given by the expected value of the random variable of  
249 the return  $R$  of a portfolio of molecules selected for testing.

250 (c) The risk  $\sigma$  associated with the expected return is to be minimized.

### 251 3.1. *A posteriori Markowitz model with fixed size portfolio*

252 The problem corresponding to a posteriori Markowitz model with limited  
253 budget and fixed size of portfolio constitutes a two-objective optimization  
254 problem that is formulated as follows:

$$\begin{aligned}
E(\mathbf{x}) &\rightarrow \max; & (4) \\
\sigma^2(\mathbf{x}) &\rightarrow \min; \\
\text{s.t. } \mathbf{x}^\top \cdot \mathbf{c} &\leq B; \\
\mathbf{x}^\top \cdot \mathbf{e} &= N_{Portfolio}; \\
x_i &\in \{0, 1\}, i = 1, \dots, N_{Total},
\end{aligned}$$

255 and is referred in the text as the *Markowitz model with fixed size portfolios*.

256 Here,  $x_i$ ,  $i = 1, \dots, N_{Total}$ , denote the decision variables;  $x_i = 1$  means  
257 that the  $i$ -th molecule is selected and  $x_i = 0$  means that it is not selected;  
258  $N_{Total}$  is the number of available molecules. The search space of the problem  
259  $\mathcal{S}$  is  $\{0, 1\}^{N_{Total}}$ . The set of feasible portfolios  $\mathcal{F}$  is now the subset of portfolios  
260 in  $\mathcal{S}$  with  $\sum_{i=1}^{N_{Total}} x_i = N_{Portfolio}$ , where  $N_{Portfolio}$  is the size of the portfolio.  
261 We consider two real valued objective functions defined on  $\mathcal{S}$ ,  $\sigma^2(\mathbf{x}) = \mathbf{x}^\top \mathbf{Q} \mathbf{x}$   
262 and  $E(\mathbf{x}) = \mathbf{r}^\top \mathbf{x}$ . Each portfolio  $\mathbf{x}$  is associated with a 2-dimensional evalu-  
263 ation vector in the objective space,  $(\sigma^2(\mathbf{x}), E(\mathbf{x}))^T$ , where the risk objective  
264 is to be minimized and the return objective is to be maximized;  $\mathbf{r}$  and  $\mathbf{Q}$  are  
265 defined as before in (1).

The computation of return  $E(\mathbf{x})$  and risk  $\sigma^2(\mathbf{x})$  is discussed next. The  
return  $E(\mathbf{x})$  is defined as the gains minus the losses. For the expected return  
it is important to realize that money from the budget that is not invested in  
molecules is lost. Therefore, the losses will be  $B$  and the gains will be the



cumulated gains from molecules that become successful drugs. Hence,

$$E(\mathbf{x}) = G \mathbf{p} \cdot \mathbf{x} - B = \left( \sum_{i=1}^{N_{Total}} G p_i x_i \right) - B.$$

Due to the probabilistic nature of the return (we get it only in case of successful drug(s)), it can be modeled as a random variable. Let  $\tilde{x}_i$  denote a random variable of Bernoulli type that models the uncertain return on investment in a molecule  $i$ :

$$\tilde{x}_i = \begin{cases} \frac{G - c_i}{c_i}, & \text{return rate with probability } p_i \text{ in case of success;} \\ \frac{0 - c_i}{c_i} = -1, & \text{return rate with probability } 1 - p_i \text{ in case of no success.} \end{cases}$$

Then, the expected return of a molecule  $i$  is defined by:

$$E(\tilde{x}_i) = \frac{G - c_i}{c_i} \cdot p_i + \frac{-c_i}{c_i} \cdot (1 - p_i).$$

266 Following the classical model of Markowitz, the risk  $\sigma^2(\mathbf{x})$  can be ex-  
267 pressed by means of a covariance matrix  $\mathbf{Q}$  as follows:

$$\sigma^2(\mathbf{x}) = \mathbf{x}^\top \mathbf{Q} \mathbf{x} = \sum_{i=1}^{N_{Total}} \sum_{j=1}^{N_{Total}} x_i q_{ij} x_j,$$

268 where  $q_{ij}$  is a correlation between the return from the  $i$ -th molecule  $r_i$  and  
269 the return from the  $j$ -th molecule  $r_j$ . The computation of the covariance on  
270 the basis of a distance matrix will be derived from the Solow-Polasky model

271 as discussed in the next section.

### 272 3.2. Solow-Polasky diversity measure

273 One possible interpretation of the covariance can be done using a measure  
274 for estimating diversity of a (biological) population introduced by Solow and  
275 Polasky (see [23]). Originally, they were searching for a measure that can be  
276 used for evaluating population diversity rigorously, assuming some particular  
277 properties for this measure are respected. A measure which counts essentially  
278 different species and is used in the context of species preservation. Within  
279 the utilitarian model of Solow and Polasky the more species is considered  
280 to be more useful because of e.g. their potential future medical benefits. In  
281 general there are other reasons for species preservation, e.g. for stability of  
282 eco-system or ethical reasons. But in our context utilitarian motivation for  
283 species preservation fits well.

284 Hence, they suggested a diversity function:

$$D(\mathbf{s}) = \mathbf{e}^\top F(\mathbf{s})^{-1} \mathbf{e},$$

285 where  $\mathbf{e}$  is an  $N_{Total}$ -vector of 1's and  $F(\mathbf{s})$  is a non-singular  $N_{Total}$ -by- $N_{Total}$   
286 distance matrix  $F(\mathbf{s}) = [f(d(s_i, s_j))]$ , with a distance function  $f(d_{ij})$  taken  
287 for each pair of species  $d(s_i, s_j)$ . Each entry of the Solow-Polasky matrix  
288 indicates distance between species  $s_i$  and  $s_j$ , where  $i = 1, \dots, N_{Total}$  and  
289  $j = 1, \dots, N_{Total}$ .

290 When compared to other diversity measures, e.g. proposed in [28], Solow-

291 Polasky distance takes into account not only the distance between species in  
292 the population but also provides a measure for the number of different species  
293 in it. This model is inspired by probabilistic modeling of a set of species, but  
294 can be adapted to drug discovery.

295 Let  $S = \{s_1, s_2, \dots, s_{N_{Total}}\}$  be a set of molecules,  $|S| = N_{Total}$ . Let  $S'$  be  
296 any subset of  $S$ , then  $B(S')$  denotes the composite event that at least one  
297 molecule in  $S'$  is successful. By  $Pr(B(S'))$  we denote the probability of this  
298 composite event. The expected benefit of  $S'$  can be measured by the product  
299  $Pr(B(S')) \cdot V$ , where  $V$  is a fixed unit value of benefit. Based on this benefit  
300 measure different subsets of  $S$  can be compared.

301 Knowing a priori information on the performance of different molecules  
302 with respect to the specified goal(s), the probability of their benefit can be  
303 defined as  $Pr(B_i) = p_i$ , where  $B_i$  denotes the event that the  $i$ -th molecule is  
304 successful. Otherwise, if probabilities are unknown, they may be considered  
305 as equal  $Pr(B_i) = p$  for all  $B_i$ ,  $i = 1, \dots, N_{Total}$ . For the event  $B_j$  being  
306 successful, the conditional probability for the event  $B_i$  is defined in [23] as  
307  $Pr(B_i|B_j) = p + (1-p)f(d_{ij})$ , where  $f$  is a function selected with the following  
308 properties:  $f(0) = 1$ ,  $f(\infty) = 0$ ,  $f' \leq 0$ . Here, as remarked by Solow and  
309 Polasky,  $f$  can be interpreted as a correlation function.

310 Finding the  $N_{Total}$ -variate distribution  $Pr(B(S))$  from univariate and bi-  
311 variate probabilities is not possible. However, the lower bound on it was  
312 defined in [10]. One example of the distance function for computing this  
313 distance matrix is provided in [23]:  $f(d) = e^{-\theta d(s_i, s_j)}$ , and it will be used

314 here.

### 315 3.3. A posteriori Markowitz model with Solow-Polasky diversity

316 In addition to difficulties with computing an exact solution for a fixed  
317 size portfolio, in some cases the tendency of selecting the cheapest solutions  
318 in the portfolio may be observed if enough diversity is reached at the cost of  
319 cheapest assets. Hence, relaxing the constraint on the number of assets in  
320 the portfolio may be beneficial.

321 Fortunately, Solow and Polasky specified a set of requirements which a  
322 biological diversity measure should satisfy; see [23] for more details. One  
323 of the requirements is *monotonicity in species*, which suggests that the di-  
324 versity of a set increases with adding new elements to it and decreases with  
325 removing elements. This property is taken into account in the next portfolio  
326 optimization model.

327 Since minimizing risk of selecting similar assets into a portfolio can also be  
328 interpreted as maximizing diversity of selected portfolio of assets, different  
329 formulations of diversity can be taken in the portfolio selection problem.  
330 Here, we propose to use Solow-Polasky diversity as a second objective instead  
331 of the risk measure calculated as a variance of the returns.

332 The Solow-Polasky diversity measure is calculated as the sum of the en-  
333 tries of the inverse of the correlation matrix for selected assets:

$$D(\mathbf{x}) = \mathbf{e}^\top F(\mathbf{x})^{-1} \mathbf{e} = \sum_{i=1}^{N_{Total}} \sum_{j=1}^{N_{Total}} F(\mathbf{x})_{ij}^{-1},$$

334 where  $F(\mathbf{x})_{ij}^{-1}$  is the inverse of the correlation matrix for all selected assets.

335 Then, the two objectives to be optimized are: the return and the diversity  
336 of the portfolio, which can be presented in the following model:

$$\begin{aligned} E(\mathbf{x}) &\rightarrow \max; & (5) \\ D(\mathbf{x}) &\rightarrow \max; \\ \text{s.t. } \mathbf{x}^\top \cdot \mathbf{c} &\leq B; \\ x_i &\in \{0, 1\}, i = 1, \dots, N_{Total}. \end{aligned}$$

337 Even though both a posteriori approaches use the correlation function sug-  
338 gested by Solow and Polasky, the Markowitz model minimizes the sum of  
339 the correlation matrix entries, while the Solow-Polasky diversity model max-  
340 imizes the sum of the entries of the inverse of the correlation matrix. Hence,  
341 the former model favors smaller size portfolios, while the latter one gives  
342 preference to larger portfolios.

#### 343 4. Solution algorithms

344 Different methods can be used to compute efficient portfolios to the given  
345 portfolio selection problem. In this section, the methods that proved to be  
346 robust solvers are presented. In general, the difficulty of finding efficient  
347 portfolios depends on the number of candidate molecules  $N_{Total}$  and the size  
348 of the subset that is selected  $N_{Portfolio}$ .

349 Portfolio optimization problems belong to the class of NP hard problems

350 and, under the  $P \neq NP$  assumption, the effort needed to solve them ex-  
351 actly is growing exponentially with increasing  $N_{Total}$ . Portfolio optimization  
352 problems can be formulated either as discrete or continuous/parametric op-  
353 timization problems. The former presentation is more common due to faster  
354 performance on small and medium size problems (with up to 500 assets) with  
355 interior-point optimizers. However, in [24], it was shown that for large-scale  
356 problems (in the range of 1,000 to 3,000 assets) continuous formulation may  
357 be computationally more efficient when solved with some optimizers. Re-  
358 cently, new exact solvers such as Gurobi (see [12]) show fast performance for  
359 large instances (at least with datasets with up to 5,000 assets considered in  
360 this work) with branch and bound method.

361 However, for finding the Pareto fronts in Markowitz models and com-  
362 puting Sharpe ratio, some adaptations to the formulations presented earlier  
363 need to be performed before applying exact solvers. This will be discussed  
364 next, first for the Pareto front computation in the Markowitz model with  
365 fixed size portfolio and then for the Sharpe ratio maximization. For the case  
366 of the Markowitz model with Solow-Polasky diversity optimization instead  
367 of the original risk objective, exact solvers cannot be applied due to the  
368 complexity of the risk objective function. But approximate algorithms, such  
369 as meta-heuristics and multiobjective evolutionary algorithms in particular,  
370 could and will be applied to find approximate solutions.

371 4.1. Markowitz model with fixed size portfolio computation using  $\epsilon$ -constraint  
 372 method

373 To find the Pareto front and efficient set of the problem, it is proposed  
 374 to use the  $\epsilon$ -constraint method, see e.g. [18]. This is done by formulating a  
 375 series of single objective constrained optimization problems (SOCOPs) with  
 376 moving constraint on one of the objective function values. Then one objective  
 377 is optimized subject to the other objective fixed and expressed as a constraint.  
 378 To obtain, say  $N_{Pareto}$ , points on the Pareto front, we solve the following series  
 379 of  $N_{Pareto}$  SOCOPs for ascending expected returns  $E_j$ ,  $j = 1, \dots, N_{Pareto}$ :

$$\begin{aligned}
 \sigma^2(\mathbf{x}) &\rightarrow \min; & (6) \\
 \text{s.t. } E(\mathbf{x}) &\geq E_j; \\
 \mathbf{x}^\top \cdot \mathbf{c} &\leq B; \\
 \mathbf{x}^\top &= N_{Portfolio}; \\
 x_i &\in \{0, 1\}, i = 1, \dots, N_{Total}.
 \end{aligned}$$

380 The resulting optima will be called  $\mathbf{x}_j^*$ , and their risk  $\sigma_j^{2*}$  and return  $E_j^*$   
 381 values. The values of  $E_j^*$  are taken evenly spaced between lower bound  $E^{\min}$   
 382 and upper bound  $E^{\max}$ . The computation of the lower and upper bounds,

383  $E^{\min}$  and  $E^{\max}$ , is done by solving the SOCOs, respectively:

$$\begin{aligned}
 E(\mathbf{x}) &\rightarrow \min; & (7) \\
 \text{s.t. } \mathbf{x}^\top \cdot \mathbf{c} &\leq B; \\
 x_i &\in \{0, 1\}, i = 1, \dots, N_{Total},
 \end{aligned}$$

384 and

$$\begin{aligned}
 E(\mathbf{x}) &\rightarrow \max; & (8) \\
 \text{s.t. } \mathbf{x}^\top \cdot \mathbf{c} &\leq B; \\
 x_i &\in \{0, 1\}, i = 1, \dots, N_{Total}.
 \end{aligned}$$

385 Let  $\mathbf{x}^{\min}$  denote the solution obtained for the first problem (7) and  $\mathbf{x}^{\max}$   
 386 denote the solution obtained for the second problem (8). Then, the lower  
 387 bound for the return is  $E^{\min} = E(\mathbf{x}^{\min})$  and the upper bound for the return  
 388 is  $E^{\max} = E(\mathbf{x}^{\max})$ .

389 *4.2. Sharpe ratio with fixed size portfolio computation using quadratic pro-*  
 390 *gramming*

391 In order to maximize the Sharpe ratio, it would be beneficial to get rid of  
 392 the nonlinear and non-quadratic term  $\frac{E(\mathbf{x})-r_f}{\sigma(\mathbf{x})}$  in the problem formulation (3),  
 393 and then use a quadratic solver. For this, homogenization has been suggested  
 394 in [5]. However, our experience was that the resulting mixed integer quadratic  
 395 programming (QP) problem was difficult to solve due to resulting covariance



396 matrix being not of a semidefinite type.

Alternatively, it is also possible to compute the Pareto front with (1) and find the point on the Pareto front that maximizes the Sharpe ratio computed with (3). Given a sufficiently dense approximation of the Pareto front this is accomplished by evaluating the Sharpe ratio of all points on the Pareto front i.e.:

$$\mathbf{x}^{sharpe} = \arg \max \{Sh(\mathbf{x}_1), \dots, Sh(\mathbf{x}_{N_{Pareto}})\}.$$

397 It is important in this context that points that maximize the Sharpe ratio  
398 are part of the efficient set.

#### 399 *4.3. Markowitz model with Solow-Polasky diversity computation using multi-* 400 *objective genetic algorithms*

401 In case of Markowitz model with Solow-Polasky diversity considered as  
402 a risk objective (5), the need of obtaining the inverse of the distance matrix  
403 makes the application of quadratic programming difficult. An alternative ap-  
404 proach is to use approximate methods, for instance, meta-heuristics. While  
405 meta-heuristics do not guarantee reaching an optimal solution, they can typ-  
406 ically obtain good approximations to optima fairly quickly even for NP hard  
407 combinatorial problems, which is the case of knapsack / portfolio optimiza-  
408 tion problems considered in this work.

409 Among many meta-heuristics developed so far, multiobjective evolution-  
410 ary algorithms (MOEAs) are particularly common for solving multi-objective  
411 optimization problems. In this study, two common MOEAs are considered:

412 NSGA-II (see [6]) and SMS-EMOA (see [7]). Using otherwise standard imple-  
413 mentations of these meta-heuristic solvers, we introduce two problem specific  
414 adaptations. These are the mutation and the recombination operators, which  
415 were specifically designed for the subset selection problem.

416 MOEAs maintain a population (multi-set) of individuals that is changing  
417 over time due to the application of variation and selection operators. From a  
418 given population  $P(t)$  at time  $t$  pairs of parents are selected – in the so-called  
419 mating selection step – and offspring are then generated by recombination  
420 and mutation based on these parents. Then from the offspring and the  
421 individuals of previous population  $P(t)$  a set of individuals is selected – in  
422 the so-called environmental selection step – that forms the next population  
423  $P(t+1)$ . While the two selection steps are based on choosing individuals with  
424 the best objective function values, the two variation steps – recombination  
425 and mutation – seek to generate new individuals that resemble some of the  
426 traits of their parents. Recombination combines the information of parents,  
427 and mutation does a small random modification of a solution.

428 The NSGA-II and SMS-EMOA algorithms differ in their selection steps:  
429 In NSGA-II, a new offspring population of the same size as the population  
430  $P(t)$  is generated, and, subsequently, the new population  $P(t + 1)$  is se-  
431 lected based on so-called non-dominated sorting and crowding-distance. In  
432 SMS-EMOA, only one offspring is generated based on  $P(t)$  and the next  
433 population  $P(t + 1)$  is obtained by non-dominated sorting and selecting the  
434 subset that maximizes the hypervolume indicator. Here, the hypervolume in-

435 dicator, which the SMS-EMOA seeks to maximize, is a measure computed to  
436 show how well a population serves to mark the boundary between the dom-  
437 inated and non-dominated spaces, and, thus, how well it serves to represent  
438 the true Pareto front. The MOEA for the portfolio subset selection problems  
439 represents individuals (that are portfolios) as *sorted index lists*. For instance,  
440 the sequence (1, 4, 6, 29) represents the portfolio that selects the 1st, the 4th,  
441 the 6th and the 29th molecules.

442 The mutation is done by (1) deleting a single randomly chosen molecule  
443 from the portfolio, (2) adding a randomly chosen new molecule, and (3)  
444 replacing a molecule inside the portfolio by a molecule outside the portfolio.  
445 Each of these mutation operators is applied with a certain probability for each  
446 molecule, which is denoted by  $p_{MD}$ ,  $p_{MA}$ , and respectively  $p_{MR}$ . In case of a  
447 fixed number of molecules in the portfolio, only replacement is used. While  
448  $p_{MD}$  and  $p_{MA}$  determine probability of adding and deleting a single molecule  
449 *per portfolio*, the replacement probability  $p_{MR}$  is defined *per molecule in the*  
450 *portfolio*.

451 As a recombination operator  $m$ -point crossover is applied. This means  
452 we randomly select  $m$  points for the number of molecules. After each point  
453 we change the parent we use to copy from. To make it applicable for subsets,  
454 the subset membership is interpreted as a bit-string (one means a molecule is  
455 a member of portfolio, zero means a molecule is not a member of portfolio),  
456 and the crossover determines membership based on either one of the two  
457 parents selected randomly. The probability of crossover is  $p_{CO}$ . If crossover

458 is not applied, then one of the two parents chosen randomly is copied and  
459 will serve as offspring (before mutation).

## 460 5. Experimental results

### 461 5.1. Molecular portfolio selection model assumptions

462 First, the information on a covariance matrix  $Q$  needs to be formulated.  
463 In chemistry, the distance between molecules can be defined by evaluating  
464 similarities/differences in the structure of two molecules. Being able to mea-  
465 sure the distance  $d(x_i, x_j)$  between each pair of molecules  $i$  and  $j$ , provides  
466 means for defining the matrix  $F(\mathbf{x})$  of  $N_{Total}$ -by- $N_{Total}$  size, e.g. as suggested  
467 in [23] with elements  $f_{ij} = e^{-\theta d(x_i, x_j)}$ , where  $\theta$  is set to  $\theta = 0.5$ . The distance  
468 between molecules can be computed based on their similarity, e.g. according  
469 to the Tanimoto similarity, see [25] also used in [22].

470 Tanimoto similarity  $Sim_T$  is a measure of similarity between two bit  
471 vectors  $A$  and  $B$ . The bit-vectors used here are the molecular fingerprints.  
472 A molecular fingerprint is a bit vector, where each bit represents whether a  
473 chemical substructure is part of the molecule (1) or not (0). The Tanimoto  
474 similarity can be defined as:

$$Sim_T(A, B) = \frac{\sum_z A_z \wedge B_z}{\sum_z A_z \vee B_z},$$

475 where the index  $z$  corresponds to a particular property of molecule structure.

476 In this study, circular fingerprints ( $FCFP_4$ ) calculated with Pipeline

477 Pilot 9.0.2.1 were used [21]. To predict activity of molecules, we used a  
478 Proteochemometric model as published by van Westen et al. in [26]. The  
479 molecules selected here originated from the Enamine building blocks [8] with  
480 prices defined per 100mg.

481 Then, we can calculate the distance between two molecules as a dissimi-  
482 larity measure, which is diversity:

$$d(x_i, x_j) = 1 - Sim_T, \quad (9)$$

483 where  $Sim_T$  is Tanimoto similarity.

484 Second, the information on (bio-)activities of the candidate molecules  
485 needs to be translated into success probabilities. Activity  $a_i$  is normally  
486 given as logarithmized activity  $l_i$ ; in this case, we can use  $a_i = e^{l_i}$ .

487 Moreover, from experience chemists know an average probability of suc-  
488 cess  $\bar{p}$ , for the sake of the argument estimated as  $\bar{p} = 1/100$ . Let us consider  
489 a vector of  $N_{Total}$  activities (exponentiated)  $\mathcal{A} = \{a_1, \dots, a_{N_{Total}}\}$  and let  
490  $\mathcal{P} = \{p_1, \dots, p_{N_{Total}}\}$  denote the success probabilities. Then, the average  
491 probability of success can be calculated as:

$$\bar{p} = \frac{1}{N_{Total}} * \sum_{i=1}^{N_{Total}} p_i, \quad (10)$$

492 and we know that activities are proportional to success probability. Hence,

493 for some constant  $k$  it holds:

$$p_i = k * a_i, \forall i = 1, \dots, N_{Total}. \quad (11)$$

494 By substituting  $p_i$  in (10) as defined in (11), we can obtain  $k$ :

$$k = \bar{p} * \frac{N_{Total}}{\sum_{i=1}^{N_{Total}} a_i}. \quad (12)$$

495 Combining 11 and 12 we get:

$$p_i = a_i * \bar{p} * \frac{N_{Total}}{\sum_{i=1}^{N_{Total}} a_i}. \quad (13)$$

496 Third, the gain from a new lead compound (i.e. a molecule that may  
497 lead to a new drug) may vary between, e.g.  $G_L = 10,000$  and  $G_U = 100,000$   
498 USD.

499 Fourth, several findings for the current drug portfolio selection model are  
500 based on the analysis of these model assumptions. Since the return of each  
501 molecule, which is equal to the product of gain and probability of success  
502 (for  $G_L$   $r_i = 10,000 * 0.0001 = 1$  or for  $G_U$   $r_i = 100,000 * 0.0001 = 10$ ), is  
503 very small, it turns out that it is not profitable to invest into molecules in the  
504 early stages of drug discovery. Besides having economical profitability, it is  
505 often the case that a budget for drug discovery is made available in research  
506 projects to stimulate medical innovation.

507 Fifth, for the fixed size portfolio model, we assume that 100 molecules

508 need to be selected out of each dataset into the portfolio of molecules to be  
509 tested *in vitro*:  $N_{Portfolio} = 100$ .

510 Sixth, the budget to be spent and to be taken into account as a constant in  
511 the model is calculated assuming a fixed number of molecules  $N_{Portfolio} = 100$   
512 will be bought. Hence, the budget can be obtained as an average cost  
513 multiplied by the number of molecules to be bought:  $B = N_{Portfolio} * \sum_{i=1}^{N_{Total}} c_i / N_{Total}$ .

514  
515 Seventh, the budget is set to a hundred times the average cost of molecules  
516 in the dataset:  $B = 100 * \bar{c}$ . For the dataset of 1000 molecules this yields  $B =$   
517 34,502USD, for the dataset of 2500 molecules this yields  $B = 34,400$ USD  
518 and for the dataset of 5000 molecules this results in  $B = 34,622$ USD.

519 Eighth, based on comparison of performance of the algorithms on all three  
520 datasets, it was observed that larger datasets perform better, when compared  
521 to smaller datasets, assuming the same fixed number of 100 molecules is  
522 selected from all three datasets. One could argue that this may be the result  
523 of applying more iterations in the bigger datasets. However, this is not  
524 the case as all datasets converge after 100,000 iterations, which means that  
525 running the algorithms for more iterations will not be effective.

526 The reason for this behavior is the way success probabilities of molecules  
527 are computed: The success probability of a molecule is calculated inversely  
528 proportional to average activity of all molecules belonging to the dataset.  
529 It would be a correct approach if the datasets would be uniformly selected  
530 from the vendor database. However, in our case the datasets were sorted by

531 activity before selection and from the same sorted dataset top 1000, 2500 and  
532 5000 most active molecules were selected. The datasets are sorted based on  
533 activity because chemists usually do not consider molecules below the cut-off  
534 activity. Thus, larger datasets, e.g. with 2500 and 5000 molecules, contain  
535 molecules with lower activity on average, when compared to smaller datasets  
536 with higher average activity, e.g. with 1000 molecules.

537 To avoid the situation when the average success probability of a given  
538 molecule is lower in a bigger dataset when compared to the average suc-  
539 cess probability of the same molecule in a smaller dataset, its calculation is  
540 adjusted. In particular, the average probability of success of a molecule is  
541 computed in such a way that it is independent of the size of the considered  
542 dataset, and in such a way it suits better to datasets with a non-uniform  
543 distribution of activities.

544 As before it is assumed that success probability is proportional to the ac-  
545 tivity. However, now the average probability  $\bar{p}_{1000}$  is fixed to be proportional  
546 to the average activity  $\bar{a}_{1000}$  of the 1000 molecules dataset:

$$\bar{p}_{1000} = k * \bar{a}_{1000}, \forall i = \{1, \dots, N_{Total}\}. \quad (14)$$

547 which leads to the  $k$  computed as:

$$k_{1000} = \bar{p}_{1000} * \frac{1}{\bar{a}_{1000}}, \quad (15)$$



548 Thus, the probability of success of each molecule can be computed as:

$$p_i = a_i * \bar{p}_{1000} * \frac{1}{\bar{a}_{1000}}. \quad (16)$$

549 Hence, this fixed average probability  $\bar{p}_{1000}$  of the 1000 molecules dataset will  
550 be used for computing the probabilities of success of molecules  $p_i$  in the  
551 datasets with 2500 and 5000 molecules.

### 552 5.2. Molecular compounds datasets

553 For testing efficiency of the proposed models for molecule subset selection,  
554 we have used 3 datasets of 1000, 2500 and 5000 molecules taken from the  
555 ZINC database of molecular compounds (see [13]), as available at vendor  
556 Enamine. Each molecule was provided with its known structure and its  
557 cost per 100mg. The Tanimoto similarity was calculated for each pair of  
558 molecules.

559 These three datasets are demonstrated in Figure 2 (a) with activity and  
560 cost of the 1000 molecules set depicted in (dark) green, the 2500 molecules  
561 set depicted in green and (light) pink, and the 5000 molecules set depicted  
562 in green, pink and blue.

### 563 5.3. Experimental settings for MOEAs

564 In the experiments of this study, the following settings are used:  $p_{MA} =$   
565 0.5 (per portfolio),  $p_{MD} = 0.1$  (per portfolio), and  $p_{MR} = 0.01$  (per molecule).  
566 The number of crossover points was set to 1, and the probability of crossover

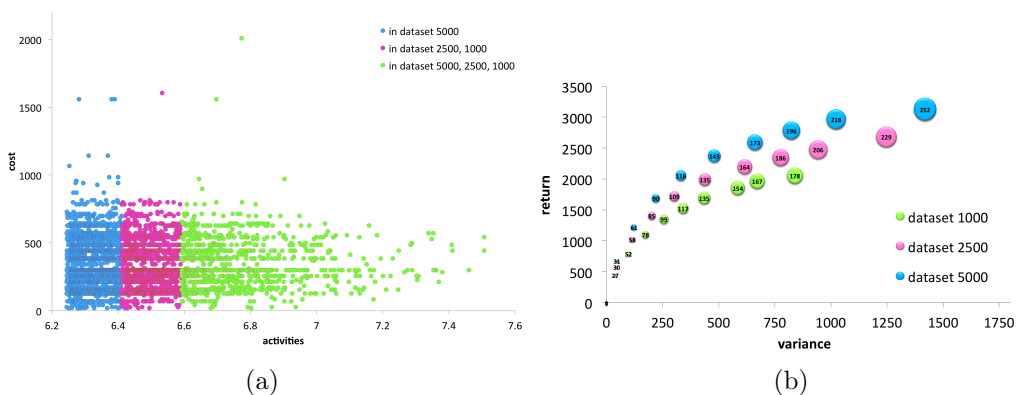


Figure 2: (a) Cost and activity of molecules in three data sets. (b) Size of the population of portfolios in a single run (depicted in a circle) of SMS-EMOA for three data sets

567 was set to  $p_{CO} = 0.2$ . In other words, for every 10 offspring there are 2 that  
 568 have been created using 2 parents, while the other 8 offspring are copies of  
 569 some parents. Replacing a molecule in the portfolio with  $p_{MR} = 0.01$  means  
 570 replacing one molecule per offspring on average. A molecule is added to half  
 571 of the offspring on average:  $p_{MA} = 0.5$ , and is removed from a portfolio  
 572 once per 10 offspring on average:  $p_{MD} = 0.1$ . The size of the population of  
 573 portfolios  $P(t), t = 1, 2, \dots$  was set to 10 in order to conform with the setting  
 574 we used to sample the Pareto front by means of quadratic programming.

575 For fair comparison of MOEAs, in all experiments we run NSGA-II for  
 576 10,000 iterations and SMS-EMOA for 100,000 for the dataset of 1000 molecules.  
 577 This is due to the fact that SMS-EMOA creates 1 offspring at each iter-  
 578 ation, whereas NSGA-II creates 100 offspring at each iteration. For the  
 579 larger datasets, we increased the number of iterations with the same factor  
 580 as the dataset size. That is, for the 2500 molecules dataset we ran NSGA-

581 II for 25,000 iterations and SMS-EMOA for 250,000 iterations, whereas for  
582 the 5000 molecules dataset, we ran NSGAII for 50,000 iterations and SMS-  
583 EMOA for 500,000 iterations.

584 Due to the design of mutation operator used in this work, which allows not  
585 only replacing molecules in portfolio, but also adding or removing portfolios  
586 in the population, the population size varies. Figure 2 (b) gives insight into  
587 the cardinality of the sets of portfolios in the population obtained after a  
588 typical run of MOEAs (SMS-EMOA in this case, but similar results were  
589 obtained for NSGA-II).

590 A problem with the model (2) that minimizes risk without a cardinality  
591 constraint can be observed. In particular, this model allows selection of very  
592 small subsets of portfolios, and even the empty set of portfolios, as a part of  
593 the optimal front. Given the model this makes sense as there is no subset of  
594 portfolios with a higher return other than the one with a variance of 0USD.  
595 However, in practice this is undesirable.

#### 596 *5.4. Discussion of the experimental results*

597 We tested the portfolio selection problem models formulated in sections  
598 2 and 3 with the algorithms presented in section 4 on all three datasets  
599 discussed above.

600 All experiments were performed on a desktop PC with an i5 core 3.2 GHz  
601 processor and 4 GB memory under Windows XP operating system. Gurobi  
602 MIP solver version 4.0 was used and MOEAs were encoded in Python version

603 3.3.

604 5.4.1. *Markowitz model with fixed portfolio size*

605 **Gurobi MIP results** In the first experiment we computed Pareto front  
606 of portfolios optimal from the point of view of their return and risk according  
607 to the Markowitz model with fixed size portfolios ( $N_{Portfolio} = 100$ ) using the  
608 formulation (6) as discussed in section 2.2. Here it is assumed that probability  
609 of success for each molecule is proportional to its activity and is computed  
610 by (11), and covariance between molecules is computed based on a distance  
611 as defined in (9).

612 The  $\epsilon$ -constraint approach to MOO was used. In particular, the return  
613 objective was set to a constraint (computed for 15 different points between  
614 lower and upper return bounds,  $E^{\min}$  and  $E^{\max}$ , respectively) and Gurobi  
615 MIP solver utilizing branch and bound method was applied to the three  
616 datasets with 1000, 2500 and 5000 molecules. The results of runs for all  
617 three datasets are presented in Figures 4 (a), (b) and (c), respectively, in  
618 black color.

619 Next, we analyze the content of portfolios belonging to the Pareto front  
620 of optimal portfolios with 100 molecules selected in each portfolio using the  
621 dataset with 2500 molecules as an example. In particular, we show four  
622 heat-maps indicating the similarity of selected molecules for portfolios with  
623 three different return values equal to 0USD, 1104USD and 1449USD and  
624 one randomly selected portfolio of 100 molecules demonstrated in Figure 3

625 (a), (b), (c) and (d), respectively. (Random selection was performed using a  
626 random percent filter in Pipeline Pilot using seed 333.) The darker the color  
627 the more similar the molecules are: the blue color gradient corresponds to a  
628 similarity equal to 1, dodger-blue to a similarity equal to 0.5, and white to a  
629 similarity equal to 0.

630 When compared to the baseline portfolio with 100 randomly selected  
631 molecules depicted in Figure 3 (d), the portfolio with 0USD return value  
632 depicted in Figure 3 (a) looks much more diverse, the portfolio with 1104USD  
633 return value depicted in Figure 3 (b) is slightly more diverse, and the portfolio  
634 with 1449USD return value depicted in Figure 3 (c) is much less diverse.  
635 The portfolio with 1104USD return value shown in Figure 3 (b) shows better  
636 diversity when compared to the baseline and relatively high return portfolios,  
637 being either close to or exactly the portfolio with optimal Sharpe ratio. This  
638 output is in line with the portfolio selection theory, according to which higher  
639 return portfolios are less diverse, since they also have higher risk, and the  
640 lower return portfolios are more diverse and have lower risk.

641 **MOEAs results** Comparison of MOEAs results is not trivial: On the  
642 one hand, due to the randomness of the population initialization and of the  
643 application of the crossover and mutation operators for individuals of the  
644 population, not a single run, but some averaged performance of MOEAs'  
645 several runs should be compared for evaluating performance of each MOEA.  
646 On the other hand, comparison of the convergence of each algorithm is dif-  
647 ficult due to the fact that no true Pareto front is known. Therefore, only

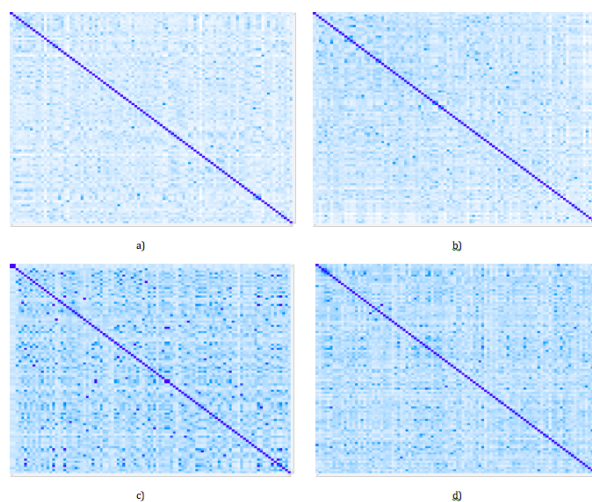


Figure 3: Similarity of 100 molecules portfolios belonging to the Pareto front and selected from the set of 2500 molecules with (a) 0USD return, (b) 1104USD return, (c) 1449USD return and (d) portfolio with 100 randomly selected molecules.

648 a visual comparison can be made based on the attainment surfaces [9] (or  
 649 attainment curves for bi-objective optimization, which is our case) covered  
 650 by each MOEA. This approach allows the comparison of lowest and highest  
 651 Pareto front solutions achieved by each algorithm as well as their average  
 652 performances. For computing the attainment surface a generalization of the  
 653 median as an average is used, which is robust against outliers. In the next ex-  
 654 periments only best front is taken from all runs of an algorithm for comparing  
 655 to other algorithms performance.

656 **Comparison of Gurobi MIP solver and MOEAs result:** We com-  
 657 pared results obtained by the Gurobi MIP solver using branch and bound  
 658 method to the results obtained by two MOEAs, NSGA-II and SMS-EMOA.  
 659 To make comparison fair we run all algorithms for circa 10 minutes for the

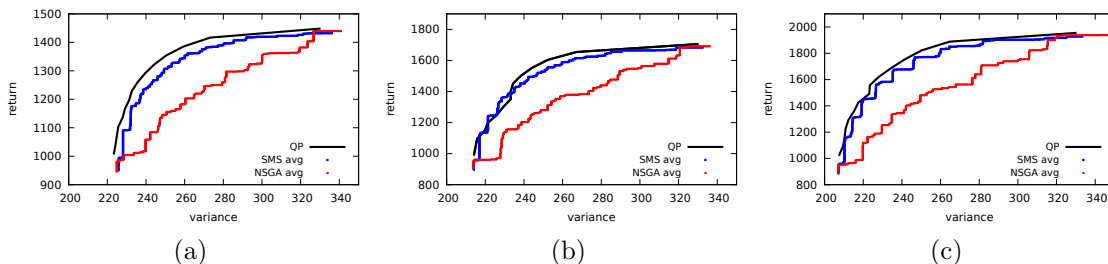


Figure 4: Comparison of the performance of the Gurobi MIP solver and the two MOEAs: NSGA-II and SMS-EMOA (on the plot denoted as QP, NSGA avg and SMS avg, respectively) for the 1000, 2500, 5000 molecules dataset, (a) (b) and (c), respectively.

660 dataset with 1000 molecules, for circa 20 minutes for the dataset with 2500  
 661 molecules, and for circa 30 minutes for the dataset with 5000 molecules.  
 662 The results of this comparison presented in Figures 4 (a), (b) and (c) show  
 663 the best performance of the exact Gurobi MIP solver for the datasets with  
 664 1000 and 5000 molecules, and better performance of the SMS-EMOA when  
 665 compared to NSGA-II on all three datasets. As can be seen from Figure 4  
 666 (b) in some concave regions of the Pareto front SMS-EMOA outperformed  
 667 Gurobi MIP solver, which means that specified time limit was not sufficient  
 668 for branch and bound method of Gurobi MIP solver to find optimal solution.

#### 669 5.4.2. Sharpe ratio with fixed portfolio size

670 We now show and discuss the results obtained for model (3). Figure 5 (a)  
 671 demonstrates values of Sharpe ratio computed for 100 molecules selected from  
 672 the 1000-molecule dataset at each of the 15th iterations of the  $\epsilon$ -constraint  
 673 method. These portfolios belong to the Pareto front of optimal portfolios  
 674 and are obtained with the Gurobi MIP solver. In this case the portfolio

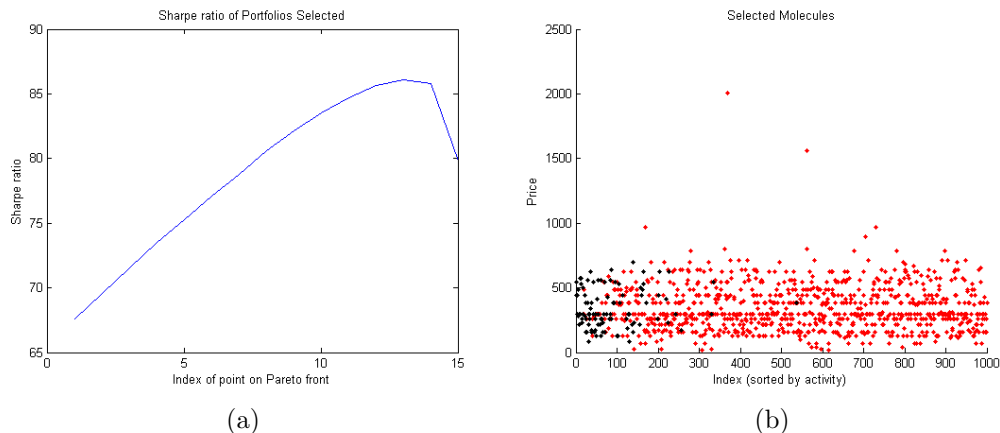


Figure 5: (a) Sharpe ratios of 15 portfolios of 100 molecules belonging to the Pareto front and selected from the set of 1000 molecules. (b) Prices and activities of the 1000 dataset molecules (in red) and of the Sharpe optimal portfolio molecules (in black).

675 obtained at the 13th iteration has the highest Sharpe ratio value and should  
 676 be selected as the most promising one for potential drug discovery. Next, we  
 677 will analyze the content of this portfolio.

678 In Figure 5 (b), the molecules of the 1000 dataset are presented. Here,  
 679 the molecules are allocated according to their activity (see X-axis) and price  
 680 (see Y-axis), respectively. The molecules selected in the Sharpe optimal  
 681 portfolio are marked in red and the non-selected molecules are depicted in  
 682 black. As can be observed from this figure, not only the cheapest molecules  
 683 are selected and not only the most active ones, but some balance between  
 684 price and activity is reached for the portfolio of molecules as a whole.



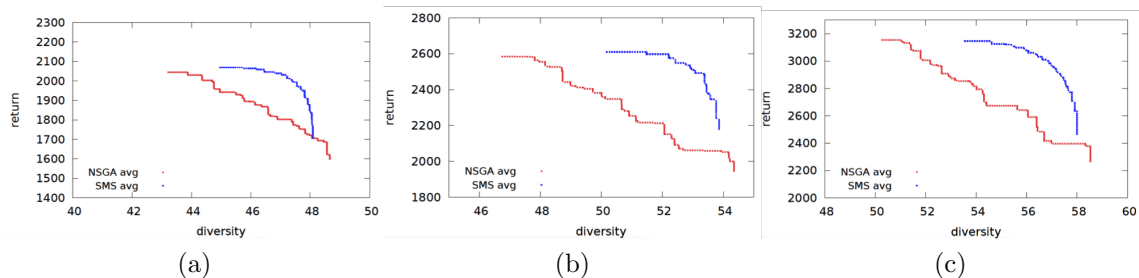


Figure 6: Comparison of NSGA-II and SMS-EMOA for Solow-Polasky diversity model with 1000, 2500 and 5000 molecules dataset, (a), (b), and (c), respectively.

685 *5.4.3. Markowitz model with Solow-Polasky diversity*

686 We now show and discuss the results obtained for model (5). Note that  
 687 these results are for MOEAs only, as application of the MIP solver is com-  
 688 plicated due to the need of obtaining the inverse of distance matrix.

689 Comparison of Pareto fronts obtained by NSGA-II and SMS-EMOA for  
 690 Solow-Polasky diversity model provided in Figures 6 (a), (b), and (c) for  
 691 datasets with 1000, 2500 and 5000 molecules, respectively, show outperfor-  
 692 mance of SMS-EMOA when compared to NSGA-II.

693 The formulation of the Solow Polasky diversity measure includes the in-  
 694 version of a matrix making it difficult to optimize this measure by means  
 695 of an exact solver, unlike the Sharpe ratio maximization formulation, which  
 696 can be solved by quadratic programming. However, it might be possible to  
 697 construct an approximation algorithm with an exact error bound for com-  
 698 puting Solow Polasky diversity measure. Based on numerical experiments we  
 699 conjecture that the Solow Polasky diversity is a submodular set function. If  
 700 this is true, a greedy subset selection heuristic would yield an approximation

701 with approximation ratio  $(1 - 1/e)$ . We were not able to provide a formal  
702 proof for submodularity and leave this question to the future work.

## 703 **6. Conclusion and future research**

704 In this work, we presented a new approach to formulating the selection  
705 of molecules for de-novo drug discovery. In particular, the well-known in  
706 finance portfolio-based approach was used to model molecular subset selec-  
707 tion for drug discovery as a portfolio selection. In addition to taking into  
708 account (bio-)activity of the molecules selected in the portfolio, the model  
709 considers the diversity of such portfolio. Moreover, it respects the limited  
710 budget provided for buying molecules and the fixed size of the portfolio as  
711 constraints. Molecules selected in the portfolio are balanced in terms of their  
712 price, expected individual performance and diversity.

713 Three models were proposed and tested on three molecular compounds  
714 datasets, in particular, classical Markowitz portfolio selection model, Sharpe  
715 ratio optimization and diversity optimization models. For solving Markowitz  
716 model with fixed size portfolio that optimizes return and risk simultaneously,  
717 we used  $\epsilon$ -constrained approach in combination with Gurobi MIP solver and  
718 applied approximation approaches, in particular, multiobjective evolution-  
719 ary algorithms, NSGA-II and SMS-EMOA. As expected QP solver was most  
720 efficient in calculating Pareto fronts except for some parts, which is due to a  
721 QP’s fixed exploration time threshold. SMS-EMOA outperformed NSGA-II  
722 for Markowitz portfolio selection model. For the single objective Sharpe ratio

723 maximization model we adjusted Gurobi MIP solver and analyzed content  
724 of the selected optimal portfolios. Finally, for solving diversity optimization  
725 model only approximate algorithms, NSGA-II and SMS-EMOA, were used,  
726 with SMS-EMOA performing better than NSGA-II on all datasets. Solv-  
727 ing this model with a quadratic solver requires obtaining inverse of distance  
728 matrix, which is difficult in practice as initial research shown. The pre-  
729 sented preliminary test results of these novel formulations obtained for three  
730 molecular compounds datasets look promising and encourage us to do future  
731 research.

732 We have also discerned a number of future research topics that could be  
733 investigated further. In particular, different formulations of risk could be  
734 tested. For instance, other popular risk measures, such as Value-at-Risk (or  
735 return-to-standard deviation index) and the diversity inversely proportional  
736 to the number of species in the population can be investigated further. It  
737 would also be interesting to construct Sharpe ratio as a tangential point to  
738 the Pareto front (with CAL) and directly compute the Sharpe ratio opti-  
739 mum via homogenization. As the initial trials indicated the later approach  
740 is really challenging, but it might turn out to be easier for alternative risk  
741 formulations. Furthermore, alternative diversity measures, e.g. Weitzman  
742 diversity [28], can be considered in optimization models. A sensitivity analy-  
743 sis for some parameters of the models (e.g., theta parameter in Solow-Polasky  
744 diversity measure) will be of value for the proposed portfolio approach.

745 Current results show that application of existing QP solvers to large size

746 problems (bigger than 5000 molecules) is difficult due to large run times. To  
747 improve exact solvers' performance for large models, relaxation of integrity  
748 constraints can be applied. This would lead to rounding-off running time to  
749 polynomial, but will require covariance matrix to be positive definite. Alter-  
750 natively, a MOEA could be used for preselection and QP solver for the final  
751 portfolio selection. It should be noted, however, that it takes approximately  
752 10 minutes for SMS-EMOA to find Pareto front of portfolios for a dataset  
753 of 1000 molecules. Hence, in this case either parallelization or fast heuristic  
754 filters can be used for preselection as well.

755 An important task for future work is not only to scale up the models  
756 proposed in this work for larger portfolios, but also to further investigate the  
757 availability of exact solvers and performance for smaller portfolios. Moreover,  
758 experience with actual performance of the models in drug discovery practice  
759 needs to be assessed by comparing data of outcomes of a larger number of  
760 *in vitro* drug discovery studies with what has been predicted by the models.

761 In this work, only structural similarity of molecules was taken into ac-  
762 count. Recent research [19] has shown that biological similarity plays an  
763 important roles in comparison of molecules. Similarly to structural diversity,  
764 biological diversity can be maximized as a third objective in the last pro-  
765 posed model. Two other models can also be adjusted to take into account  
766 biological similarity in the risk calculation. Moreover, at the later stages of  
767 drug discovery process additional objectives can be considered for molecular  
768 portfolio selection, such as minimizing side effects of the discovered lead can-

769 didates. The experimental validation of the discovered molecular portfolio  
770 via *in vitro* testing and chemists feedback on the results of such testing will  
771 be a natural next stage for the proposed in this work molecular portfolio  
772 selection approach.

## 773 **7. Acknowledgements**

774 Adriaan P. IJzerman and Eelke B. Lenselink thank the Dutch Research  
775 Council (NWO) for financial support (NWO-TOP #714.011.001).

## 776 **References**

- 777 [1] Allmendinger, R., Simaria, A., Farid, S.. Multiobjective evolution-  
778 ary optimization in antibody purification process design. *Biochemical*  
779 *Engineering Journal* 2014;91:250–264.
- 780 [2] Belton, V., Stewart, T.. *Multiple Criteria Decision Analysis: An*  
781 *Integrated Approach*. Dordrecht, Netherlands: Kluwer Academic Pub-  
782 lishers, 2001.
- 783 [3] Bonham, S.S.. *IT Project Portfolio Management*. Norwood: Artech  
784 House Inc., 2005.
- 785 [4] Brown, N.. *Chemoinformatics – An introduction for computer scien-*  
786 *tists*. *ACM Computing Surveys* 2009;41(2):8:1–8:38.
- 787 [5] Cornuejols, G., Tutuncu, R.. *Optimization Methods in Finance*. Cam-  
788 bridge, UK: Cambridge University Press, 2007.
- 789 [6] Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.. A fast and eli-  
790 tist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on*  
791 *Evolutionary Computation* 2002;6(2):182–197.
- 792 [7] Emmerich, M.T.M., Beume, N., B., N.. An EMO algorithm us-  
793 ing the hypervolume measure as selection criterion. In: Coello Coello,  
794 C., Hernández Aguirre, A., Zitzler, E., editors. *Proceedings of the*

- 795 3d International Conference on Evolutionary Multi-Criterion Optimiza-  
796 tion (EMO 2005). Berlin Heidelberg, Germany: Springer-Verlag; volume  
797 3410; 2005. p. 62–76.
- 798 [8] Enamine Ltd., . Enamine screening compounds and building blocks.  
799 2014. URL: [www.enamine.net](http://www.enamine.net).
- 800 [9] Fonseca, C.M., Grunert da Fonseca, V., Paquete, L.. Exploring the  
801 performance of stochastic multiobjective optimisers with the second-  
802 order attainment function. In: Coello Coello, C., Hernández Aguirre,  
803 A., Zitzler, E., editors. Proceedings of the 3d International Confer-  
804 ence on Evolutionary Multi-Criterion Optimization (EMO 2005). Berlin  
805 Heidelberg, Germany: Springer-Verlag; volume 3410; 2005. p. 250–264.
- 806 [10] Gallot, S.. A bound for the maximum of a number of random variables.  
807 Journal of Applied Probability 1966;3(2):556–558.
- 808 [11] Ghasemzadeh, F., Archer, N.. Project portfolio selection through  
809 decision support. Decision Support Systems 2000;29(1):73–88.
- 810 [12] Gurobi Optimization Inc., . Gurobi optimizer reference manual. 2014.  
811 URL: <http://www.gurobi.com>.
- 812 [13] Irwin, J.J., Shoichet, B.K.. ZINC - A free database of commercially  
813 available compounds for virtual screening. Journal of Chemical Infor-  
814 mation and Modeling 2005;45(1):177–182.
- 815 [14] Kirkwood, C.W.. Strategic Decision Making: Multiobjective Decision  
816 Analysis with Spreadsheets. Belmont, USA: Duxbury Press, 1997.
- 817 [15] Markowitz, H.. Portfolio selection. Journal of Finance 1952;7(1):77–91.
- 818 [16] Martello, S., Toth, P.. Knapsack Problems: Algorithms and Computer  
819 Implementations. Chichester, UK: John Wiley & Sons Ltd., 1990.
- 820 [17] Meinl, T., Ostermann, C., Berthold, M.R.. Maximum-score diversity  
821 selection for early drug discovery. Journal of Chemical Information and  
822 Modeling 2011;51(2):237–247.
- 823 [18] Miettinen, K.. Nonlinear Multiobjective Optimization. Dordrecht,  
824 Netherlands: Kluwer Academic Publishers, 1999.

- 825 [19] Paricharak, S., IJzerman, A.P., Bender, A., Nigsch, F.. Analysis of  
826 iterative screening with stepwise compound selection based on Novartis  
827 in-house HTS data. *ACS Chemical Biology* 2016;11(5):1255–1264.
- 828 [20] Pisinger, D.. The quadratic knapsack problem - A survey. *Discrete*  
829 *Applied Mathematics* 2007;155(5):623–648.
- 830 [21] Rogers, D., Hahn, M.. Extended-connectivity fingerprints. *Journal of*  
831 *Chemical Information and Modeling* 2010;50(5):742–754.
- 832 [22] Rogers, D.J., Tanimoto, T.T.. A computer program for classifying  
833 plants. *Science* 1960;132(3434):1115–1118.
- 834 [23] Solow, A.R., Polasky, S.. Measuring biological diversity. *Environmental*  
835 *and Ecological Statistics* 1994;1(2):95–103.
- 836 [24] Steuer, R.E., Qi, Y., Hirschberger, M.. Comparative issues in large-  
837 scale mean-variance efficient frontier computation. *Decision Support*  
838 *Systems* 2011;51(2):250–255.
- 839 [25] Tanimoto, T.T.. An Elementary Mathematical theory of Classification  
840 and Prediction. Technical Report 8; IBM Internal Report; 1958.
- 841 [26] van Westen, G.J.P., van den Hoven, O.O., van der Pijl, R., Mulder-  
842 Krieger, T., de Vries, H., Wegner, J.K., IJzerman, A.P., van Vlijmen,  
843 H.W.T., Bender, A.. Identifying novel adenosine receptor ligands by  
844 simultaneous proteochemometric modeling of rat and human bioactivity  
845 data. *Journal of Medicinal Chemistry* 2012;55(16):7010–7020.
- 846 [27] Walters, W.P., Stahl, M.T., Murcko, M.A.. Virtual screening – An  
847 overview. *Drug Discovery Today* 1998;3(4):160–178.
- 848 [28] Weitzman, M.L.. On diversity. *The Quarterly Journal of Economics*  
849 1992;107(2):363–405.
- 850 [29] Yevseyeva, I., Guerreiro, A.P., Emmerich, M.T.M., Fonseca, C.M..  
851 A portfolio optimization approach to selection in multiobjective evolu-  
852 tionary algorithms. In: Bartz-Beielstein, T., Branke, J., Filipič, B.,  
853 Smith, J., editors. *Proceedings of the 13th International Conference*  
854 *on Parallel Problem Solving from Nature - PPSN XIII (PPSN 2014)*.  
855 Switzerland: Springer International Publishing; volume 8672; 2014. p.  
856 672–681.