Learning Parameter Sets for Alignment Advising

Dan DeBlasio
John Kececioglu

Department of Computer Science
University of Arizona

COMPUTER SCIENCE
UA SCIENCE

THE UNIVERSITY OF ARIZONA®

Motivation

Multiple sequence alignment is a fundamental problem in bioinformatics.

- multiple sequence alignment is NP-Complete
- many popular aligners for multiple sequence alignment
- each aligner has many parameters whose values affect the accuracy of the alignment

alternate

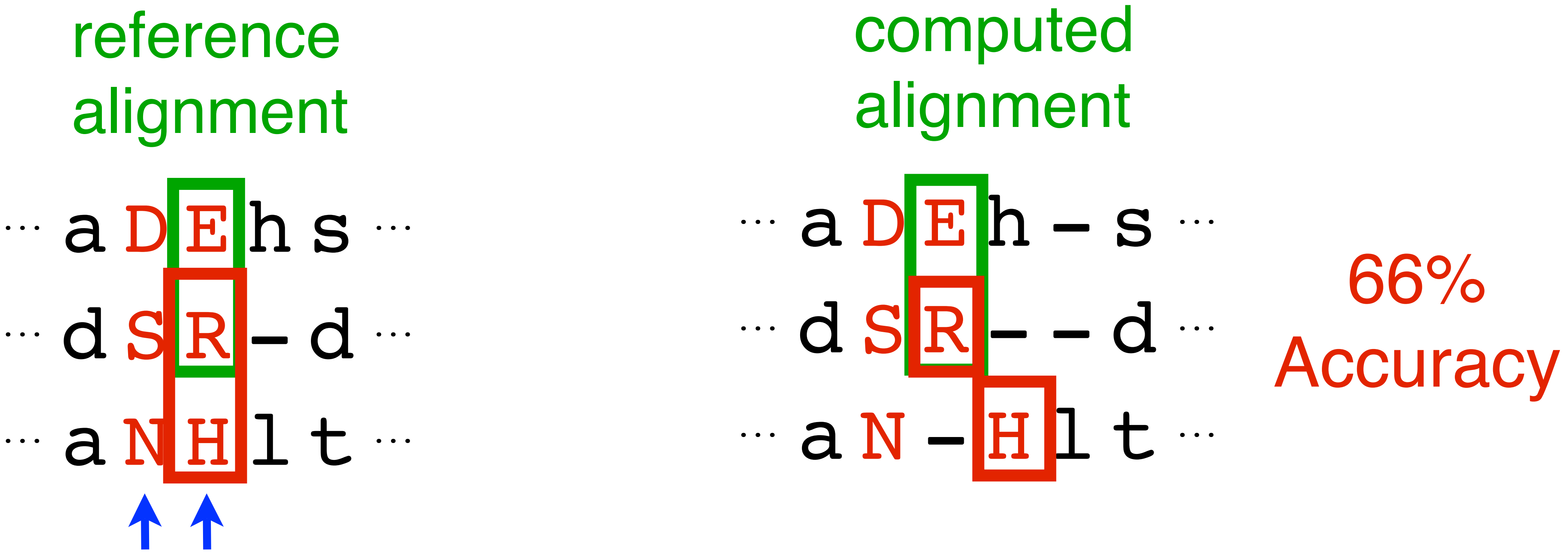
```
... gsvenrarlvlevvdavcnewsad-RIGIRVSPigtfqnvdngpnee--adalyl--- ...
... ydfeatekllke-----vftfftk-PLGVKLPPyf---------------dlvhfdim ...
... gsienrarftlevvdalveaighe-KVGLRLSPygvfnsmsggaetgivaqyayvage ...
... gslenrarfwletlekvkhavgsdcAIATRFGV-----------------dtvygpgq ...
... tdpevaaalvka-----ckavskv-PLYVKLSPnvt--------------divpiaka ...
```

default

```
... yl-lhqflspssnqrtdqyggsvenrarlvlevvdavcnewsad-RIGIRVSPigtfq ...
... kP-LGVKLPPyf--dlvhfdimaeilnqfpltyvsnv-nsig----nglfidpeaesv ...
... yl-lnqfldphsntrtdeyggsienrarftlevvdalveaighe-KVGLRLSPygvfn ...
... yl-plqflnpyynkrtdkyggslenrarfwletlekvkhavgsdcAIATRF---GVdt ...
... kvPLYVKLSPnv-tdivpiakaveaagadgltmintl---------mgvrfdlktrqp ...
```

Motivation

Alignment accuracy is measured with respect to a reference alignment.



reference alignment

```
··· a D E h s ···
··· d S R – d ···
··· a N H l t ···
```

computed alignment

```
··· a D E h – s ···
··· d S R – – d ···
··· a N – H l t ···
```

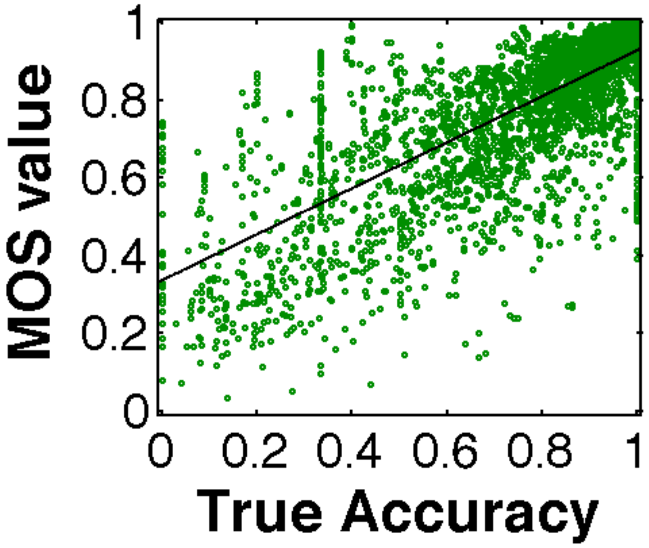
66% Accuracy

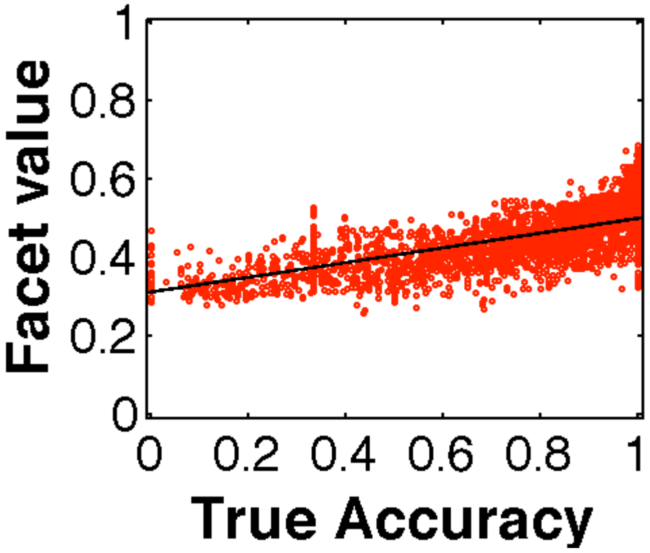
- accuracy is the fraction of substitutions from the reference that are in the computed alignment,
- measured on the core columns of the reference.

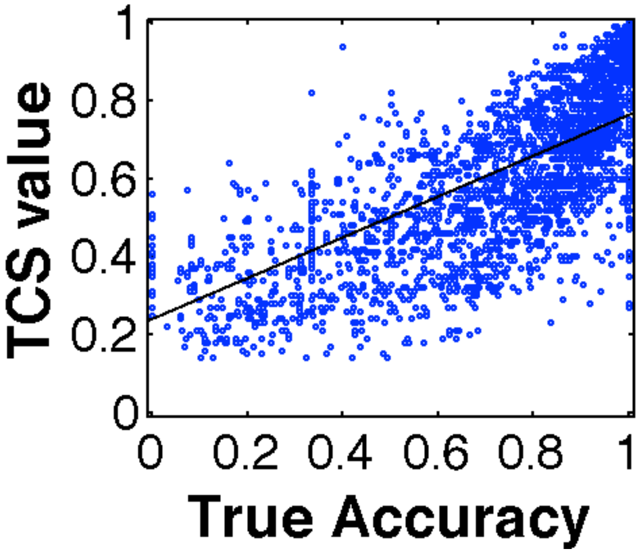
Accuracy estimators

The best estimators of alignment accuracy *without a reference* include:

- MOS [Lassmann and Sonnhammer, 2005]
- PredSP [Ahola, *et al.*, 2008]
- Guidance [Penn, *et al.*, 2010]
- Facet [Kececioglu and DeBlasio, 2013]
- TCS [Chang, Tommaso and Notredame, 2014]







Parameter advising

Aligners often use *one* default parameter choice for *all* inputs.

- The default has good *average accuracy* across all benchmarks.
- The optimal default choice can be found by inverse alignment [Kececioglu and Kim 2007].
- The default may be a poor choice for specific inputs.

Parameter advising

Parameter advising for input sequences S is

- selecting the parameter choice p from a set P
- for which the alignment output by aligner $\mathbb{A}$
- has the highest value under estimator E.

$$\text{Choice}(P, S) \;:=\; \operatorname*{argmax}_{p \,\in\, P} \; E\Big(\mathbb{A}_p(S)\Big)$$

An oracle is a *perfect* advisor whose “estimator” is true accuracy.

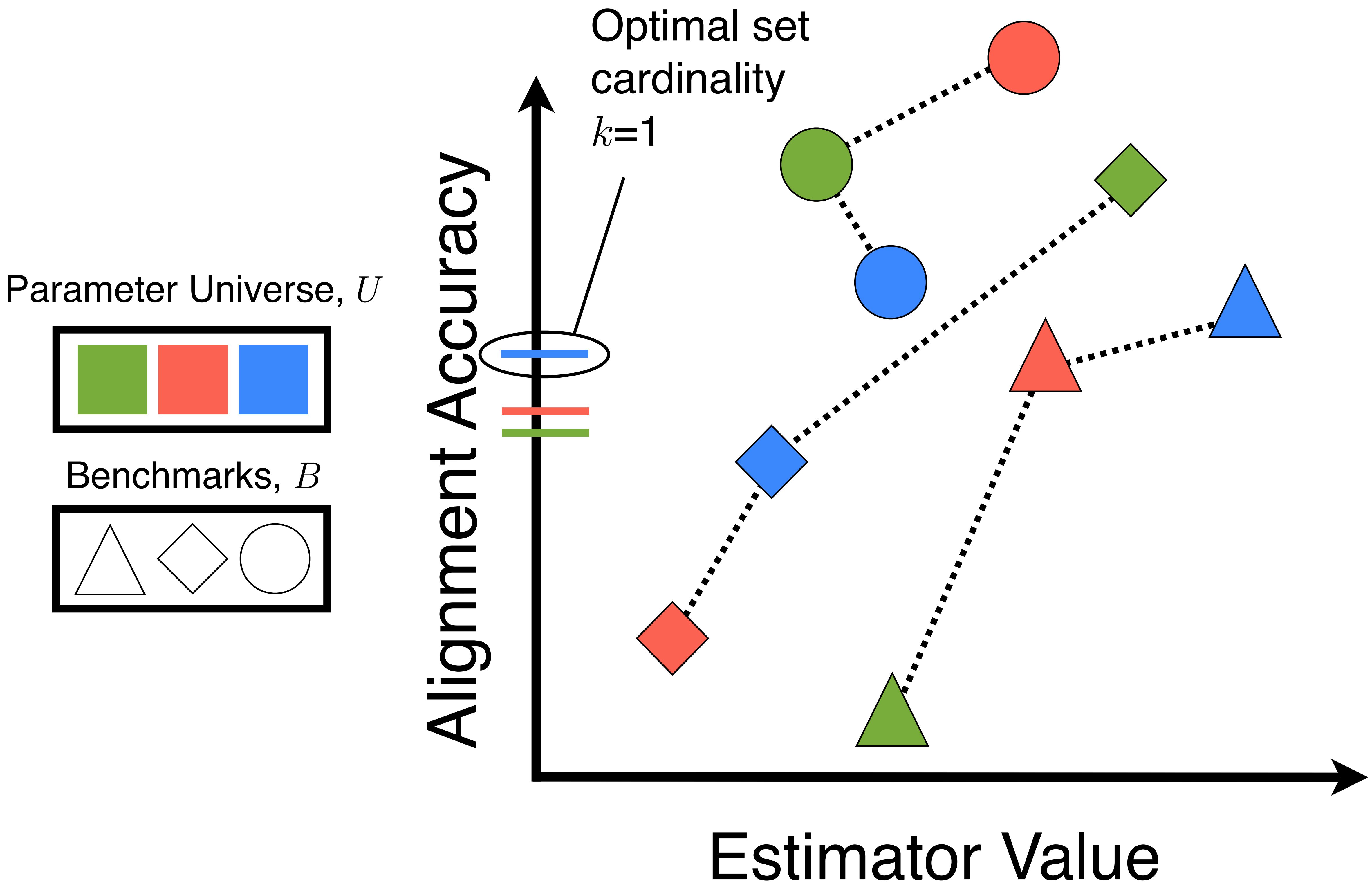
Parameter advising

A parameter advisor has two components:

- an accuracy estimator $E(A)$, and
- a set of candidate parameter choices P.

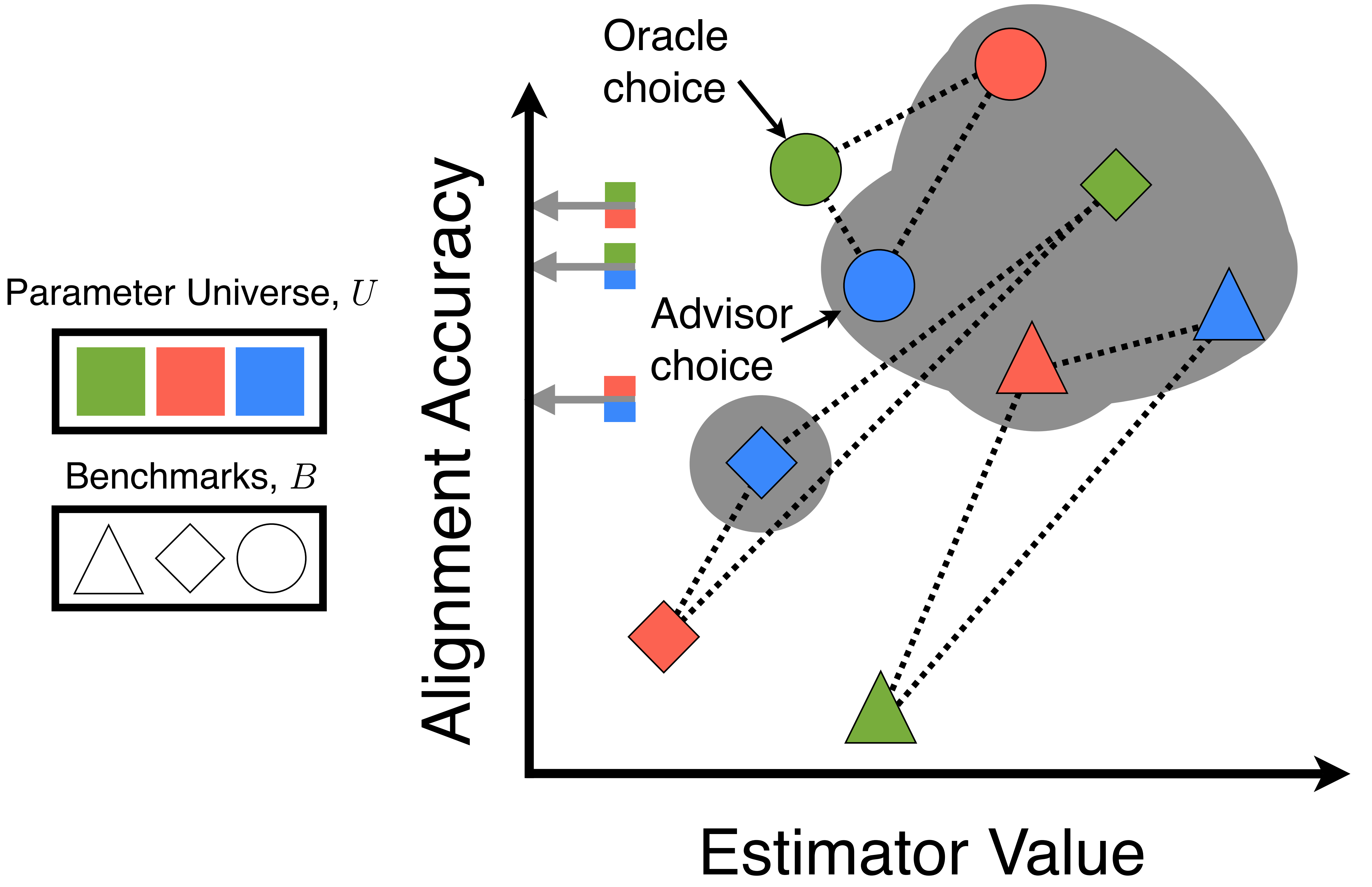
Given accuracy estimator E,
what is the *optimal set*
of parameter choices P?

Advisor Set problem



Optimal set cardinality k=1

Advisor Set problem



Parameter Universe, U

Benchmarks, B

Oracle choice

Advisor choice

Alignment Accuracy

Estimator Value

Advisor Set problem

A parameter choice j assigns values to all parameters.

- For the `Opal` aligner, a parameter choice is a 5-tuple

$$(\sigma, \gamma_I, \gamma_E, \lambda_I, \lambda_E$$

- Universe U is the set of all parameter choices.

Advisor Set problem

Each benchmark i consists of:

- a set S_i of protein sequences, and
- its reference alignment.

To correct for bias in easy benchmarks we assign each a weight w_i.

Advisor Set problem

We learn the advising set using examples consisting of

- an alignment $A_{ij} = \mathbb{A}_j(S_i)$
- the associated estimated accuracy $e_{ij} = E(A_{ij})$,
- the true accuracy a_{ij} of A_{ij}.

Advisor Set problem

Given these examples, we would like to find:

- over all subsets P of size at most k from the universe U,
- the optimal subset P^* that has highest average advising accuracy on the benchmarks.

Advisor Set problem

For ties in the estimator, the advisor accuracy is not well-defined.

- Consider the parameter choices that are tied for maximizing the estimator.
- We take the advisor's accuracy to be its expected value on these choices.
- To aid generalization, we include choices that are close to maximizing the estimator.

$$\mathrm{Accuracy}_i(P) := \left(\begin{array}{l} \text{Average accuracy of alignments} \\ \text{of benchmark } i \text{ using parameters } j \in P \\ \text{where } e_{ij} \text{ is within } \epsilon \text{ of the maximum} \end{array} \right)$$

Advisor Set problem

For the Advisor Set problem the input is

- cardinality bound k,
- universe of parameters choices U,

along with the error tolerance, and for all examples, their estimator values, accuracies, and weights.

Advisor Set problem

The output is

- an optimal set $P \subseteq U$ of parameter choices with $|P| \leq k$, that maximizes the objective function

$$\sum_i w_i \ \mathrm{Accuracy}_i(P)$$

Advisor Set problem

THEOREM (Problem Complexity)

The Advisor Set problem is NP-complete.

- Polynomial-time solvable for fixed k
- Reduction is from the Dominating Set problem
- Oracle sets can be found for all k in practice

Approximation algorithm

A natural greedy procedure finds good sets.

(1) Start with an optimal set $\tilde{P}$ of size at most ℓ

(2) Find parameter choice p^* such that

$$p^* = \operatorname*{argmax}_{p \in U - \tilde{P}} \left\{ \sum_i w_i \ \mathrm{Accuracy}_i \left(\tilde{P} \cup \{p\} \right) \right\}$$

(3) Update $\tilde{P} := \tilde{P} \cup \{p^*\}$

(4) Repeat (2) and (3) until $|\tilde{P}| = k$

(5) Of all these $\tilde{P}$, return the best one under the objective function

Approximation algorithm

An α-approximation algorithm

- finds a feasible solution in polynomial-time
- whose objective value is at least α times the optimal solution
- where $\alpha < 1$ for a maximization problem
- α is called the approximation ratio

Approximation algorithm

THEOREM (Approximation Algorithm)

The greedy procedure is an $\frac{\ell}{k}$-approximation algorithm for Advisor Set, with constant ℓ and ϵ = 0.

The approximation ratio $\frac{\ell}{k}$ is tight.

Experimental results

To evaluate the accuracy of advising, we consider:

- `PredSP`, `MoS`, `Guidance`, `Facet`, and `TCS` estimators,
- over 800 benchmarks from `BENCH` and `PALI`,
- a universe of over 200 parameter choices,
- evaluated with k-fold cross validation,
- advising for the `Opal` aligner.

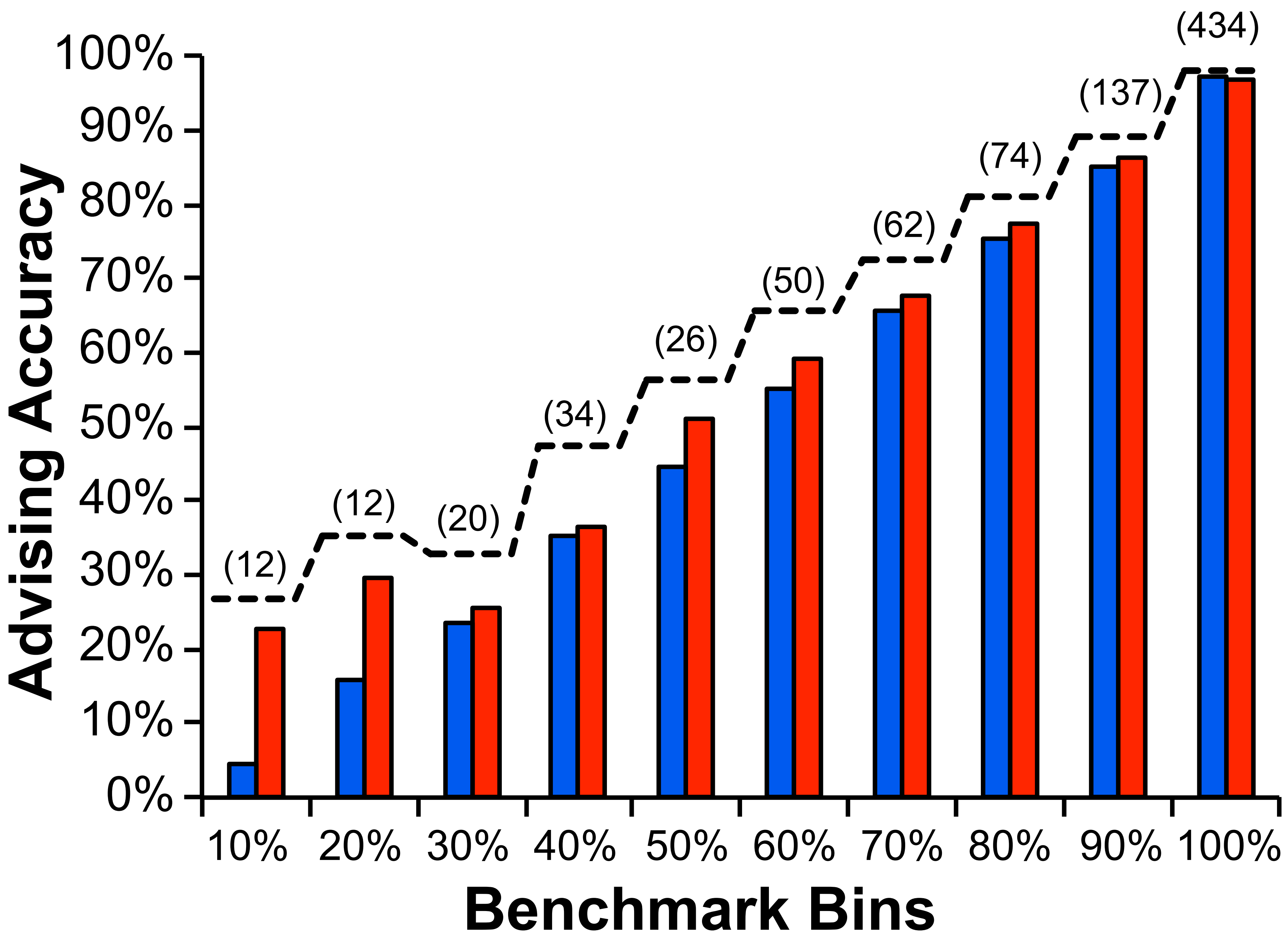
Experimental results

We correct for the bias in over-representation of easy-to-align benchmarks.

- The difficulty of a benchmark is its accuracy under the default parameter setting.
- Split the range of difficulties [0,1] into 10 bins.
- Report advisor accuracy as the average across bins.

Experimental results

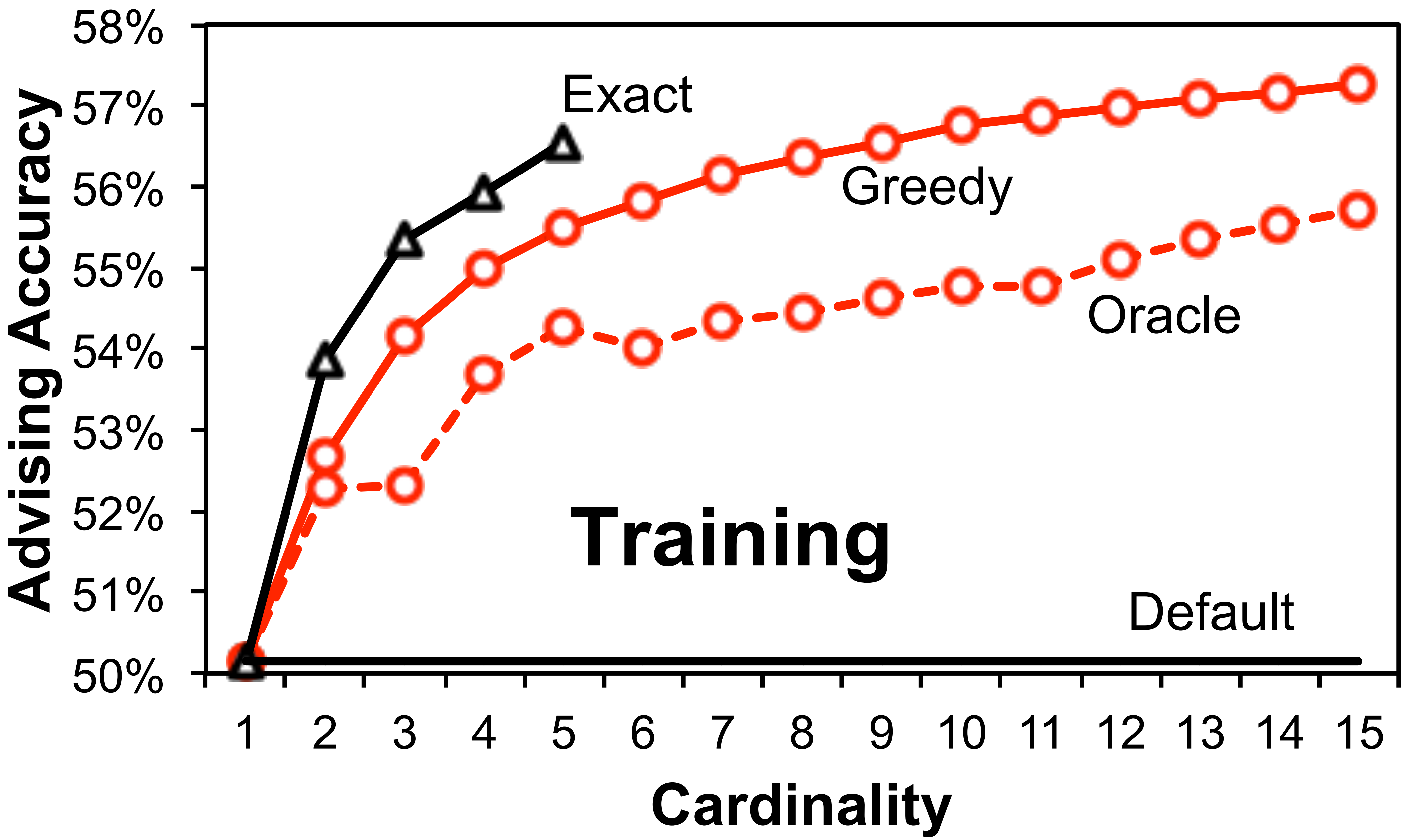
Average accuracy of advisors by difficulty bin



Boosts the accuracy on the hardest bins by almost 20%

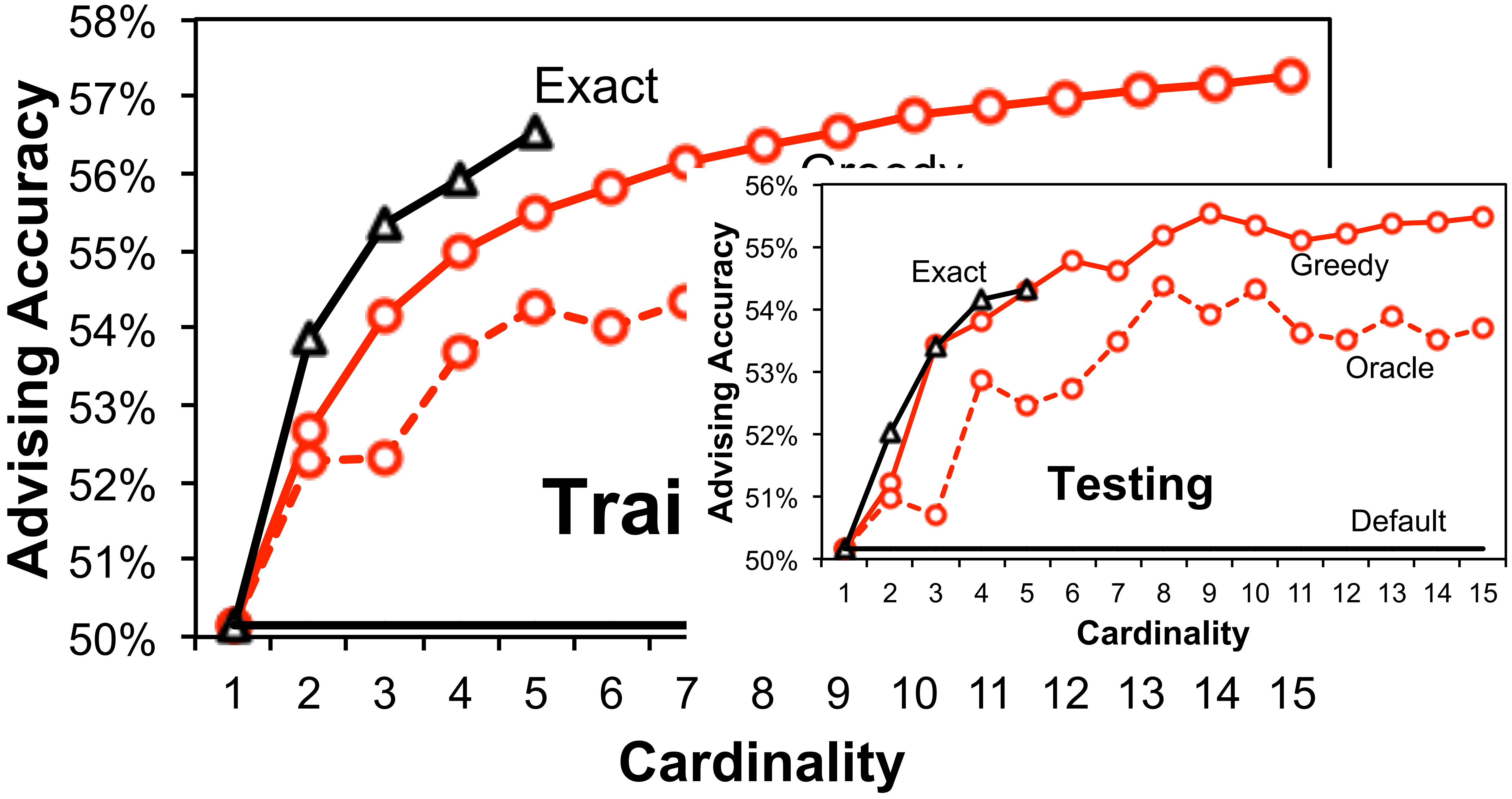
Experimental results

Advisor performance versus parameter set cardinality



Experimental results

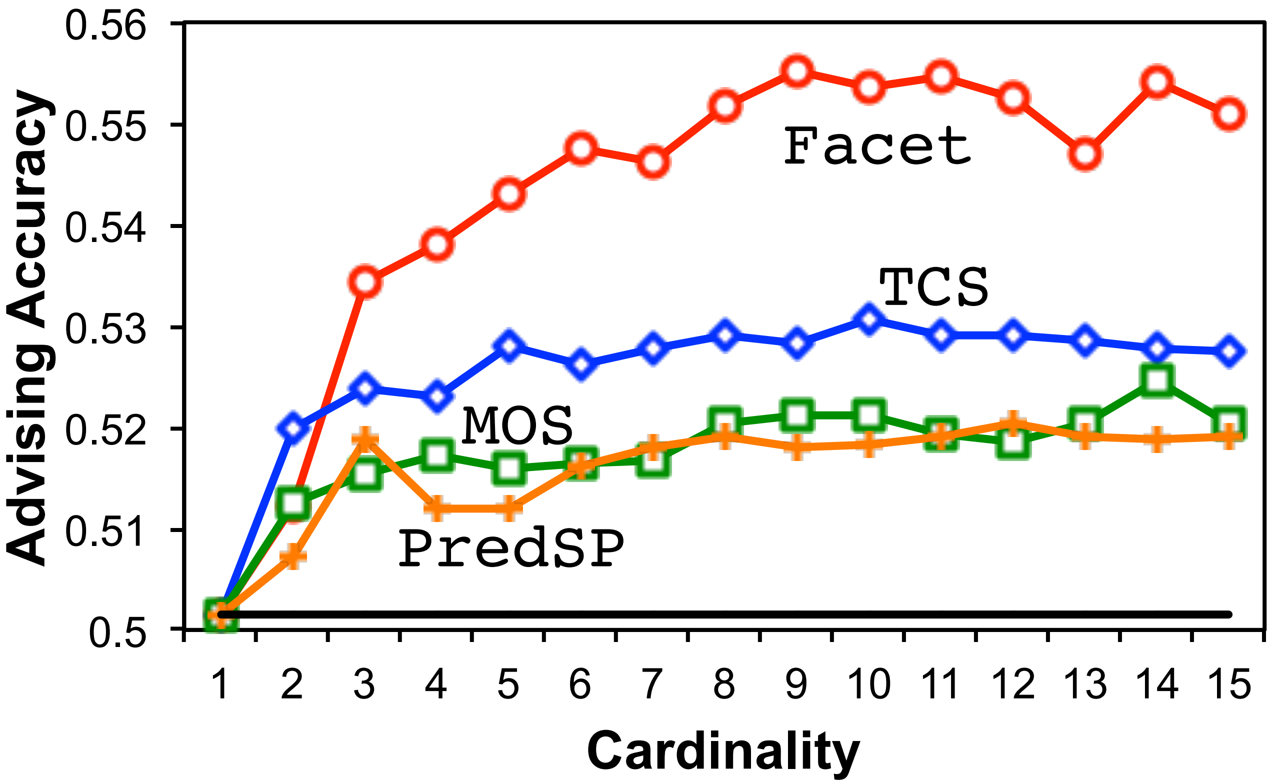
Advisor performance versus parameter set cardinality

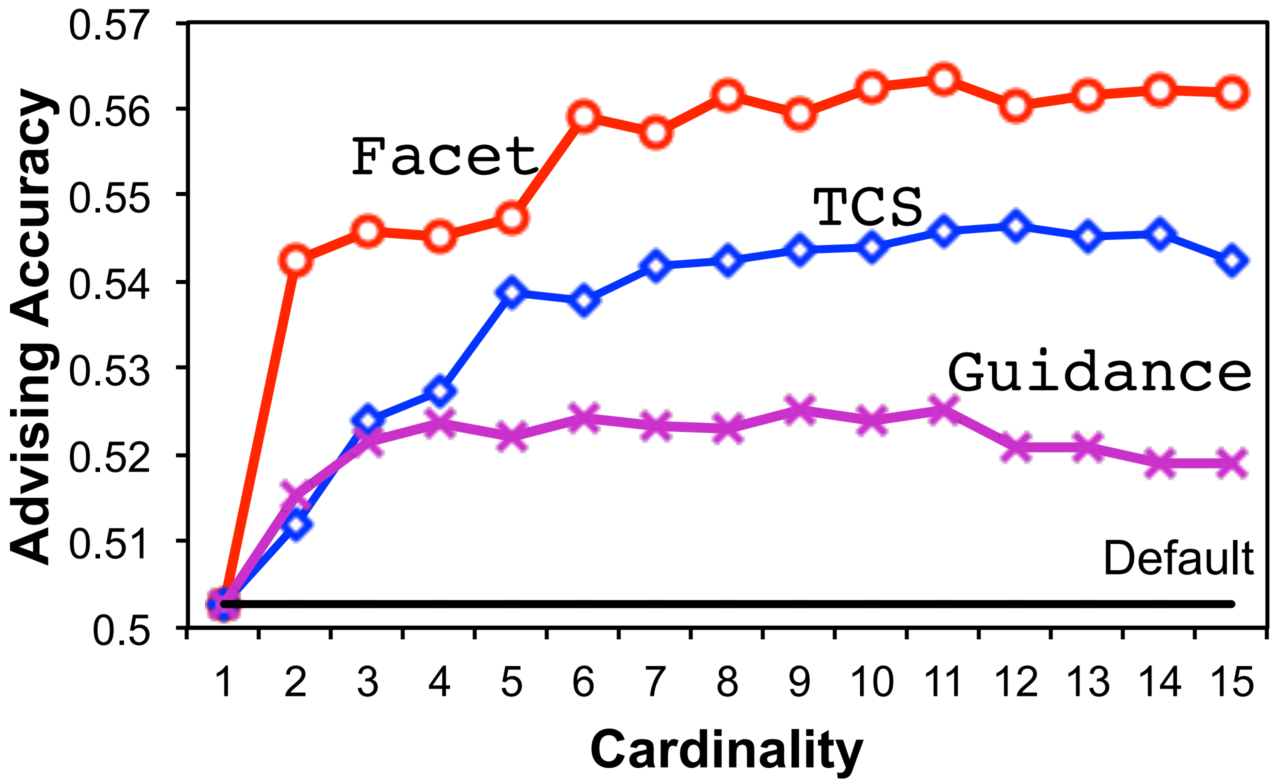


Greedy sets generalize better than exact sets

Experimental results

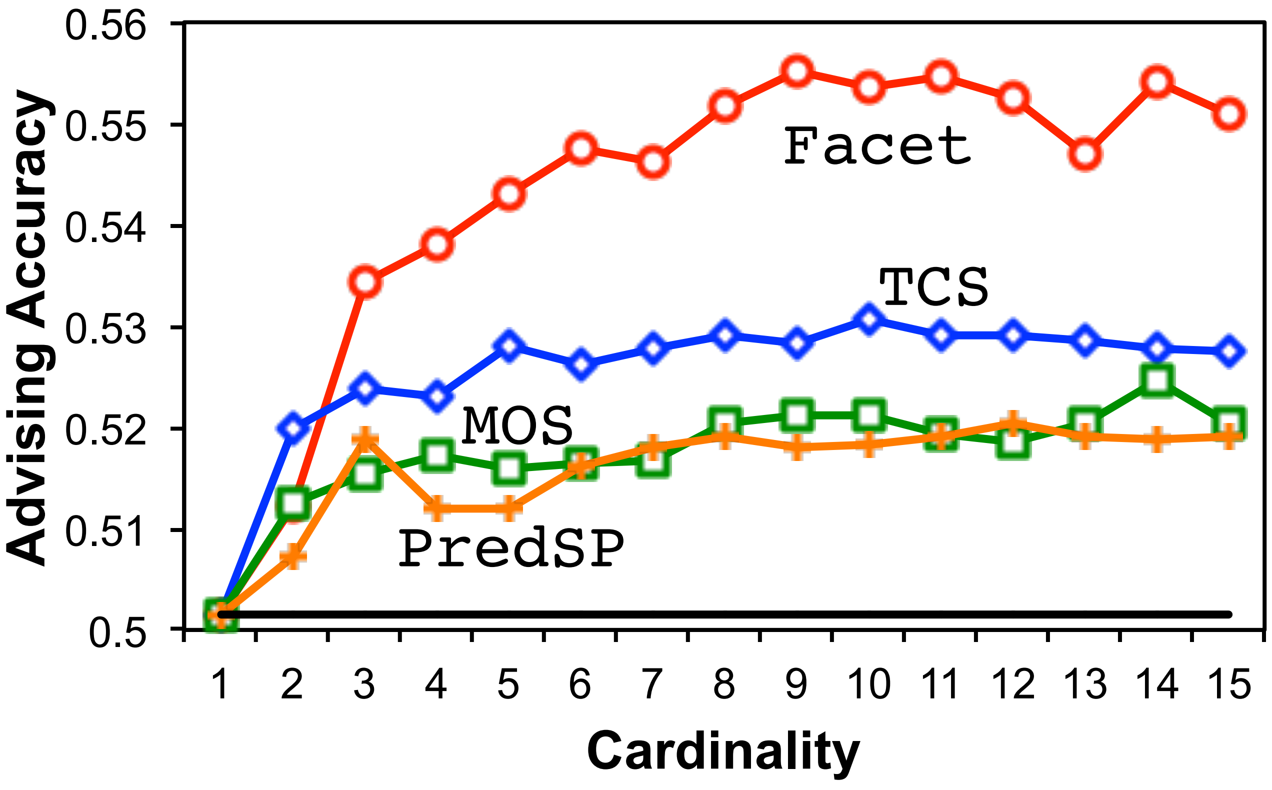
Advising performance for various estimators

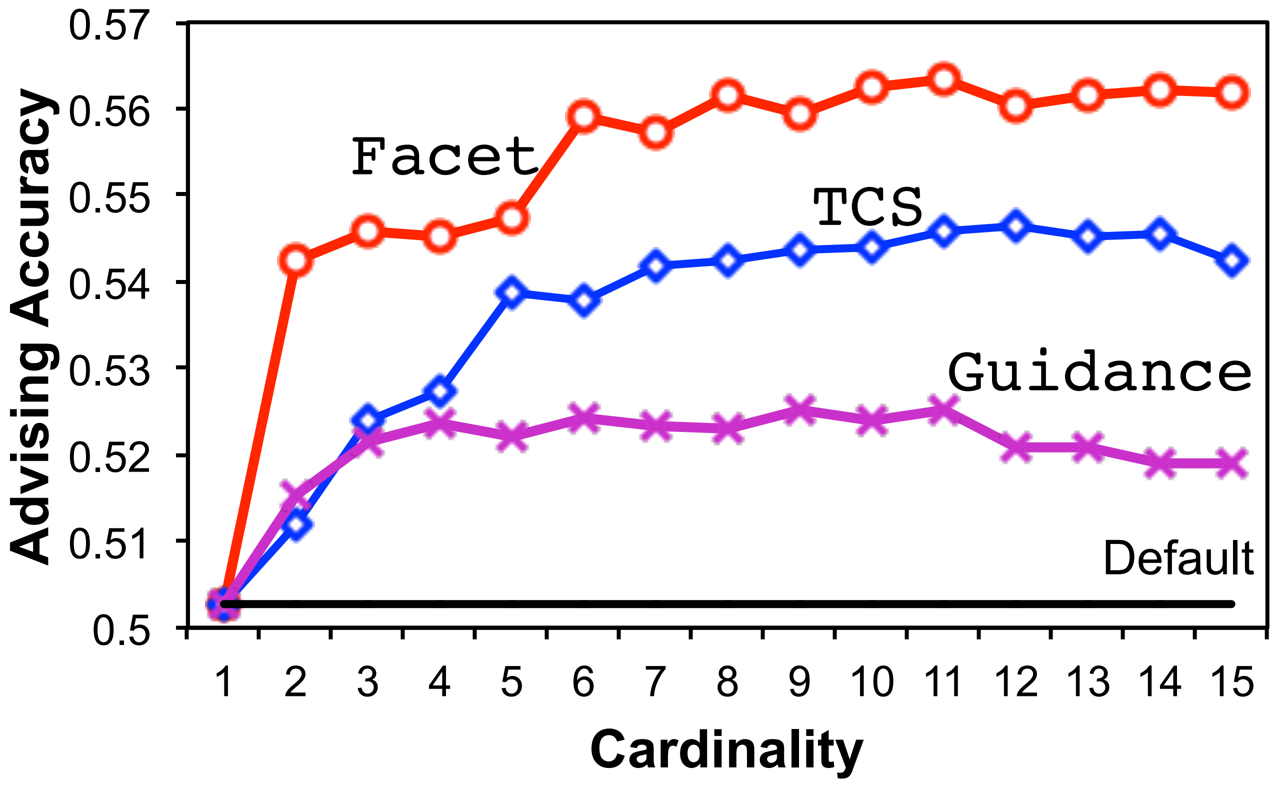




Experimental results

Advising performance for various estimators





Facet outperforms other accuracy estimators

Experimental results

Greedy parameter sets for `Opal` using `Facet`

Cardinality	Parameter choices (σ, γ_I, γ_E, λ_I, λ_E)	Average advising accuracy
1	(VTML200, 50,17, 41,40)	51.2%
2	(VTML200, 55, 30, 45, 42)	53.4%
3	(BLOSUM80, 60, 26, 43, 43)	54.5%
4	(VTML200, 60, 15, 41, 40)	55.2%
5	(VTML200, 55, 30, 41, 40)	55.6%

Sets include different families of substitution matrices

Conclusions

Parameter advising gives a significant improvement in alignment accuracy.

- Learning an optimal set for advising is NP-complete.
- A greedy approach yields an $\frac{\ell}{k}$-approximation algorithm.
- Greedy sets generalize better than exact sets.
- On the hardest benchmarks, boosts the accuracy by almost 20%.

Further research

Further improvement in advising will not come from learning better parameter sets.

Promising directions include,

- Developing estimators that better correlate with true accuracy
- Extending to DNA sequence alignments
- Extending parameter advising to aligner advising

Software distribution

Available for download:

- `Facet` estimator tool
- Precomputed parameter sets for `Opal` aligner
- Benchmark suites with structure predictions

`facet.cs.arizona.edu`

Acknowledgments

People

William Pearson

Travis Wheeler

Funding

- University of Arizona NSF IGERT in Genomics Grant DGE-0654435
- US NSF Grant IIS-1217886
- ACM-BCB Conference Travel Grant